감리 점검 가이드 전 195건 수록



AI SERVICE AUDIT GUIDE · 2026

AI 서비스 구축 사업
감리 점검 가이드

생성형 AI 정보화사업을 위한
단계별·영역별 감리 점검체계

사업관리 · 품질보증 · 응용시스템 · 데이터베이스 · 시스템구조·보안

정보시스템감리협회
KOREA INFORMATION SYSTEM SUPERVISION ASSOCIATION

발간사



우리 사회는 지금 인공지능(AI)을 중심으로 빠르게 변화하고 있습니다. AI는 특정 기술 분야를 넘어 국가 경쟁력과 사회 전반의 혁신을 좌우하는 핵심 기반으로 자리 잡았으며, 국가 간 경쟁도 더욱 치열해지고 있습니다.

정부 또한 'AI 3대 강국 도약'을 주요 목표로 대규모 투자를 추진하고 있으며, 공공분야 정보시스템 구축사업에서도 AI 기술의 도입과 활용이 빠르게 확대되고 있습니다.
이러한 변화는 정보시스템 감리에도 새로운 역할을 요구하고 있습니다.

AI가 정보시스템의 우리 사회는 지금 인공지능(AI)을 중심으로 빠르게 변화하고 있습니다. AI는 특정 기술 분야를 넘어 국가 경쟁력과 사회 전반의 혁신을 좌우하는 핵심 기반으로 자리 잡았으며, 국가 간 경쟁도 더욱 치열해지고 있습니다.

정부 또한 'AI 3대 강국 도약'을 주요 목표로 대규모 투자를 추진하고 있으며, 공공분야 정보시스템 구축사업에서도 AI 기술의 도입과 활용이 빠르게 확대되고 있습니다.
이러한 변화는 정보시스템 감리에도 새로운 역할을 요구하고 있습니다. AI가 정보시스템의 핵심 요소로 자리 잡으면서 감리 역시 기존의 시스템·데이터·사업관리 중심의 점검을 넘어, 데이터의 적정성, AI 모델의 성능과 신뢰성, 운영 · 모니터링, 보안과 안전성, 프라이버시와 책임성 등 AI 특유의 요소까지 체계적으로 검증할 수 있어야 합니다.

정보시스템감리협회는 이러한 환경 변화에 대응하기 위해 AI 분야 정보시스템 감리 가이드에 관한 연구를 수행하였습니다. 특히 생성형 AI와 대규모 언어모델(LLM)의 확산으로 새로운 기술적·관리적 위험이 빠르게 등장하고 있어, 현장에서 활용할 수 있는 체계적인 감리방법과 점검기준 마련의 필요성이 더욱 커지고 있습니다.

이번에 발간하는 「AI 기반 정보시스템 감리 가이드」는 그동안의 연구와 세미나, 실제 AI 정보시스템 감리 사례를 바탕으로 마련한 실무 중심의 가이드입니다.
본 가이드는 기존 정보시스템 감리체계를 기반으로 AI 생명주기와 데이터 · 모델 · LLMOps · 응용서비스 · 거버넌스 등 AI 기술요소를 연계하여 감리 프레임워크를 구성하였습니다. 또한 보안 · 안전성, 프라이버시, 책임성, 투명성, 공정성 등 AI 신뢰성 확보를 위한 주요 요소를 체계화하고, 현장에서 활용할 수 있는 점검항목과 점검방법을 제시하였습니다.

이번 가이드는 AI 감리의 완결된 기준이라기보다 지속적으로 발전시켜 나갈 출발점입니다. 협회는 앞으로 현장의 감리사례와 기술 발전, 국내외 정책과 제도의 변화를 지속적으로 반영하여 가이드를 보완해 나가겠습니다. 아울러 감리교육과 연계한 AI 감리 전문인력 양성체계를 마련하여 감리인들이 AI 정보시스템을 효과적으로 검증할 수 있는 전문역량을 갖추도록 지원할 계획입니다.

아무쪼록 이번 가이드가 AI 정보시스템 감리를 수행하는 현장의 실질적인 지침이 되고, AI 정보시스템의 품질과 신뢰성을 높이는 데 도움이 되기를 기대합니다.
나아가 정보시스템 감리가 신뢰할 수 있는 AI 사회를 구현하는 데 중요한 역할을 수행하고, 우리나라의 AI 경쟁력 강화에 기여하는 데 이번 가이드가 의미 있는 디딤돌이 되기를 바랍니다.

2026년 8월
정보시스템감리협회
회장 조 병 휘

감수의 글



AI 시대, 기술과 감리에 대한 신뢰할 수 있는 이정표

이제 인공지능(AI) 기술은 단순한 업무 자동화를 넘어 기업의 의사결정과 국가기반시설의 핵심 엔진으로 자리 잡았으며 공공기관들이 발주하는 정보시스템 구축사업에는 예외 없이 AI의 적용을 요구하고 있습니다. 그러나 AI 도입이 확대될수록 데이터 편향성, 모델의 블랙박스화(해석 불가성), 환각 현상(Hallucination), 보안 및 윤리적 리스크 등 기존 IT 시스템에서는 보지 못했던 새로운 위험 요인들이 수면 위로 떠오르고 있습니다.

이러한 시점에 발간되는《AI 서비스 구축사업 감리 점검가이드》는 AI가 광범위하게 활용되고 있는 현실에서 AI의 특장점과 위협요소 모두를 고려한 감리 체크리스트 및 기법을 소개하고 있어 매우 적시성 있고 뜻깊은 시도라고 하겠습니다. 기존 정보시스템 감리의 프레임워크가 성능, 기능, 안정성에 초점을 맞추었다면, AI 기술이 광범위하게 활용된 정보시스템 감리는 AI 특유의 생애주기, 즉 데이터 수집 및 전처리, 모델 학습, 검증, 배포 및 모니터링 전반을 관통하는 새로운 감리기준과 체크리스트를 필요로 합니다

이 책은 이론적 설명에 치우치지 않고, 감리 현장에서 즉시 적용할 수 있는 데이터 품질 검증 방법, 모델 공정성 평가 기법, AI 윤리 가이드라인 준수 여부 판정 기준 등을 구체적인 사례와 함께 다루고 있습니다 예를 들어 AI 적용 정보시스템 구축사업의 위험요인으로 드는 학습데이터의 원본을 검증하는 데 한계가 있다는 점에 대하여는 현행 감리기법이 산출물과 원자료를 대조하여 확인하는 반면 AI 학습데이터는 사업자의 자산이거나 외부에서 수집된 것이어서 원본을 전수 대조하기 어렵다는 차이를 설명한 후, AI 감리에서는 데이터 자체가 아니라 수집·정제·라벨링 과정의 기록을 증적으로 삼아야 한다는 점과 출처와 라이선스, 계보, 교차검수 이력이 남아 있는지를 확인하여야 한다는 점을 접근방안으로 제시하고 있습니다.

AI 기술발전 속도가 너무 빨라져 감리와 영역에서도 패러다임 자체가 달라지는 상황에서, 이 책은 감리원에게는 명확한 검증의 잣대를 제공하고, AI 시스템 구축 사업자와 발주기관에게는 안전하고 효율적인 정보시스템 도입을 보장하는 든든한 나침반이 되어 줄 것입니다. AI가 적용된 정보시스템의 신뢰성과 안전성 확보를 위해 고심하는 모든 IT 전문가와 감리원들에게 일독을 권합니다.

감수자 김 남 석

前 행정안전부 차관 , 전자정부본부 본부장, 우즈베키스탄 ICT위원회 부위원장

감수의 글



AI 시대에 부합하는 감리 체계의 진화

인공지능은 이제 정보시스템에 부가되는 하나의 기능을 넘어 업무와 서비스의 설계 방식 자체를 변화시키고 있습니다. 특히 생성형 AI는 새로운 가능성과 함께 기존 정보시스템과는 다른 불확실성과 위험을 가져오고 있습니다. 구축 이후에도 성능이 변화할 수 있습니다. 따라서 AI 시대에는 기술의 도입 못지않게 성과를 어떻게 검증하고, 위험을 어떻게 관리하며, 변화에 어떻게 지속적으로 대응할 것인가가 중요합니다. 이러한 점에서 「AI 서비스 구축사업 감리 점검가이드」의 발간은 매우 뜻깊습니다. 사업목표와 성과지표 에서부터 데이터, 모델, 응용서비스, 인프라와 보안, 품질관리, 인간의 감독과 운영 모니터링까지 유기적으로 연결되어야 합니다. 감리 역시 완성된 시스템의 오류나 산출물의 미비점을 확인하는 것을 넘어, 사업의 목표와 성능이 검증 가능한 형태로 정의되고 주요 위험이 체계적으로 관리되고 있는지를 살펴보는 역할로 확장될 필요가 있습니다.

본 가이드의 가장 큰 강점은 이러한 문제의식을 실제 현장에서 활용할 수 있는 구체적인 점검체계로 구현했다는 점입니다. 사업관리, 품질보증, 응용시스템, 데이터베이스, 시스템구조·보안에 걸쳐 195개의 세부 점검가이드를 마련하여 AI 서비스 구축사업의 전반을 폭넓게 다루고 있습니다. 새로운 감리체계를 별도로 만드는 대신 기존 정보시스템 감리의 단계와 영역을 준용하면서 AI 생명주기와 신뢰성 관점을 연계한 점도 실용적입니다. 특히 각 점검가이드를 국내외 법령, 가이드라인과 표준 등 다양한 근거자료에 기반하여 구성하고, 목적과 필요성뿐 아니라 구체적인 검토방법, 판정기준, 유의사항과 실무 Tip까지 제시함으로써 현장에서 무엇을 확인하고 어떠한 근거로 판단할 것인지를 구체화하였습니다.

또한 빠르게 변화하는 AI의 특성을 감안하여 현재 가이드가 다루는 기술의 범위와 한계도 명확히 제시하고 있다는 점이 인상적입니다. 앞으로 외부 모델의 활용과 조합, 검색증강생성(RAG), 멀티모달 AI, 에이전틱 AI 등 기술과 활용 방식이 더욱 다양해지는 만큼, 현장의 경험과 기술·제도의 변화를 지속적으로 반영한다면 다양한 AI 사업의 특성을 더욱 정교하게 포착하는 기준으로 발전할 수 있을 것입니다.

이번 「AI 서비스 구축사업 감리 점검가이드」가 감리인에게는 AI 사업을 보다 체계적으로 이해하고 점검하는 데 도움이 되고, 발주기관과 사업자에게도 성과와 위험을 함께 고려하며 사업을 추진하는 데 유용한 참고자료가 되기를 기대합니다. 앞으로 현장의 경험과 기술·제도의 변화를 꾸준히 담아내며, 우리나라 AI 사업의 품질과 신뢰성을 높이는 데 오래 활용되는 가이드로 자리 잡기를 바랍니다.

감수자 김 숙 경

KAIST 기술경영학부 초빙교수, 한국AI서비스학회 정부공공AI서비스위원회 위원장

일러두기

본 문서는 생성형 인공지능(AI) 기술이 적용된 정보화사업의 감리에서 확인하여야 할 사항을 단계별·영역별로 정리한 점검 가이드이다. 다음의 사항을 먼저 확인한 뒤 본문을 활용하기 바란다.

1. 본 문서의 성격

본 문서는 「정보시스템 감리기준」에 따라 감리를 수행하는 데 필요한 공식적인 감리 수행가이드가 아니다. 감리의 절차와 보고서 양식, 점검결과의 판정 체계는 현행 감리기준과 감리지침을 그대로 따르며, 본 문서는 그 위에서 생성형 AI 특유의 점검 관점을 정리한 참고용 자료이다.

본 문서는 연구용역의 결과물로서, 감리 현장에서 참고할 수 있도록 점검항목과 검토 방법을 예시한 것이다. 따라서 본 문서의 점검항목은 감리인이 반드시 점검하여야 할 의무 사항이 아니며, 개별 사업의 특성에 따라 취사선택하여 적용한다.

구 분	내 용
문서의 종류	감리 점검 가이드 (감리 수행가이드가 아님)
문서의 성격	연구용역 결과물 — 감리 현장 참고용 자료
판 본	Ver 1.0
적용의 강제성	없음. 개별 사업의 특성에 따라 취사선택하여 적용
현행 기준과의 관계	「정보시스템 감리기준」 및 현행 감리지침을 대체하지 않고 보완
점검항목 수	195건 (사업관리·품질보증·응용시스템·데이터베이스·시스템구조·보안)

2. 본 문서의 활용 방법

감리인은 현행 감리 점검항목으로 감리를 수행하되, 사업에 생성형 AI가 적용된 부분에 한하여 본 문서의 점검항목을 함께 참고한다. 점검 결과는 별도의 보고서 양식을 신설하지 않고 현행 감리 보고서 체계에 반영한다.

사업 특성에 따라 어떤 점검항목을 적용할 것인지는 부록 D의 조정 매트릭스를, 현행 감리 점검항목과의 관계는 부록 E의 연계표를 참고한다.

3. 판본 관리

본 문서는 Ver 1.0이다. 인공지능 관련 법령과 지침은 제정·개정이 계속되고 있으므로, 본 문서의 인용 근거는 발간 시점을 기준으로 한다. 각 점검항목의 「근거」 항목과 부록 B.5의 근거–점검항목 대조표는 근거가 개정되었을 때 갱신하여야 할 점검항목을 특정하기 위한 것이다.

아래 사항은 본 판본의 범위에 포함되지 않으며, 차기 판본의 과제로 둔다.

차기 판본 과제	사 유
AI 관련 법령의 시행령·시행규칙 단위 적용요건 정리	법령 및 하위법령의 제정·개정이 진행 중이어서 본 판본의 작성 시점에 확정할 수 없음
공식 가이드라인의 추가 분석 및 점검항목 반영	관련 가이드라인이 계속 발간되고 있어 본 판본에서는 근거·출처 매핑까지만 수행
가이드 개정 절차 및 갱신 주기의 제도화	개정 체계의 운영은 발간 주체의 정책 사항이므로 본 판본에서는 체계만 제시
AI 개발방법론 및 산출물 표준의 제시	방법론·산출물 표준은 현행 공공 감리 가이드에서도 다루지 않는 영역

※ 본 문서의 내용에 대한 문의와 개선 의견은 발간 주체에 전달하여 차기 판본에 반영한다.

Contents

1 개요 ---------- 8

1.1. 배경 및 목적 8
1.2. 적용 범위 및 대상 10
1.3. 가이드 구성 원칙 11
1.4. 감리 점검 프레임워크 12
1.5. 가이드 유지·개정 체계 13
1.6. 참고문헌 14

2 가이드의 구성 ---------- 18

2.1. 가이드 구성 체계 18
2.2. 가이드 구성 목록 18
2.3. 감리 점검 가이드 목차 구성 20
2.4. 감리 점검 가이드 코드 체계 20

3 감리 점검항목 ---------- 24

3.1. 착수단계 24
가. 사업관리 24
3.2. 요구정의단계 26
가. 응용시스템 26
나. 데이터베이스 29
다. 시스템구조·보안 32
라. 품질보증 37
3.3. 설계단계 39
가. 응용시스템 39
나. 데이터베이스 43
다. 시스템구조·보안 45
3.4. 구현단계 50
가. 응용시스템 50
나. 데이터베이스 51
다. 시스템구조·보안 53
3.5. 시험단계 55
가. 응용시스템 55
나. 데이터베이스 57
다. 시스템구조·보안 61
라. 품질보증 65
3.6. 전개단계 66
가. 시스템구조·보안 66
3.7. 실행통제단계 68
가. 사업관리 68
3.8. 종료단계 69
가. 사업관리 69

4 개별 감리 점검 가이드 ---------- 74

4.1. 사업관리 75
- 착수단계 (10건) 75
- 실행통제단계 (6건) 75
- 종료단계 (4건) 75

4.2. 품질보증 137
- 요구정의·분석 (5건) 137
- 시험 (2건) 137

4.3. 응용시스템 159
- 요구정의·분석 (15건) 159
- 설계 (16건) 159
- 구현 (6건) 160
- 시험 (12건) 160

4.4. 데이터베이스 302
- 요구정의·분석 (11건) 302
- 설계 (11건) 302
- 구현 (10건) 303
- 시험 (13건 303

4.5. 시스템구조·보안 440
- 요구정의·분석 (24건) 440
- 설계 (19건) 441
- 구현 (5건) 441
- 시험 (21건) 442
- 전개 (5건)

부록 ---------- 660

부록 A. 전체 가이드 목록표 (195건) 661

부록 B. 인용 법령·표준 현황 672
- B.1 감리 수행의 근거 672
- B.2 AI 관련 법령 672
- B.3 공공 AI 관련 가이드라인·안내서 673
- B.4 국내외 표준·프레임워크 674
- B.5 근거–점검항목 대조표 674

부록 C. 약어·용어집 678
- C.1 약어표 678
- C.2 용어표 682
- C.3 카테고리 정의표 688

부록 D. 사업특성별 점검항목 조정 매트릭스 691
- D.1 조정 절차 691
- D.2 조정 매트릭스 서식 692
- D.3 작성 예시 693
  - 예시 1. 대민 민원상담 챗봇 693
  - 예시 2. 내부 문서검색 서비스 695

부록 E. 현행 감리 점검항목 연계표 (97건) 696

부록 F. 기술 유형별 커버리지 현황 703
- F.1 반영 수준 구분 703
- F.2 기술 유형별 커버리지 703
- F.3 미반영 기술이 포함된 사업의 대응 706

CHAPTER

1

개요

1. 개요

본 가이드는 생성형 AI 기술이 반영된 정보화사업의 감리 활동을 단계별·영역별 점검 가이드 195건으로 정리한 실무지침서이다. 본 장에서는 가이드의 개발 배경과 목적, 적용 범위, 구성 원칙, 감리 점검 프레임워크, 가이드의 유지·개정 체계 및 참고문헌을 소개한다.

1.1. 배경 및 목적

최근 AI 기술 발달에 따라 공공·민간 부문에서 생성형 AI 기반 정보화사업이 활발하게 추진되고 있으며, 국내외 연구기관과 유관기관에서는 AI 관련 다양한 특성(기술, 위험, 통제, 거버넌스 등)에 따른 개발 및 접근 방안을 안내서·점검표 등의 형식으로 활발하게 발표·발간하고 있다. 반면 공공 부문의 기존 감리지침은 AI(특히 생성형 AI)의 기술적 특성을 충분히 감안하지 못하고 있다.

이에 우리 협회에서는 2025년에 'AI(인공지능) 분야 정보시스템 감리 가이드' 연구용역을 통하여 AI 감리 표준화의 필요성과 국제표준 정합성, 생명주기 기반 감리절차, 점검항목 개발 방향과 필요성을 우선 제시한 바 있으며, 본 가이드는 이에 대한 후속 작업으로 실제 감리 현장에서 활용할 수 있는 점검체계를 마련한 것이다.

▣ 공공부문 AI 도입 현황

공공부문의 AI 도입은 이미 시범 도입의 단계를 지났다. 소프트웨어정책연구소가 조달청 조달정보개방포털의 조달계약 자료를 분석한 「2025년 공공부문 AI 도입현황 연구」(2026.4)에 따르면, 최근 10년간 공공부문의 AI 도입 계약은 6,975건에 이르고 412개 공공기관 가운데 268개 기관(65.0%)이 AI를 도입한 것으로 나타났다.

그러나 확산의 속도와 감리 준비의 속도는 같지 않다. 같은 자료에서 생성형 AI 도입 계약은 2023년 24건에서 2024년 42건으로 늘었을 뿐이어서, 감리 현장이 생성형 AI 사업을 충분히 경험하고 점검 방법을 축적할 시간은 아직 확보되지 못하였다. 반면 2026년도 정부 AI 예산은 2025년 본예산의 3배 수준인 9.9조원으로 편성되어, 생성형 AI가 반영된 사업의 발주는 단기간에 크게 늘어날 것으로 예상된다. 본 가이드는 이러한 시차를 메우기 위한 것이다.

▼ 공공부문 AI 도입 현황 (공표 자료 인용)

지 표	수 치	출 처
누적 AI 도입 계약 건수 (2015~2024년)	6,975건	ⓐ
AI 도입 공공기관	412개 기관 중 268개 (65.0%)	ⓐ
〃 **(직전 조사)**	401개 기관 중 243개 (60.6%)	ⓑ
연간 계약 건수	221건(2015년) → 1,215건(2024년), 약 5.5배	ⓐ

지 표	수 치	출 처
연간 계약 금액	2,443억원(2015년) → 2조 8,207억원(2024년), 약 11.5배	ⓐ
생성형 AI 도입 계약 건수	24건(2023년) → 42건(2024년), 75.0% 증가	ⓐ
2026년도 정부 AI 예산	9.9조원 (2025년 본예산 대비 약 3배)	ⓒ

ⓐ 소프트웨어정책연구소「2025년 공공부문 AI 도입현황 연구」(2026.4) ⓑ 소프트웨어정책연구소「2024년 공공부문 AI 도입현황 연구」(2025.4) ⓒ 국가인공지능전략위원회「AI 예산사업 통합 설명자료」(2026.3)

※ 위 수치는 본 협회가 자체 조사한 결과가 아니라 공표된 연구·발표 자료를 인용한 것이다. ⓐ·ⓑ는 조달청 조달정보개방포털의 조달계약 자료를 분석한 것이므로, 설문 방식의 실태조사 결과와는 측정 방법이 달라 직접 비교할 수 없다.

제도 환경도 함께 바뀌었다.「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」이 2026년 1월 22일 시행되어 국가기관등의 의사결정에 활용되는 AI가 고영향 인공지능의 영역으로 명시되었고,「인공지능 및 데이터 기반 행정 활성화에 관한 법률」이 2026년 8월 28일 시행을 앞두고 있다. 행정안전부는 2026년 6월「공공부문 AI 도입·활용 가이드」를 배포하여 기획부터 운영까지의 절차를 표준화하였다. 발주기관에 새로운 의무가 부과된 만큼, 감리도 그 준수 여부를 확인할 수 있는 점검 기준을 갖추어야 한다.

▣ AI 사업 특유의 위험과 감리 접근방안

AI 사업의 감리가 어려운 이유는 점검할 것이 많아서가 아니라, 현행 감리가 전제하고 있는 검증 방식이 그대로 성립하지 않는 경우가 있기 때문이다. 아래는 그러한 위험요인을 여섯 가지로 정리하고, 현행 감리 방식과의 차이 및 이를 대신할 점검 방법을 제시한 것이다. 표의「관련 점검항목」은 제4장 개별 가이드의 관리코드이다.

위험요인	현행 감리 방식과의 차이	감리 접근방안	관련 점검항목
학습데이터 원본의 검증 한계	현행 감리는 산출물과 원자료를 대조하여 확인한다. 그러나 AI 학습데이터는 사업자의 자산이거나 외부에서 수집된 것이어서 원본을 전수 대조하기 어렵다.	데이터 자체가 아니라 수집·정제·라벨링 과정의 기록을 증적으로 삼는다. 출처와 라이선스, 계보, 교차검수 이력이 남아 있는지를 확인한다.	AG_PM04-1-10 AG_PM04-2-04 AG_SD41-3-03 AG_SD41-3-09
외부 기반 모델의 내부 검증 불가	상용·공개 파운데이션 모델은 내부 구조와 학습데이터를 발주기관이 확인할 수 없다. 소스코드·설계서를 열람하여 검증하는 방식이 성립하지 않는다.	모델 제공자가 공표한 자료(모델 명세·이용약관·서비스 수준)와 발주기관이 자체 평가셋으로 측정한 결과를 함께 확인하고, 공급망 구성요소를 목록으로 관리하게 한다.	AG_SD41-1-15 AG_SD41-2-02 AG_SD44-1-13
성능의 비결정성	현행 시험은 동일한 입력에 동일한 출력이 나오는 것을 전제로 합격·불합격을 판정한다. 생성형 AI는 같은 입력에도 출력이 달라진다.	개별 실행의 정오가 아니라, 사전에 합의된 평가셋에 대한 정량 지표의 달성 여부로 판정한다. 목표값은 요구정의 단계에서 확정되어 있어야 하며, 확정되지 않았다면 그 자체가 지적사항이다.	AG_SD41-1-03 AG_SD44-1-02 AG_SD44-4-02

위험요인	현행 감리 방식과의 차이	감리 접근방안	관련 점검항목
환각·오답의 상시성	현행 감리에서 결함은 시정하여 제거하는 것이 원칙이다. 그러나 생성형 AI의 환각은 완전한 제거가 불가능하며 잔여위험으로 관리하여야 한다.	제거 여부가 아니라 억제 설계와 사후 통제의 존재 여부를 점검한다. 근거 문서 기반 응답, 출처 표시, 신뢰도 노출, 인간 개입 및 안전모드 절차가 설계에 반영되었는지를 확인한다.	AG_SD41-2-05 AG_SD42-2-04 AG_SD44-2-03
전개 이후의 성능 변화	현행 감리는 전개 시점의 상태를 기준으로 종료된다. 그러나 AI 모델은 입력 데이터의 분포가 달라지면 운영 중에 성능이 저하된다.	감리 범위는 구축 단계이나, 성능 저하를 감지할 기준선과 재학습 판단 기준, 모니터링 체계가 전개 이전에 설계되어 있는지를 확인한다.	AG_SD41-1-23 AG_SD44-1-16 AG_SD45-1-03
규제 환경의 유동성	현행 감리의 준거 법령은 대체로 안정적이다. 그러나 AI 관련 법령은 제정·시행이 진행 중이고 하위법령이 계속 정비되고 있어, 점검 시점에 따라 준거가 달라진다.	점검 시점의 시행 여부를 확인하고, 강제 의무와 노력 의무를 구분하여 지적 수준을 정한다. 고영향 AI 해당 여부에 대한 사전 검토 기록이 남아 있는지를 확인한다.	AG_PM04-1-03 AG_PM04-1-05 AG_PM04-3-02

※ 위 표는 감리 접근의 방향을 제시한 것이며, 단계별·영역별로 무엇을 어떻게 확인할 것인지는 제3장 점검항목과 제4장 개별 가이드에 담았다. 사업 특성에 따른 점검항목의 취사선택 기준은 부록 D를, 현행 감리 점검항목과의 연계 관계는 부록 E를 함께 참조한다.

1.2. 적용 범위 및 대상

본 가이드는 생성형 AI 기술이 적용되는 공공·민간 부문 정보화사업의 감리에 적용한다. 주된 사용자는 감리법인과 감리원이며, 발주기관과 사업자도 사업 준비·수행 과정에서 점검 기준을 사전에 확인하는 용도로 활용할 수 있다.

본 가이드는 기존 정보시스템 감리기준과 감리지침을 대체하는 것이 아니라, 생성형 AI 특유의 위험과 점검 관점을 보완하는 실무 참고 지침으로 활용한다. 개별 사업의 특성에 따라 점검항목을 취사선택하거나 조정하여 적용할 수 있다.

본 가이드가 다루는 범위는 다음 세 가지 축으로 한정된다. 축을 벗어나는 사업에도 참고 적용할 수 있으나, 점검항목의 산출물 명칭과 점검 시점은 해당 사업의 특성에 맞추어 대응시켜야 한다.

❑ 정보화 생애단계: 정보화사업의 기획(ISP) - 구축 - 운영·유지보수 중 구축 단계를 대상으로 한다. 기획 단계의 사업 타당성 검토와 운영 단계의 서비스 관리 활동은 본 가이드의 점검항목에 포함하지 않는다.

❑ AI 기술유형: 인공지능 기술 중 생성형 AI, 특히 대규모 언어모델(LLM)을 기반으로 구축되는 시스템을 대상으로 한다. 컴퓨터 비전·예측 분석 등 비생성형 AI가 적용된 사업에는 해당되는 점검항목을 선별하여 참고 적용한다.

❑ 개발방법론: 정보공학 기반의 단계별 개발방법론(SDLC)을 준거로 점검항목을 배치하였다. 애자일 등 다른 방법론을 적용하는 사업은 산출물의 명칭과 작성 시점을 해당 방법론의 것으로 대응시켜 적용한다.

▣ 본 가이드의 위상

본 가이드는 「정보시스템 감리기준」에 따라 감리를 수행하는 데 필요한 공식적인 감리 수행가이드가 아니다. 감리의 절차와 보고서 양식, 점검결과의 판정 체계는 현행 감리기준과 감리지침을 그대로 따르며, 본 가이드는 그 위에서 생성형 AI 특유의 점검 관점을 정리한 연구용역 결과물로서 감리 현장의 참고용 자료에 해당한다. 따라서 본 가이드의 점검항목은 감리인이 반드시 점검하여야 할 의무 사항이 아니다.

공공 부문의 AI 구축사업은 대부분 기존 정보시스템 구축사업에 AI 기능이 결합된 형태로 발주된다. 따라서 감리원은 현행 감리 점검항목을 기본으로 삼고, 본 가이드의 점검항목을 사업 특성에 맞게 취사선택·조정하여 병행 적용한다. AI 기능이 사업의 일부에만 적용되는 복합 사업에서는 해당 기능이 관여하는 영역·단계의 점검항목만 선별하여 적용할 수 있다.

점검 결과는 별도의 보고서 양식을 신설하지 않고 현행 감리보고서 체계에 반영한다. 본 가이드가 점검항목별로 제시한 판정기준은 현행 감리기준의 판정 체계를 준용하여 작성한 참고 기준이며, 보고서의 지적유형과 판정은 현행 감리기준이 정한 바에 따른다.

※ 본 가이드의 성격과 판본, 차기 판본의 과제는 책머리의 「일러두기」에 정리하였다.

1.3. 가이드 구성 원칙

생성형 AI 사업의 '생명주기'와 '기술영역'은 전통적인 '생명주기' 또는 '기술영역'과 다르게 구성되는 특성을 가진다. 그럼에도 본 가이드는 감리 현장의 연속성과 실무 적용성을 확보하기 위하여 기존 정보시스템 감리지침의 단계·영역 체계를 준용하되, AI 특유의 생명주기와 기술영역을 매핑할 수 있도록 아래와 같은 원칙을 정하여 마련·적용하였다.

❑ 단계 준용 원칙: 기존 정보시스템 감리지침의 단계(분석-설계-구현-시험-전개/운영준비)를 준용하되, 생성형 AI 특유의 생명주기(생명주기 관리 → 데이터 수집 및 처리 → 인공지능 모델 개발 → 시스템 구현 → 운영 및 모니터링)와 매핑할 수 있도록 접근한다.

❑ 영역 준용 원칙: 기존 정보시스템 감리지침의 4대 영역(시스템구조, 응용시스템, 데이터베이스, 사업관리 및 품질보증활동)을 준용하되, AI 특유의 기술적 영역(AI Full Stack 등)을 매핑할 수 있도록 접근한다.

▼ 감리지침 준용 개념도 (단계·영역)



1.4. 감리 점검 프레임워크

본 가이드는 생성형 AI 정보화사업의 감리 점검 가이드를 체계적으로 도출·배치하기 위하여 점검의 축을 세 가지로 정의하고, 이를 'AI 감리 점검 프레임워크'라 한다.

❑ ① 감리영역: 기존 정보시스템 감리지침의 4대 영역을 준용하되, 본문 편성 시 사업관리와 품질보증을 구분하여 5개 영역(사업관리, 품질보증, 응용시스템, 데이터베이스, 시스템구조·보안)으로 운영하며, 각 영역은 AI Full Stack 기술영역에 매핑된다(1.3절 영역 준용 원칙).

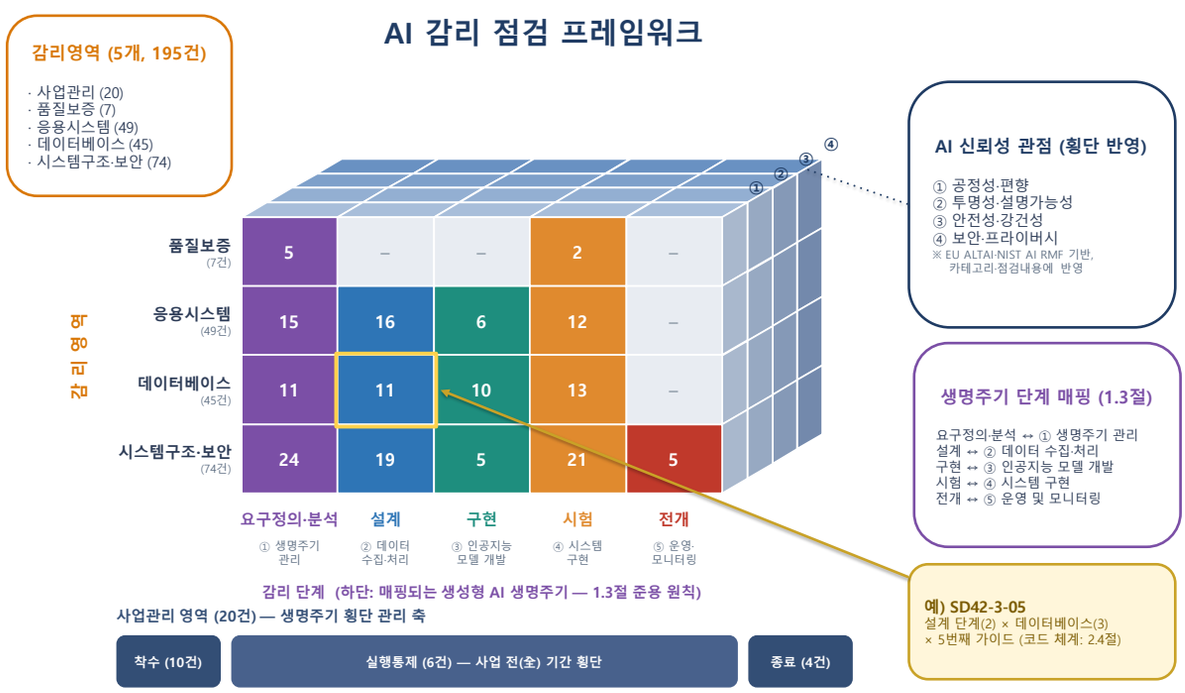
❑ ② 감리 단계: 기존 감리지침의 단계(요구정의·분석-설계-구현-시험-전개)를 준용하며, 각 단계는 1.3절 단계 준용 원칙에 따라 생성형 AI 생명주기(생명주기 관리 → 데이터 수집 및 처리 → 인공지능 모델 개발 → 시스템 구현 → 운영 및 모니터링)와 매핑된다. 사업관리 영역의 착수·실행통제

·종료 단계는 특정 생명주기 단계에 귀속되지 않고 사업 전(全) 기간을 횡단하는 관리 축으로 프레임워크에 포함한다.

❑ ③ AI 신뢰성 관점: 공정성·편향, 투명성·설명가능성, 안전성·강건성, 보안·프라이버시의 4개 관점으로 정의하며, EU 신뢰할 수 있는 AI 자체평가목록(ALTAI), NIST AI 위험관리 프레임워크(AI RMF) 등 1.6절의 국내·외 표준에서 도출하였다.

195개 점검 가이드는 감리영역×감리 단계의 셀에 배치되며(영역·단계별 분포는 2.1절 참조), AI 신뢰성 관점은 별도의 편성 축으로 사용하지 않고 각 점검 가이드의 카테고리와 점검 내용에 횡단적으로 반영한다(가이드별 주된 신뢰성 관점은 부록 A 목록표 참조). 개별 가이드의 구성 체계는 2.3절을 따른다. 이러한 3차원 점검 체계를 아래 그림과 같이 'AI 감리 점검 프레임워크'로 정의한다.

▼ AI 감리 점검 프레임워크



감리원은 감리계획 수립 시 본 프레임워크를 기준틀로 하여 사업의 특성(적용 AI 기술, 데이터 민감도, 서비스 위험도 등)에 따라 점검 대상 셀과 점검항목을 선정·조정할 수 있다(1.2절).

1.5. 가이드 유지·개정 체계

본 가이드는 인공지능 관련 법령과 행정 지침, 국내외 표준의 변화에 따라 지속적으로 갱신되어야 한다. 갱신이 일회성으로 끝나지 않도록 개정의 주기와 착수 기준, 갱신 대상을 특정하는 절차를 다음과 같이 정한다.

❑ 개정 주기: 연 1회 정기 개정을 원칙으로 하되, 아래 개정 착수 기준에 해당하는 사유가 발생한 경우에는 정기 개정 시점과 무관하게 해당 부분을 우선 갱신한다.

❑ 모니터링 대상: 부록 B에 등재한 법령·가이드라인·표준의 제·개정 여부를 상시 확인한다. 특히 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 및 그 시행령, 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」, 「공공부문 AI 도입·활용 가이드」는 공공 AI 사업 감리에 직접 영향을 미치므로 우선 확인 대상으로 한다.

❑ 개정 착수 기준: ① 부록 B 등재 법령·지침의 제정·개정 또는 시행, ② 국내외 표준·프레임워크의 제정·개정, ③ 본 가이드가 다루지 않는 AI 기술유형이 공공사업에 확산되는 경우, ④ 감리 현장에서 동일한 점검항목에 대한 보완 요구가 반복하여 제기되는 경우.

❑ 갱신 절차: 개정 대상 근거가 확인되면 부록 B.5 근거-점검항목 대조표에서 해당 근거를 인용하고 있는 점검항목을 특정하고, 그 점검항목의 「근거」와 검토방법·점검결과 판정기준을 함께 갱신한다. 갱신 후에는 부록 B의 현황과 대조표를 다시 일치시킨다.

❑ 이력 관리: 개정 시에는 개정 사유, 갱신한 점검항목의 관리코드, 근거의 변경 내역을 기록하여 다음 개정의 판단 자료로 활용한다.

1.6. 참고문헌

본 절은 가이드 개발 과정에서 참고한 표준·문헌을 정리한 것이며, 본문에서 직접 인용한 법령·표준의 현행성 정보는 부록 B를 참조한다.

❑ 국내·외 표준 참고

- ISO/IEC 42001(인공지능 경영시스템(AIMS) 요구사항)

- OECD AI 시스템 분류 프레임워크

- 인공지능에 대한 이사회 권고(OECD, 2019)

- Ethics guidelines for trustworthy AI(EU, 2019)

- EU ALTAI(The Assessment List for Trustworthy Artificial Intelligence) 자체점검 툴

- IIA(Institute of Internal Auditors) AI auditing framework

- ISACA(Information Systems Audit and Control Association) 인공지능 감사

- NIST 인공지능 위험관리 프레임워크(AI RMF 1.0)

❑ 선행 연구개발 문헌 참고

- 생성형 인공지능 정보시스템 데이터 품질 감리 점검

- ATL 10: 인공지능 기술 수준 정의(ETRI, 2020)

- 생성형 AI 윤리 가이드북(NIA, 2023)

- AI 기술을 적용한 정보시스템의 감리 점검방안(한국IT정책경영학회, 2019)

- 인공지능 신뢰 체계 정립 방안 연구(소프트웨어정책연구소(SPRI), 2022)

- 인공지능 윤리기준 실천을 위한 자율점검표(과기정통부·KISDI, 2024)

- 인공지능 학습용 데이터 품질관리 가이드라인 v2.0(NIA, 2022)

- 신뢰할 수 있는 인공지능 개발안내서(과기정통부·TTA, 2024)

CHAPTER

2

가이드의 구성

2. 가이드의 구성

2.1. 가이드 구성 체계

본 가이드는 영역 준용·단계 준용 원칙에 따라 195개 점검 가이드를 감리영역과 단계(요구정의·분석·설계·구현·시험·전개, 사업관리의 착수·실행통제·종료 포함)의 두 축으로 구성하였다. 아래 표는 각 감리영역이 어느 단계에 몇 건의 점검 가이드를 두는지를 교차 집계한 것으로, 감리원이 담당 영역·단계별 점검 범위와 전체 분포를 한눈에 파악할 수 있다.

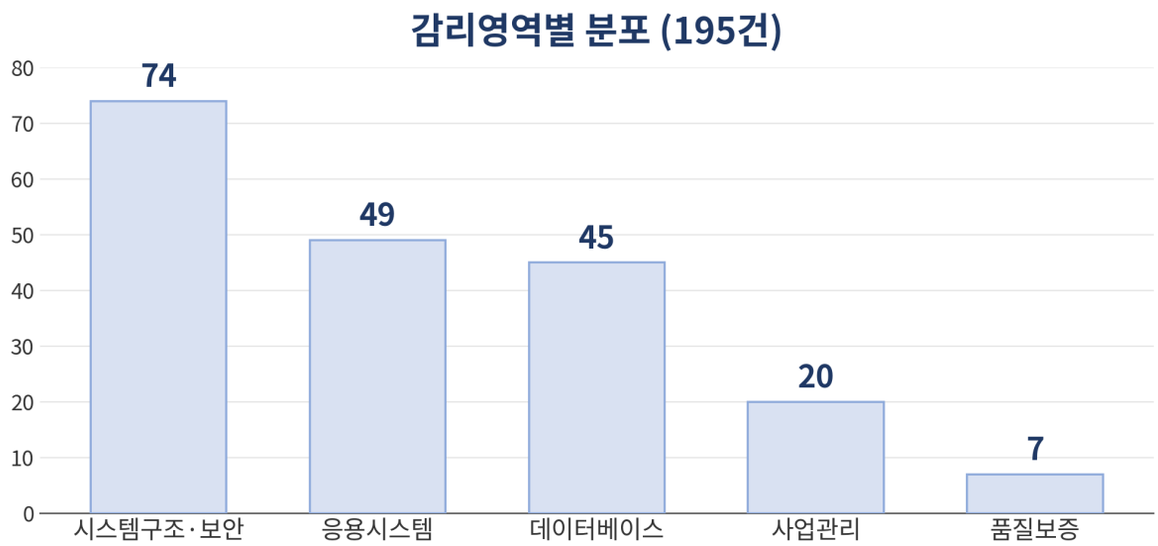
▼ 감리영역 × 단계 점검 가이드 분포 (195건)

단계 / 감리영역	시스템 구축					사업관리			총합계
	요구정의	설계	구현	시험	전개	착수	실행통제	종료	
사업관리						10	6	4	20
품질보증	5			2					7
응용시스템	15	16	6	12					49
데이터베이스	11	11	10	13					45
시스템구조·보안	24	19	5	21	5				74
총합계	55	46	21	48	5	10	6	4	195

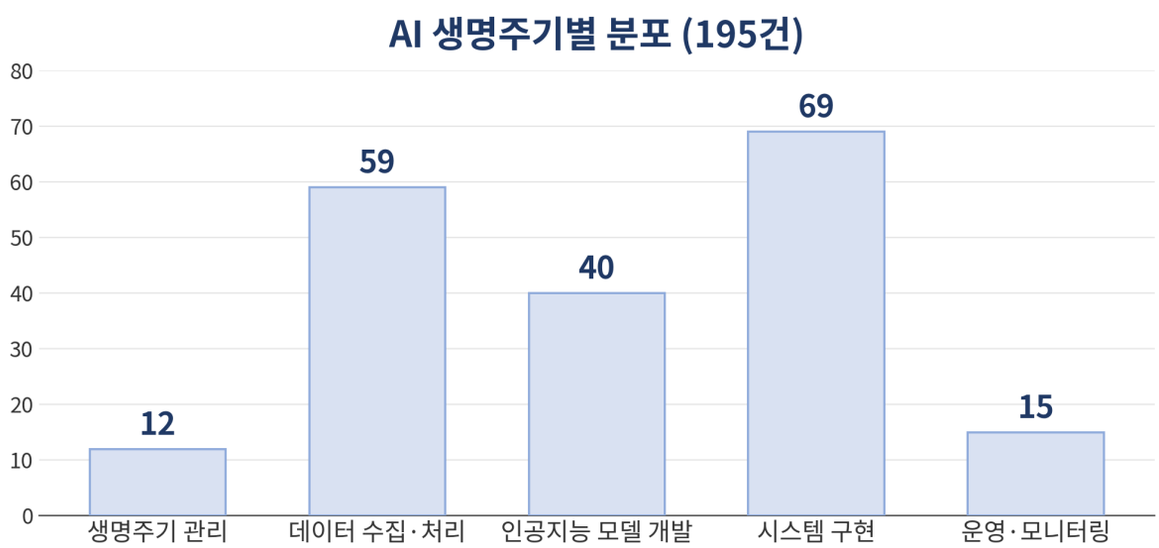
2.2. 가이드 구성 목록

본 가이드의 195개 점검 가이드는 전통적 4대 감리영역(시스템구조·보안, 응용시스템, 데이터베이스, 사업관리·품질보증) 전반과 AI 생명주기 전(全) 단계에 걸쳐 분포한다. 특히 상당수 항목이 환각·편향·프롬프트 인젝션·모델 공격 등 생성형 AI 고유의 위험을 직접 점검하여, 전통 정보시스템 감리로는 다루기 어려운 AI 특유의 신뢰성·안전성 요건을 포괄한다.

▼ 감리영역별 분포 (195건)



▼ AI 생명주기별 분포 (195건)

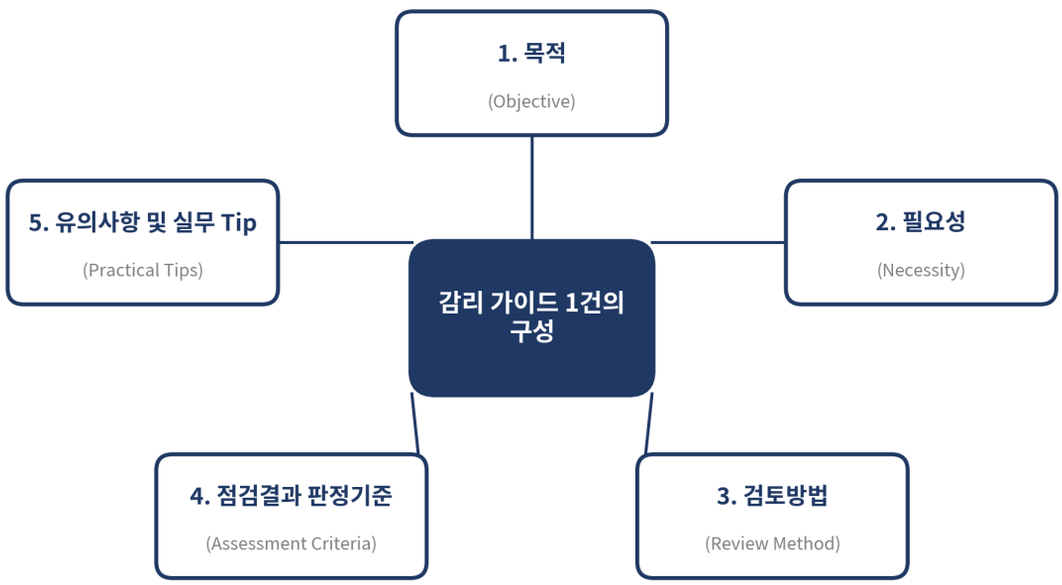


감리영역별 주요 점검 내용은 다음과 같다.

감리영역	건수	주요 점검 내용
사업관리	20	이해관계자·위험관리, AI 리스크 분류, 라벨링 도급·하도급 보안, 인수
품질보증	7	데이터 품질·편향 정량지표, AI 신뢰성 테스트·자동화 평가
응용시스템	49	XAI·환각·프롬프트 방어, AI 결정 오버라이드, 신뢰점수 노출
데이터베이스	45	데이터 품질·비식별화·계보, 벡터 DB·RAG, 학습·검증 데이터 분할
시스템구조·보안	74	망분리·DDoS/WAF, 모델 저장소·컨테이너 보안, 성능·이중화
합계	**195**	**— (전체 상세 목록은 부록 A 참조)**

2.3. 감리 점검 가이드 목차 구성

각 감리 점검 가이드의 구성 목차는 기존 감리지침의 구성과 유사한 구조를 유지하기 위하여 아래와 같은 구조로 구성된다. (☞ 목적 - 필요성 - 검토방법 - 점검결과 판정기준 - 감리 시 유의사항 및 실무 Tip)



2.4. 감리 점검 가이드 코드 체계

본 가이드의 195개 감리 점검 가이드에는 기존 정보시스템 감리 점검 체계의 시스템 개발(SD) 유형 구분을 확장한 AI 분야 점검 체계 코드(SD4)를 부여한다. 코드는 [SD4+단계]-[영역]-[순번]의 구조를 가지며, 사업관리 영역은 별도의 PM04 체계를 사용한다.

구성 요소	의미
SD4□	AI 분야 시스템 개발(SD) 점검 체계. 기존 감리 점검 체계의 SD1(구조적/정보공학)·SD2(객체지향/컴포넌트)·SD3(빅데이터)을 확장한 것으로, 넷째 자리(□)는 감리 단계를 나타낸다. (1 분석/요구정의 · 2 설계 · 3 구현 · 4 시험 · 5 전개)
-□ (영역)	감리영역 구분: 1 시스템구조·보안 · 2 응용시스템 · 3 데이터베이스 · 4 품질보증
-□□ (순번)	동일 단계·영역 내 점검항목 일련번호 (01부터 부여)
PM04-□-□□	사업관리(PM) 점검 체계. 첫째 자리(□)는 단계(1 착수 · 2 실행통제 · 3 종료), 이후 두 자리는 일련번호

예를 들어 AG_SD42-3-05는 설계 단계(2)·데이터베이스 영역(3)의 다섯 번째 감리 점검 가이드를, AG_PM04-1-03은 사업관리 착수단계의 세 번째 감리 점검 가이드를 의미한다.

CHAPTER

3

감리 점검항목

3. 감리 점검항목

본 장은 개별 감리 점검 가이드 195건을 감리 현장에서 활용하기 쉽도록 유사 검토항목을 기준으로 141개 점검항목으로 통합·재구성한 점검항목 체크리스트이다. 점검항목은 감리단계·감리영역별로 배열하고, 각 영역 안에서는 업무 진행(카테고리) 순서로 정렬하였다. 각 점검항목의 '관련 가이드' 열에는 해당 항목이 도출된 개별 감리 점검 가이드의 코드를 표기하였으며, 세부 검토 근거와 판정기준은 그 코드로 4장(개별 감리 점검 가이드) 및 부록 A에서 확인할 수 있다. 각 점검항목의 카테고리는 AI 신뢰성 관점(공정성·편향, 투명성·설명가능성, 안전성·강건성, 보안·프라이버시)과 AI 기술·관리 요소를 감안하여 구성하였으며, 가이드별 주된 신뢰성 관점은 부록 A 목록표에 표기하였다.

3.1. 착수단계

가. 사업관리

점검항목	검토항목	카테고리	관련 가이드
1. 사업목표 달성 지표(KPI) 적정성과 AI 시스템 도입 효과 지표 명확성 여부	- 측정 가능한 KPI와 AI 도입 효과 지표가 사업 목적과 정합되게 정의되었는가? • KPI 항목별 목표값·기준값·측정 주기·방법 명시 및 AI 특화 KPI와 일반 IT KPI 구분·수치 설정 • 도입 전후 정량 개선 목표와 KPI-사업 목적·Pain Point 간 인과 정합성 및 Baseline 근거	사업타당성/ROI	AG_PM04-1-01
2. 이해관계자별 요구사항의 체계적 식별· 정의 여부	- 사용자·개발자·운영자 등 이해관계자 요구사항이 체계적으로 식별·정의되었는가? • 사용자·개발자·운영자 등 이해관계자 요구사항 명시 및 AI 서비스 대상 사용자·취약계층 정의	의사소통	AG_PM04-1-02
3. AI 고유 위험 관리 프로세스 수립 및 AI 점검 제약 대체 검증 방법 수립 여부	- AI 점검 제약 및 대체 검증 방법이 명시되었는가? • AI 법적·기술적 제약 및 외부 기반모델 활용 제약 식별, 점검 불가 항목의 사전 협의·대체 검증 방법	AI 위험관리	AG_PM04-1-03

점검항목	검토항목	카테고리	관련 가이드
	- AI 고유 위험에 대한 관리 프로세스가 수립되었는가? • 환각·편향·개인정보·보안 등 AI 고유 위험의 식별 목록과 원인·결과 분석 구체화 • 위험 심각도·발생가능성 평가 기준 적정성 및 제거·완화·전가·수용 대응 조치와 잔여 위험 재평가 절차 - AI 시스템 위험 수준 분류 및 위험 등급별 관리·규제 요건이 식별·반영되었는가? • 리스크 분류 요구사항 명시 및 발주기관 내부 AI 윤리 정책·위험관리 가이드 • 위험 분류 매트릭스의 작성 시점·근거 자료 적정성		AG_PM04-1-04 AG_PM04-1-05
4. 데이터 라벨링 도급·하도급 관리 요구사항 명시 여부	- 데이터 라벨링 도급 관리와 하도급 작업자 보안 요건이 명시되었는가? • 데이터 라벨링 도급·하도급 요구사항 명시 및 대상 데이터의 개인정보·민감정보 포함 여부 사전 파악 • 하도급 작업자 신원 확인·접근 통제·보안서약 등 사전 통제 방안	조직체계	AG_PM04-1-06
5. 데이터 라벨링 품질을 확보하기 위한 가이드라인, 라벨러 교육 및 품질 검증 계획 적정성	- 라벨링 가이드라인 작성·일관성 규칙과 라벨러 교육·편향 방지 교육 계획이 반영되었는가? • 라벨링 가이드라인 작성 일정·담당 조직·교육 계획·품질 기준 • 학습 데이터 유형·라벨링 특성 및 파인튜닝·RLHF 등 생성형 AI 특화 라벨링 방식 파악 - 다수 검수자 교차 검수·검수자 간 일치율 기준·불일치 해소 절차가 반영되었는가? • 교차 검수 대상 범위·검수자 인원·자격과 검수자 간 일치율 측정 방법·목표 수치 정의 • 불일치 라벨 해소 절차·최종 라벨 결정 기준과 임계 초과 시 가이드라인 재검토 트리거	라벨링	AG_PM04-1-07 AG_PM04-1-08

점검항목	검토항목	카테고리	관련 가이드
	- 주관적 판단 배제 라벨링 기준·경계 사례 처리와 작업자 간 일치도 검증 요건이 정의되었는가? • 라벨 유형별 정의·판단 기준의 객관적·예시 기반 명세 및 경계 사례 처리 기준 사전 • 라벨러간 일치도 목표·측정 방법과 라벨 불일치 해소·편향 어노테이터 탐지·검수 요건		AG_PM04-1-09
6. 학습데이터 수집 시점 편향 배제 및 수집 후 편향 탐지·사후 보정 방안 정의 여부	- 학습데이터 수집 시 출처·인구통계·시점 다양성 확보와 편향 탐지·보정 요건이 수치로 정의되었는가? • 수집 출처 목록과 출처별 비율 상한·인구통계적 다양성 균형 기준 • 수집 시점 편향 배제와 수집 후 편향 탐지·사후 보정 방안	편향/공정성	AG_PM04-1-10

3.2. 요구정의단계

가. 응용시스템

점검항목	검토항목	카테고리	관련 가이드
1. 도메인 특화 편향 요인 사전 식별과 화면 요건에 편향 배제 및 UI/UX 중립성 요건 포함 여부	- AI 시스템 편향 발생 가능성 식별·정의 여부 및 편향 완화 계획 요건이 식별되었는가? • AI 편향 관련 요구사항 유무 및 AI 적용 업무 도메인의 편향 민감도 특성 파악 여부 • AI 편향 관리 활동 포함 여부 및 편향 식별 결과의 데이터·모델 요건 연계	편향/공정성	AG_SD41-2-01
2. 설명가능·인간감독 관련한 기능 도출 및 구체적 정의 여부	- AI 모델 작동 방식·성능 특성·한계점 명세 및 AI 시스템 안내·불확실성/오류 가능성 고지 화면 요건이 반영되었는가? • AI 모델 명세·사용자 안내 관련 요구사항 유무 및 AI 투명성 항목 포함 • 사용자 안내 요건 정의 현황	AI윤리/xAI	AG_SD41-2-02

점검항목	검토항목	카테고리	관련 가이드
	- AI 설명가능성 기법 적용과 추론 근거를 제공하는 대상·기능·방식이 정의되었는가? • XAI·설명가능성 관련 요구사항 유무와 구축 대상 AI 서비스의 위험 수준·사용자 유형 - 신뢰점수·확률·불확실성 정보를 사용자에게 제공하는 방식·임계값 요건이 정의되었는가? • AI 출력 신뢰점수·불확실성 관련 요구사항 유무 • 서비스 도메인 고위험 여부 파악을 통한 노출 요건 중요도 사전 평가 • AI 모델 유형과 발주기관의 불확실성 고지 정책		AG_SD41-2-03 AG_SD41-2-04
3. 환각 방지 기능·절차 요건 적정성	- 환각 방지를 위한 허위정보 필터링·출처 명시·불확실성 고지·HITL 기능·절차 요건이 정의되었는가? • AI 출력 정확성·신뢰성 관련 요구사항 유무와 허위정보 방지 관련 별도 요구 사항 면담 확인 여부 • 환각 방지 대응 방안 언급 여부	할루시네이션	AG_SD41-2-05
4. 부적절 입력 차단 방어 요건과 편향 탐지·차단 가드레일 정의 여부	- 프롬프트 인젝션·탈옥·금칙어 등 부적절한 입력 차단 요건이 명시되었는가? • 차단 대상 입력 유형·금칙어 목록 관리 절차와 프롬프트 인젝션 탐지 범위·탐지율·오탐률 기준·탈옥 대응 방식 • 신규 공격 패턴 대응 갱신 절차·오탐 처리·차단 이력 로깅 요건 - 편향 탐지·차단 동작·오탐 처리·갱신 절차·검사 기준이 정의되었는가? • 금칙어 사전 차단 범주·예외 처리 기준과 문맥 기반 편향 탐지·후처리·탐지율·오탐률 임계값 요건부 • 금칙어 갱신 주기·담당 주체·이력 관리·오탐 처리와 우회 시도 포함 검사 시나리오·모니터링 요건	금칙어/가드레일	AG_SD41-2-06 AG_SD41-2-07

점검항목	검토항목	카테고리	관련 가이드
5. 필수 설명의무, 금지표현 차단 등 컴플라이언스 가드레일 요건 정의 적정성	- 설명의무·부당권유 금지 등 컴플라이언스 통제 요건이 정의되었는가? • 필수 설명의무 정량 검증 및 투자성향 적합성 Rule 기반 판정 기준·등급체계 정의 • (금융) 원금보장·확정수익 등 금지표현·오인표현 차단	컴플라이언스/가드레일	AG_SD41-2-08
6. 윤리적 우려사항 보고 채널 기능 요건 정의 적정성	- 윤리적 우려사항(차별·편견·불공정 출력 등)을 이용자가 직접 신고·보고할 수 있는 채널에 관한 요구사항이 도출되었는가? • 신고 채널 유형·익명 신고 가능 여부와 처리 절차·담당자·응답 기한 • 신고 내용 기록·보존·이력 관리와 검수 측정 지표·UI 반영 요건	법령/AI윤리	AG_SD41-2-09
7. 사용자가 AI와 상호작용 중임을 고지, 웹 접근성 요건, 안내 메시지 제공 요건 분석 적정성	- 장애인·노약자·외국어 사용자 대상 접근성과 다국어 지원 요건이 검증 방법과 함께 반영되었는가? • KWCAG 2.2 이상 준수·스크린리더 호환성과 TTS·STT·고대비·글자 크기 조정·키보드 접근성 등 대체 입출력 수단 요건 - AI 특화 오류 유형별 예외 안내 메시지 표출 요건이 정의되었는가? • 오류 유형 (일반·AI 특화) 분류·귀책 주체별 안내 차별화와 메시지 구성 요소·표시 방식	UI/UX, 접근성	AG_SD41-2-10 AG_SD41-2-11
8. 사용자 식별·프롬프트·접속 로깅 요건이 개인정보 처리기준에 부합 여부	- 식별된 개인정보에 대해서 마스킹·보관 기간·접근 통제·무결성·오남용 탐지 요건이 정의되었는가? • 사용자 식별 항목의 로그인·비로그인 구분과 프롬프트·응답 로깅 방식·요청 식별자 연계·개인정보 마스킹 방식 • 법적 기준 기반 보관 기간·역할별 접근 통제·자동 파기·로그 무결성·오남용 탐지·조회 기능 요건	개인정보보호	AG_SD41-2-12

점검항목	검토항목	카테고리	관련 가이드
9. 추론 실패 시 폴백 및 추론결과 신뢰도가 부족한 상황 발생 시 안전모드로 전환 요건 정의 적정성	- AI 오작동·장애 유형 분류와 유형별 안전모드로 전환·복구 절차 요건이 정의되었는가? • AI 특유 장애 유형 분류와 탐지 기준 및 유형별 디폴트 행동·자동·수동 전환 기준 • 장애 시 사용자 안내·대체 채널·단계별 확대와 이력 로깅·관리자 알림·복구 전환 기준·로그 보관 기간 요건 - 오류 발생 시 안전 모드 자동 전환 요건이 정의되었는가? • Fail-safe 요건 구체적 기준 설정과 외부 AI 모델 의존 시 장애 대체 처리 요건	가용성/이중화	AG_SD41-2-13 AG_SD41-2-14
10. 모델 평가 지표 및 운영 절차 수립 요건 정의와 사용자 데이터 수집·모니터링·재학습 요건 정의 여부	- 사용자 경험 로그 수집·분석 요건 및 사용자 행동 데이터 수집 요건이 정의되었는가? • 사용자 입력·응답·오류·재시도·평가 등 주요 로그 항목·멀티턴 대화 구조와 수집 목적·모델 개선 피드백 루프 설계 • 개인정보 비식별화·가명처리 요건과 사용자 동의·고지·거부권 보장 절차	모니터링/재학습	AG_SD41-2-15

나. 데이터베이스

점검항목	검토항목	카테고리	관련 가이드
1. 보호변수, 민감변수 및 특성변수 선정 기준·배제 관리 요건 적정성	- 특성 변수 선정 기준 및 관리 요건이 정의되었는가? • 후보 특성 변수 목록·선정 근거와 통계적 중요도 평가 방법·기준 수치 • 데이터 누수·보호 속성·프록시 변수 배제 요건과 임베딩 모델·차원·거리척도 선정 근거 문서화 요건	편향/공정성	AG_SD41-3-01

점검항목	검토항목	카테고리	관련 가이드
2. 학습데이터/메타데이터/공통코드 표준 요건 및 데이터 계보 추적 요건 정의 적정성	- AI 모델 입출력에 사용되는 공통 코드 설계 및 관리 요건이 정의되었는가? • AI 입출력 공통 코드 집합 목록·코드 값·의미·유효 기간 명세 완전성과 업무 시스템 연계 코드 일치·변환 매핑 요건 • 코드 추가·변경·폐기 관리 절차 및 요건 완전성 추적	거버넌스/데이터계보	AG_SD41-3-02
	- 데이터 계보 추적 요건이 정의되었는가? • 데이터 출처·수집 방법·이력·버전 관리 요건과 파이프라인 단계별 처리 기준·책임자·품질 임계값·추적 단위·메타데이터 정의 • 파이프라인 오류·변경 시 추적·임베딩 데이터 계보 및 요건 완전성 추적		AG_SD41-3-03
	- AI 학습데이터 메타데이터 명세 및 관리, 벡터 메타데이터 요건이 정의되었는가? • 메타데이터 필수 항목 정의·학습데이터-메타데이터 표준·형식 정의 및 관리 요건 • 벡터 데이터 메타데이터 확보 요건		AG_SD41-3-04
	- AI 감사 추적 데이터 항목·보존기간·저장구조가 정의되었는가? • AI 자동 결정의 영향 범위에 따른 감사 추적 요건 적용 강도 및 설명가능성 요구사항 반영 일관성 • 동일 입력에 대한 출력 재현을 위한 모델버전·입력·출력 보존 및 보존기간		AG_SD41-3-05
3. 개인식별정보 비식별화 요건 정의 여부	- 개인식별정보 가명·익명처리·마스킹 비식별화 요건이 법적 근거와 함께 정의되었는가? • 개인식별정보 대상 항목과 항목별 처리 법적 근거 • 가명·익명·마스킹 처리 방법·검증 절차와 임베딩 생성 이전 비식별화 수행·역추적 위험 평가 요건	비식별화	AG_SD41-3-06

점검항목	검토항목	카테고리	관련 가이드
4. 정제 전·후 데이터 명세와 마이그레이션 데이터 검증, 학습/검증/테스트 데이터셋 요건 정의 여부	- 데이터 마이그레이션 요건 및 레거시 데이터 이관·검증 요건이 정의되었는가? • 이관 대상·제외 범위 정의와 항목별 변환 규칙 매핑·이관 완전성·정합성 검증 기준 수치 정의 • 이관 실패 시 롤백·재이관 기준·재임베딩 검증 및 요건 완전성 추적 - 학습/검증/테스트 세트 분할 기준이 명시되었는가? • 세트 분할 비율·분할 방법 정의와 테스트 세트 독립성 보장·데이터 누수 방지·랜덤 시드 고정 요건 • RAG 검색 평가셋 분리 요건 및 요건 완전성 추적	정제(클렌징)/변환	AG_SD41-3-07 AG_SD41-3-08
5. 데이터 분포·이상값 탐색과 학습데이터 품질 특성별 측정 기준·임계값 정의 여부	- 학습데이터 품질 특성별 측정 기준·임계값이 정의되었는가? • 완전성·유효성·정확성·정합성·유일성·적시성 품질 특성별 기준의 수치·규칙과 벡터 검색 정확도 기준 • 품질 측정 도구·주기·이력 기록 방법과 임계값 미달 시 처리 절차	데이터 품질	AG_SD41-3-09
6. 벡터 DB 및 RAG 파이프라인 요구 정의 적정성	- 벡터 DB 제품 선정 기준·RAG 전략·데이터 규모 산정·운영 SLA·데이터 거버넌스 요건이 정의되었는가? • 벡터 DB 제품 선정 기준의 정량적 명시 및 RAG 데이터 규모·증가율 기반 운영 SLA 임계값 • 망분리 등 인프라 환경 제약 및 데이터 거버넌스·보안 요건	벡터 DB/임베딩	AG_SD41-3-10

점검항목	검토항목	카테고리	관련 가이드
7. 학습/테스트 데이터 분할 계획 및 검증용(Hold-out) 데이터 확보 계획 적정성 여부	- 학습·검증·테스트 데이터 분할 계획과 독립 검증용 데이터 격리·확보 계획이 수립되었는가? • 데이터 분할 방식·비율의 수치 명시와 Hold-out 데이터의 규모·확보 시점·격리 보관·접근 통제 • 데이터 오염 방지 절차, 분포 대표성, 분할 이력·재현성	테스트시나리오	AG_SD41-3-11

다. 시스템구조·보안

점검항목	검토항목	카테고리	관련 가이드
1. 프롬프트 및 모델 응답을 통한 개인정보 유출 방지 요건 적정성	- 프롬프트 입력 및 모델 응답을 통한 개인정보 유출 차단 요건이 적정하게 정의되었는가? • AI 데이터 흐름에 맞춘 DLP 연동 지점 및 개인정보 유형별 탐지 규칙·마스킹 정책·탐지 로그 관리 요건의 구체성 • 개인정보 보호 지침과 AI 서비스 적용 범위의 정합성	개인정보보호	AG_SD41-1-01
2. AI 서비스 시스템의 백업·재해복구 요건 정의 여부	- 백업·재해복구 체계 요건이 모델 가중치·하이퍼파라미터·프롬프트 설정 등 AI 고유 자산 백업 대상 포함과 함께 명시되었는가? • 모델 가중치·하이퍼파라미터·프롬프트 설정 등 AI 고유 자산 및 GPU/NPU 서버·벡터 DB 등 AI 전용 인프라를 반영한 백업·DR 대상 포함 여부 • 기존 백업·DR 정책의 AI 확장 적용 가능성	데이터 백업/복구	AG_SD41-1-02
3. AI 모델 성능·정확도 목표의 측정 가능한 요건 정의 여부	- AI 모델 성능·정확도 목표가 AI 모델 유형에 적합한 항목이 측정 가능하게 정의되었는가? • 모델 유형별 적합한 정량 지표·합격 임계값 수치와 평가 데이터셋 규모·대표성·편향 배제·Baseline • 환각 허용·탐지 기준·운영 환경 기준 성능 목표·추적성	모델평가	AG_SD41-1-03

점검항목	검토항목	카테고리	관련 가이드
4. 목표 AI 시스템 아키텍처에 요구사항 반영 완전성 및 AI 특화 계층 도식화 적정성	- AI 서비스 특유의 구성요소(모델 서빙 계층, 데이터 파이프라인 등)를 식별 가능한 수준으로 표현하였는가? • AI 모델 유형·학습/추론 환경 분리·클라우드 구분 등 구성 요구사항 반영 및 모델 서빙 계층·데이터 파이프라인·GPU 추론 인프라 등 AI 특화 계층의 누락 없는 도식화	아키텍처	AG_SD41-1-04
5. AI 서비스 장애·오류 발생 시 서비스를 안전하게 유지하거나 복구 요건 정의 적정성	- AI 장애·오류 시 안전모드 작동 인프라 요건 및 폴백 서버·모델 이중화 요건 (트리거 조건·전환 기준·복귀 절차)이 구체적으로 정의되었는가? • 안전모드 트리거 조건·해제 복귀 기준의 수치 명시 및 폴백 대상·우선순위·전환 임계값·대체 모델 최소 성능·헬스체크 요건 • AI 특화 장애 시나리오 식별 및 이중화 아키텍처와 요구사항 간 정합성	가용성/이중화	AG_SD41-1-05
6. AI 시스템의 도입·설치 환경을 구체적으로 분석 여부	- AI 시스템 도입·설치 환경 분석이 요구사항 대비 구체적으로 수행되었는가? • 서버 사양·GPU·네트워크 등 AI 인프라 요구사항 및 전산센터 여유 공간·전력·대역폭 등 기존 인프라 현황 대비 분석 적정성	서버스토리지/네트워크	AG_SD41-1-06
7. 성능 요구사항과 인프라 사이징 기준 연관성 유지 여부	- 성능 요구사항이 인프라 사이징에 적정하게 반영되었는가? • 목표 성능과 GPU 인스턴스 수량·VRAM 산정에 일관되게 적용되고, 추론 서버 Batch Size·큐 설계 • 부하 임계값에 기반한 오토스케일링 정책의 구체성	성능	AG_SD41-1-07
8. 컨테이너 이미지 보안 요건 및 클라우드 환경 AI 인프라 보안요건 정의 여부	- 컨테이너 기반 AI 인프라 보안요건 및 컨테이너 이미지 취약점 점검 요건이 정의되었는가? • 컨테이너 보안 요구사항 및 이미지 공급망 취약점·컨테이너 탈출·오케스트레이션 권한 오남용 등 위협의 요건 • 기존 컨테이너 보안 정책 및 시스템 배치 환경	클라우드컨테이너	AG_SD41-1-08

점검항목	검토항목	카테고리	관련 가이드
9. AI 특유의 파이프라인 보안 위험과 모델 배포 환경 보안 요건 정의	- AI 모델 CI/CD 파이프라인 보안 및 모델 자동 배포 환경 요건이 정의되었는가? • CI/CD 파이프라인·MLOps 보안 요구사항 정의 여부 확인과 발주기관의 소프트웨어 공급망 보안 정책·DevSecOps 지침·CI/CD 구축 계획 • AI 모델 배포 방식 및 배포 대상 환경 파악 여부	CI/CD파이프라인	AG_SD41-1-09
10. AI Agent가 접근하는 Legacy API가 업무 목적에 기반하도록 요건 정의 여부	- AI Agent가 접근하는 Legacy API가 업무 목적에 기반하여 구축되도록 요건이 정의되었는가? • 테이블·스키마 직접 노출 안티패턴 금지 및 업무 목적에 기반한 API 명세화와 보안요건 • 특정 실행 기능의 Draft→Human 승인→확정단계 분리	MCP/Tool연계	AG_SD41-1-10
11. 대량 추론 요청(Inference Request)과 외부 API 연계 트래픽에 대한 과부하 방어 및 접근통제 요구사항 정의 여부	- AI 추론 트래픽 특성을 반영한 DDoS 방어·API Rate Limit·Graceful Degradation·외부 API 폴백·로깅 요건이 정의되었는가? • AI 추론 엔드포인트 특성을 반영한 DDoS 방어 정량 임계값, 트래픽 유형별 차등 임계값 및 외부 LLM·검색 API Rate Limit·사용자·역할별 호출 한도 • 과부하 시 Graceful Degradation 전략, 외부 API 장애 시 폴백 방안, 보안 이벤트 로깅·알림 요건	API보안	AG_SD41-1-11
12. AI 특화 공격 유형별(적대적 예제·멤버십 추론 등) 방어 요건 정의 여부	- 적대적 예제·데이터 중독·모델 추출·멤버십 추론 등 적대적 공격 유형별 방어 요건이 정의되었는가? • 위험 등급 기반 방어 대상 공격 유형·수준 정의와 데이터 중독 방어·무결성 검증 및 적대적 예제 내성·입력 검증 요건 • 모델 추출 방어·멤버십 추론 방어와 임베딩·벡터 공격 방어 요건	AI특화보안	AG_SD41-1-12

점검항목	검토항목	카테고리	관련 가이드
	- 모델 공격 위협 식별/시나리오·정량적 방어 요건·입출력 검증 방법이 구체화되고 전문가 검토가 정의되었는가? • AI 모델 공격 유형과 생성형 AI 특화 위협 식별 및 위협 시나리오 구체성·방어 요건 정량화·검증 방법 • 입력값 검증·출력 필터링 요건과 위협 분석의 제3자·보안 전문가 검토 이력		AG_SD41-1-13
	- 모델 저장소의 접근통제·무결성 검증·배포 승인 절차가 정의되었는가? • 모델 저장소의 RBAC 기반 권한 정의·파일 무결성 검증 및 스테이징→프로덕션 다단계 배포 승인·버전 이력 추적 요건 • 주요 행위 감사 로그 및 외부 도입 모델 보안 검증 요건의 구체성		AG_SD41-1-14
	- AI 모델 API 환경의 비정상 API 호출·모델 추출 공격 탐지 요건이 AI 특화 위협으로 정의되었는가? • AI 모델의 API 서비스 제공 여부 및 대량 추론 요청을 통한 의사결정 경계 역추정·파라미터 복제 공격 탐지 요건 • AI 보안 위협 분석 활동 및 API 보안 가이드라인		AG_SD41-1-15
	- AI 공급망 보안 위험 관리 요건이 명시되었는가? • AI 시스템 아키텍처의 외부 의존성 식별 및 발주기관 보안 정책의 오픈소스·외부 모델 사용 제한 • 공급망 보안 요구사항 및 AI-BOM (AI Bill of Material)		AG_SD41-1-16

점검항목	검토항목	카테고리	관련 가이드
13. 생성형 AI 추론 API 접근통제가 제로트러스트 및 최소권한 원칙에 기반한 정의 여부	- 생성형 AI 추론 API 접근 통제·인증 체계가 정의되었는가? • AI API 보안 요구사항 유무의 예비조사 단계 선행 확인과 외부 클라우드 LLM 연계 또는 자체 모델 서빙 아키텍처 유형 • 발주기관 내부 API 보안 정책	인증/인가/접근통제	AG_SD41-1-17
14. AI API 토큰·키 관리 및 구간별 통신 암호화 요건 정의 여부	- AI 데이터·모델 파라미터 통신구간 암호화 요건이 구간별로 정의되었는가? • 사용자·추론서버·모델 저장소·외부 LLM 등 구간별 TLS 적용 범위·버전 정의 및 외부 API 보안 요구사항 명시 • 암호화 표준의 AI 시스템 적용 및 보안 검증 계획	암호화	AG_SD41-1-18
15. AI 특화 위협 대응 요건 및 방화벽·WAF 요건 및 오픈소스 보안 요건 정의 여부	- 오픈소스 라이브러리 보안요건 및 의존성 취약점 요건이 정의되었는가? • 허용 라이선스·출처 제한, 최소 버전·취약점 점검·CI/CD 자동 점검 • 오픈소스 사용 목록 완전성	시큐어코딩	AG_SD41-1-19
16. 망분리 환경에서 모델 훈련에 대한 보안요건 정의 여부	- 망분리 환경에서 AI 모델 훈련에 필요한 학습 데이터 외부유출 방지 등에 대한 보안요건이 정의되었는가? • 폐쇄망 학습 환경의 데이터 유출·비인가 접근 방지 보안요구사항 및 학습 데이터 무결성 검증·데이터 오염 대응 요건 • 망분리 정책에 부합하는 학습 환경 네트워크 구성 및 데이터 보안등급 적용 적정성	망분리/폐쇄망	AG_SD41-1-20
17. AI 서비스 신뢰성 기준 및 AI 시스템의 성능 저하 임계값이 구체적이고 측정 가능한 정의 여부	- 환각·편향·탈옥·프롬프트 인젝션 등 AI 신뢰성 특화 테스트 항목과 수치화된 합격 기준·일정이 반영되었는가? • 환각·편향·탈옥·프롬프트 인젝션 등 AI 특화 테스트 항목과 정확도·차단율 등 수치화된 합격 기준·일정 • 레드팀·제3자 독립 평가 계획과 테스트 데이터셋의 대표성·편향 방지·학습/테스트 분리 계획 포함 여부	AI QA체계	AG_SD41-1-21

점검항목	검토항목	카테고리	관련 가이드
	- 서비스 성능 저하 임계값이 측정 가능하도록 정량적으로 명시되었는가? • 응답시간 상한·정확도 최솟값·가용률·RTO/RPO 등의 수치 명시 및 임계값 초과 시 알림·재훈련·서비스 제한 대응 절차의 구체성 • 모델 재훈련 트리거 조건과 성능 요구사항의 추적 연계		AG_SD41-1-22
18. 생성형 AI 시스템 모니터링 지표 및 수집 방법 등에 대한 요건 정의 여부	- AI 추론 성능저하 실시간 감지·알림 모니터링 인프라 요건이 정의되었는가? • GPU 사용률 등 AI 전용 성능 지표·임계값·에스컬레이션 및 APM 분산 트레이싱 연동·모델 품질 지표 모니터링 요건 • 로그 수집·보존, 부하 테스트 연계, 운영 이관용 모니터링 정의 - AI 장애·이상 자동 알림 라우팅 체계가 정의되었는가? • AI 시스템 모니터링·알림 요구사항 및 추론 지연·GPU OOM·환각 빈도 급증 등 AI 고유 장애 유형의 알림 대상 반영 • AI 전용 알림 체계 요구사항의 포함 여부	모니터링/재학습	AG_SD41-1-23 AG_SD41-1-24

라. 품질보증

점검항목	검토항목	카테고리	관련 가이드
1. 편향성 지표 정의 및 모델 공평성 평가 지표/목표치 명시 여부	- 편향성·공평성 정량지표 정의와 모니터링 절차가 적정하게 정의되었는가? • 편향성·공평성 정량지표와 집단별 허용 임계값·목표치 수치 • 공평성 평가 계획의 평가 방법·데이터 분리·측정 시점 구체화와 모니터링 주기·임계값 초과 시 보고·조치 절차	편향/공정성	AG_SD41-4-01

점검항목	검토항목	카테고리	관련 가이드
2. 데이터 버전관리, 변경이력 관리 요건 정의 여부	- 데이터 버전관리 요건이 정의되었는가? • 데이터버전관리시스템 도입 요건·대상 범위·명명 체계와 모델-데이터셋·파이프라인·하이퍼파라미터 버전 간 1:1 매핑·재현 요건 • 데이터 변경 감사 추적 기록과 변경/조회 권한 분리 등 접근 통제 요건	거버넌스/데이터계보	AG_SD41-4-02
3. AI 예외 대응 요건과 AI 산출물의 합격 기준·생명주기별 품질활동 계획 여부	- AI 예외 대응 요건이 측정 가능하게 명시되었는가? • 예상 사용자 오류 유형 분류와 금칙어 필터링·탈옥 탐지·프롬프트 인젝션·환각·응답 실패 대응 요건 • 탈옥 탐지율·오류 응답 비율 등 수치 목표와 테스트 케이스 작성 요건 - AI 특유 산출물이 QA 대상으로 명시되고 AI 특성별 QA 방법·정량 합격 기준·생명주기별 품질활동이 정의되었는가? • AI 산출물의 일반 IT 산출물과 구분된 QA 대상 명시 및 유형별 QA 방법·정량 합격 기준·검토 시점·담당자 • QA 계획이 AI 생명주기 전 단계 포함 여부	AI QA체계	AG_SD41-4-03 AG_SD41-4-04
4. AI 개발 방법론 테일러링과 형상관리 계획 수립 여부	- AI 개발 방법론 테일러링 및 AI 특화 산출물의 작성 표준·책임자·검토 시점이 정의되었는가? • AI 특화 방법론 또는 테일러링 계획과 그 기준·검토·승인 절차의 구체성 • AI 고유 산출물의 작성 양식·책임자·시점 정의 및 기능·비기능 요건의 정량적 정의·추적성	형상관리	AG_SD41-4-05

3.3. 설계단계

가. 응용시스템

점검항목	검토항목	카테고리	관련 가이드
1. 다크패턴 배제한 공정 UI, 편향·공정성 설계의 적정성	- 다크패턴 배제 공정 UI가 설계되었는가? • AI 추천 편향 기본값·선택지 불균형 배제와 동의·거부 버튼 시각적 균형·고지 가시성·앵커링 배제 • AI 추천 거부·이의 제기 경로의 동등한 접근성·중요 안내 방해 제거와 공정 UI·자율 선택권 요건 추적성	편향/공정성	AG_SD42-2-01
	- 편향 모니터링 정량 지표 산출 기준이 설계되었는가? • 집단별 비교 편향 지표 항목 및 각 지표 산출 기준 • 집단 정의·비교 기준·지표 산출 주기·데이터 범위 및 추적성		AG_SD42-2-02
	- 편향 유발 입력 사전차단 필터링 로직이 설계되었는가? • 편향 유발 입력 유형별 금칙어 분류·악성 프롬프트 필터링과의 독립적 구분 및 필터링 처리 흐름도·탐지 방식 설계 • 차단 후 이용자 안내 메시지·금칙어 목록 버전 관리 절차		AG_SD42-2-03
	- RAG기법에 기반한 문서 검색 및 응답 검증이 설계되었는가? • RAG 파이프라인 전체 흐름도·분기 노드 및 문서 검색 방식·관련성 임계값·Top-K 컨텍스트 구성·출력 근거 검증 로직 • 검증 실패 시 폴백 처리		AG_SD42-2-04

점검항목	검토항목	카테고리	관련 가이드
2. 설명가능·인간감독 관련한 기능 등 AI 윤리·설명가능성 UI 설계의 적정성	- AI 출력에 대한 관리자 오버라이드 기능이 적정하게 설계되었는가? • 오버라이드 권한 체계·접근 화면과 수정 가능 항목 범위·편집 UI·수정 사유 입력 설계 • 출력결과 수정자·일시·수정 전/후·사유를 포함하는 감사 추적 로그와 오버라이드 승인 단계·이력 조회	AI윤리/xAI	AG_SD42-2-05
	- AI 결과의 한계를 알리는 면책 고지 문구·한계정보 화면·저신뢰도 경고 UI가 설계되었는가? • 결과 화면 면책 고지 문구와 학습 데이터 기간·정확도 범위·적용 제외 케이스 등 한계 정보 화면·신뢰도 수준별 경고 UI 차등 설계 • 고지 문구 가독성·고위험 기능 확인 인터랙션·정합성·추적성		AG_SD42-2-06
	- 서비스·이용자 유형에 적합한 XAI 기법 선정 근거·출력 형태·이용자별 설명 수준·성능 트레이드오프가 반영되었는가? • XAI 적용 필요성·기법 선정 근거·적용 대상 모델과 출력 형태·이용자 유형별 설명 수준·범위·시점 • XAI 추적성과 설명가능성·정확도 트레이드오프 대응 설계		AG_SD42-2-07
	- AI 추론 결과 신뢰점수가 표현 형식·이용자 유형별 노출 수준·의미 설명과 함께 화면에 표시하도록 설계되었는가? • 신뢰점수 표시 UI 요소 설계 • 이용자 유형별 노출 수준 차등화·신뢰점수 의미 설명 UI설계		AG_SD42-2-08

점검항목	검토항목	카테고리	관련 가이드
	- 관리자의 AI 결정 오버라이드 수동변경 기능이 설계되었는가? • 오버라이드 접근 권한 부여된 관리자 역할 범위·변경 가능 항목·범위 제한 및 변경 사유 필수 입력 강제·변경 이력 기록하는 기능 • 수동변경 결과 구분 표시·이력 조회 기능·변경 승인 체계·이력 보관 기한 지원 구현		AG_SD42-2-09
3. 개인화 데이터 처리 알고리즘 설계(사용자 프로파일 활용 시 개인정보 마스킹 로직) 적정성	- AI 개인화 입력 프로파일의 개인정보 항목이 추론과정에서 불필요하게 노출되거나 로깅되지 않도록 설계되었는가? • AI 입력 프로파일 항목·직접/간접 식별자 구분·마스킹 방식과 추론 입력 구조·임시 버퍼·로그·캐시 잔류 방지 설계 • 조합 재식별 위험 처리·동의 철회·삭제 시 비개인화 전환 로직·마스킹 요건 추적성	비식별화	AG_SD42-2-10
4. 추론 실패 시 폴백 및 추론결과 신뢰도가 부족한 상황 발생 시 안전모드로 전환 설계의 적정성	- AI 추론 신뢰도 임계값 미달·오류 시 안전모드 Fail-safe 로직과 진입·해제 조건이 설계되었는가? • 안전모드 발동 신뢰도 임계값의 수치 정의 및 미달 시 기본값 반환·서비스 보류·관리자 전환 절차 설계 • 안전모드 진입·해제 조건·이용자·관리자 안내 메시지·Fail-safe 이력 기록 및 추적성	가용성/이중화	AG_SD42-2-11
5. 악성 프롬프트 필터링 로직 설계의 적정성	- 비정상 출력을 유도하기 위해 입력하는 악성 프롬프트를 탐지하고 차단하는 로직이 구체적으로 설계되었는가? • 프롬프트 인젝션·탈옥·시스템 프롬프트 추출 등 차단 대상 공격 유형 목록 및 정규식·룰셋·AI 분류기 필터링 방식·처리 흐름도·룰셋 갱신 절차 설계 • 차단·경고·로깅 분기 처리·이용자 친화적 안내 문구·오탐 대응 및 추적성	AI특화보안	AG_SD42-2-12

점검항목	검토항목	카테고리	관련 가이드
6. 단일 식별자에 의한 사용자 행동 추적·성능 지표 수집 등 로깅 추적성 설계의 적정성	- 이용자 행동·AI 추론 성능 지표 실시간 수집 로그 객체가 구체화되었는가? • 수집 대상 이벤트 목록 및 이벤트별 로그 객체 구조·사용자 행동·AI 추론 성능 수집 로직의 비동기·비차단 방식 설계 • 로그 내 개인정보 비식별화·로그 객체 스키마 버전 관리 및 추적성	모니터링/재학습	AG_SD42-2-13
7. 비정형 AI 출력을 정형 포맷으로 변환/정제하는 로직 설계 및 외부·공공 API 연계 설계의 적정성	- 외부 시스템·공공 API 연계 프로토콜이 설계되었는가? • 연계 인터페이스 목록·입력 전문 구조 및 AI 추론 결과를 포함한 응답 스키마 완결성과 HTTP 외 애플리케이션 오류 코드 체계	AI모델 API	AG_SD42-2-14
8. 서비스 목적 부합 AI 아키텍처 설계 여부	- 서비스 목적에 부합하는 AI 모델 유형 선정 근거·알고리즘 처리 흐름·프롬프트 전략·환각/편향 대응·품질 지표가 반영되었는가? • AI 모델 유형 선정 근거·입력→처리→출력 아키텍처와 프롬프트 설계·환각·편향 대응·외부 모델 출처·버전·라이선스·SLA 명시 • 요건↔설계 추적성 매트릭스와 측정 가능한 AI 출력 품질 지표 설계	LLM모델	AG_SD42-2-15
9. Agent 툴 호출 Loop Lock·안전정지, 워크플로우 엔진 기반 결정론적 통제 및 정형 판정을 Rule Engine 강제하도록 설계 여부	- Agent 안전정지·Loop 통제가 설계되었는가? • 툴 호출 Loop Lock·반복 호출 탐지 및 Context Mismatch 권한 우회·프롬프트 인젝션 대응 설계 • 레거시 API 타임아웃·오류 Safe Stop 및 표준 안전메시지·관리자 Alert 처리	오케스트레이션	AG_SD42-2-16

나. 데이터베이스

점검항목	검토항목	카테고리	관련 가이드
1. 편향 방지 데이터 특성(피처) 선정 근거·보호변수 프록시 변수 식별·변환 규칙 및 피처 스토어·변환 이력 테이블을 통한 특성 관리 설계의 적정성 여부	- 편향 방지를 위한 데이터 특성(피처)관리 방안이 적정하게 수립되었는가? • 최종 특성 목록·선택 근거와 컬럼별 변환 규칙 정의 및 보호변수 프록시 변수 식별·처리 방안 • 피처 스토어·변환 이력 테이블 설계와 추적 가능성	편향/공정성	AG_SD42-3-01
2. 데이터 계보 추적 스키마 설계의 적정성	- 모델 버전-데이터셋 버전 매핑이 설계되어 모델 학습 데이터 역추적이 가능한가? • 데이터셋 버전 관리 전용 테이블의 필수 컬럼 설계와 모델 버전-데이터셋 버전 매핑 또는 FK 참조 관계 설정 • RAG·파인튜닝 환경에서 LLM·임베딩 모델·청크·어댑터 버전의 추론 결과 함께 기록 설계	거버넌스/데이터계보	AG_SD42-3-02
3. 보호변수·민감정보 컬럼 식별과 비식별화 설계 및 사용자 데이터를 학습에 활용 동의 여부 관리 방안 수립 여부	- 개인정보 비식별화·가명화 View가 설계되었는가? • 원본 테이블 비식별화 View 정의·컬럼별 마스킹 함수·규칙 구체화와 역할별 View 접근 권한 분리·원본 직접 접근 차단 설계 • RAG 청크 임베딩 전 비식별화 처리 및 벡터 DB 컬렉션 역할별 접근 권한 분리 설계	비식별화	AG_SD42-3-03
4. 데이터 정합성을 주기적으로 검증하는 배치(Batch) 설계 여부	- 데이터의 정합성을 주기적으로 검증하는 배치(Batch)가 설계되었는가? • 검증 대상·규칙 완전성·스크립트 명세·배치 스케줄과 검증 결과 저장·오류 데이터 격리·KPI 임계값 설계 • EDB-벡터 DB 동기화 정합성 검증 및 벡터 차원·모델 버전 일관성 검증 배치 설계	데이터 품질	AG_SD42-3-04

점검항목	검토항목	카테고리	관련 가이드
5. 개인정보 암호화 대상 컬럼 식별 및 암·복호화 알고리즘 설계의 적정성	- 개인정보 암호화 대상 컬럼 식별 및 암·복호화 처리가 설계되었는가? • 법적 의무·AI 특성 항목 암호화 대상 식별·알고리즘 지정과 AI 전용 저장소 암호화·키 관리·학습 과정 암호화·복호화 처리 • 임베딩 자산 다중 통제·프롬프트·응답 로그 보안·KMS·키 회전 일관성 설계	암호화	AG_SD42-3-05
6. 테스트 데이터 관리 설계 및 편향·오염 방지 설계 적정성	- 테스트 데이터 수집·편향 방지·오염 방지·민감정보 처리·버전 관리·AI 평가셋 분리 및 누출 방지가 반영되었는가? • 테스트 데이터 수집·확보 계획·편향 방지 설계·학습-검증-테스트 분리 및 민감정보 처리·버전 관리·요구사항-테스트 추적성 • AI 평가셋 분리 정의, 누출 방지 격리·접근 통제 및 재현 가능성 설계	테스트시나리오	AG_SD42-3-06
7. 데이터 수집 다양성 확보 방안 적정성과, 공공데이터 적법성 요건을 반영한 학습데이터 소싱·수집 설계의 적정성 여부	- 공공데이터 연계·활용 시 공공누리 유형·이용약관 등 적법성·데이터 품질·보안 요건이 반영되었는가? • 활용 대상 공공데이터셋의 공공누리 유형·이용약관 확인 및 AI 학습·추론 활용 방식과 라이선스 제약 부합 여부 • 기관별 별도 이용약관 적용 및 데이터 품질·보안	학습데이터	AG_SD42-3-07
8. 학습·추론 저장을 위한 EDB/벡터 DB 구조 설계의 적정성	- AI 서비스 핵심 쿼리 실행계획 최적화가 설계되었는가? • 인덱스·비정규화·사전 집계·파티셔닝·실행계획 분석·힌트 전략 설계 • 벡터 검색 튜닝 이력 보존·하이브리드 검색 측정 포인트·LLM 추론 지연 최적화 설계	학습추론 및 저장	AG_SD42-3-08
	- EDB-벡터 DB 하이브리드 데이터 모델/데이터 정합성이 확보되도록 설계되었는가? • EDB·벡터 DB 하이브리드 데이터 모델 및 데이터 정합성 확보 설계 • RAG 청크-추론 결과 연계·보호변수 속성 정의		AG_SD42-3-09

점검항목	검토항목	카테고리	관련 가이드
	- AI 성능지표를 산출/분석/모니터링하기 위한 데이터마트 스키마가 설계되었는가? • 스키마 구조·성능 지표 팩트·차원 테이블·시계열 집계·원천 ETL·편향·공정성 지표 • AI 운영 지표·RAG·LLM 품질 지표 구분·에이전트 운영 지표 마트 설계		AG_SD42-3-10
	- 정제 전/후 (원천 데이터, 훈련 데이터) 저장 공간 분리가 설계되었는가? • 원천·학습용 데이터 적재 공간 분리와 ERD 서브젝트 영역 구분·이동 경로·접근 권한 차등 설계 • 원본/정제/임베딩/색인 4계층 분리·격리 영역 정의·운영/비운영 영역 분리 설계		AG_SD42-3-11

다. 시스템구조·보안

점검항목	검토항목	카테고리	관련 가이드
1. 개인정보 비식별화 요구사항 설계 반영 및 테스트 데이터 파기 절차 수립 여부	- 개인정보 비식별화 처리 설계 및 테스트 데이터 파기 절차가 설계되었는가? • 비식별화 대상 개인정보 항목 식별과 항목별 기법·처리 시점 구분 설계 • 테스트 완료 후 복구 불가 파기 절차	비식별화	AG_SD42-1-01
2. 성능 지표별 산식·정량적 목표값 확정 및 임계값이 측정 가능한 수준으로 정의 여부	- 성능 지표별 산식·목표값과 추적성·공정성 지표가 확정되었는가? • 정밀도·재현율·F1·BLEU·ROUGE 등 AI 유형별 성능 지표와 산식·수치 목표값 구체화 • 측정용 테스트 데이터셋 구성 적정성, 요구사항-지표 추적성 및 공정성·하위 그룹별 성능 편차 지표	모델평가	AG_SD42-1-02
	- 성능 평가 지표·최소 성능 임계값·테스트 데이터셋 조건이 측정 가능하게 정의되었는가? • 성능 지표와 지표별 최소 임계값의 측정 가능한 수치 • 임계값 미충족 시 처리 절차 및 테스트 데이터셋·측정 환경 명시		AG_SD42-1-03

점검항목	검토항목	카테고리	관련 가이드
3. DR 전환 모의훈련 절차, 방법 등이 구체적으로 계획 여부	- 안전모드(Fail-safe) 동작 조건과 장애 탐지 체계가 구체적으로 설계되었는가? • 안전모드 진입 조건과 폴백 응답·정상 복귀 절차의 상태 전이도 표현 및 이벤트 버스 발행·구독 기반 장애 이벤트 전파 경로 • 장애 탐지 에이전트의 탐지 대상·방법·조치 흐름, AI 고유 장애 유형별 처리 설계 및 요구사항-설계 추적 일관성	가용성/이중화	AG_SD42-1-04
4. CI/CD 배포 파이프라인 보안 요구사항 및 소스/모델 병합 시 자동 보안 검사 파이프라인 설계 여부	- CI/CD 배포 파이프라인 과정에서 발생할 수 있는 모델 변조(Model Tampering), 악의적 프롬프트 삽입 등의 위협에 대한 방어 설계가 되었는가? • CI/CD 파이프라인 보안 항목의 식별·설계 매핑 및 모델 가중치·프롬프트·학습 데이터 병합 시 모델 변조·악의적 삽입 방어 설계 • AI 모델 배포용 ML 파이프라인의 자동 보안 검사	CI/CD파이프라인	AG_SD42-1-05
5. 적대적 입력 유입 시 입력 정제·거부 로직 설계 및 AI 모의 공격 계획 수립 여부	- 모델 공격 방어 및 적대적 입력 정제·거부 로직 요구사항이 프롬프트 인젝션·탈옥 대응 방어 아키텍처로 반영되었는가? • 모델 공격 방어 요구사항 포함 및 설계 산출물 추적·매핑 • 채택한 AI 모델 유형별 적용 가능한 공격 유형과 방어 범위 사전 파악	AI특화보안	AG_SD42-1-06
6. 접근통제 및 인증체계 설계의 적정성	- 관리자 페이지 접근통제 및 인증방식이 적정하게 설계되었는가? • 관리자 페이지 접근 통제 정책·2FA/MFA 적용 대상·방식 및 허용 IP 목록·세션 타임아웃·잠금 임계값 등 계정 보안 정책 • 요구정의 단계에서 정의된 인증 요건의 추적 및 관리자 접근 감사 추적 설계	인증/인가/접근통제	AG_SD42-1-07

점검항목	검토항목	카테고리	관련 가이드
	- AI 추론 API 접근통제 구조가 적정하게 설계되었는가? • 호출 시나리오별 OAuth 2.0 Grant Type·JWT 토큰 설계 및 API Gateway 중앙 집중 인증·인가·엔드포인트별 권한 매트릭스 • 토큰 생명주기 및 비인가 접근 에러 응답·로깅 정책		AG_SD42-1-08
	- 모델 저장소 접근통제 요구사항이 RBAC 권한 매트릭스·인가된 CI/CD만 쓰기 허용·무결성 검증·상태전이 승인·감사로그로 반영되었는가? • 모델 저장소 접근 통제의 RBAC 권한 매트릭스·최소권한 및 인가된 CI/CD만 쓰기 허용·모델 파일 무결성 검증·상태 전이 승인 워크플로우 반영 • 접근 감사 로그 설계와 보안요구사항의 설계 매핑 완전성		AG_SD42-1-09
7. 전송 구간 암호화 방안과 대용량 데이터 저장 시 암호화 방안 설계 여부	- HSM 연동 및 API 키·암호화 키 관리 아키텍처 방안이 수립되었는가? • HSM 연동 구성·모델 가중치 서명 키 관리 및 추론 API 키 순환·외부 LLM API 키 Vault 분리 설계 • 학습 데이터 암호화 키 분리 보관 및 키 접근 권한·감사 로그 설계	암호화	AG_SD42-1-10
	- 민감정보가 전송되는 통신 구간별 암호화·인증서 관리 방안이 수립되었는가? • 구간별 TLS 적용 대상·최소 버전·암호 스위트 및 내부 마이크로서비스 mTLS·외부 API 인증서 검증·HSTS 설계 • 인증서 수명주기 관리 및 학습 파이프라인 구간 TLS 적용 설계		AG_SD42-1-11

점검항목	검토항목	카테고리	관련 가이드
8. AI 특화 보안약점 예방 아키텍처와 오픈소스 의존성·취약점 관리 설계의 적정성	- DDoS 공격 방어·비정상 트래픽 차단 방안이 구체적으로 수립되었는가? • DDoS 방어 항목의 설계 매핑 및 AI 추론 API 대상 L7 방어·프롬프트 인젝션 기반 연산 고갈 등 WAF 탐지 규칙 • 자동 대응 임계값 및 방어 장비 배치 위치 - 입력검증·출력인코딩·System Prompt 보안 방안이 구체적으로 수립되었는가? • SQL Injection·프롬프트 인젝션 방어 레이어와 입력 유효성 검증·출력 인코딩·System Prompt 보안관리 설계 • SQL/프롬프트 인젝션 요구사항의 설계 산출물 누락 없는 매핑	시큐어코딩	AG_SD42-1-12 AG_SD42-1-13
	- 웹셸(Webshell) 등 악성파일 업로드 차단 방안이 수립되었는가? • 화이트리스트 방식 확장자 제한과 서버 측 검증, MIME 타입·매직 바이트 이중 검증 및 업로드 디렉터리 실행 권한 분리·웹 루트 외부 저장 설계 • RAG·멀티모달 입력 등 AI 특화 경로의 악성코드 스캔 및 간접 프롬프트 인젝션 방어 설계		AG_SD42-1-14
9. 클라우드 환경 보안체계 및 연계서버·방화벽 통제 방안 설계의 적정성	- 클라우드 인프라 보안 요구사항이 VPC 아키텍처·Security Group 최소권한 포트 통제·서브넷 격리로 설계되었는가? • VPC 격리·Security Group 최소 오픈 등 요구사항의 설계 반영 및 포트 오픈 범위의 최소권한 원칙 • 모델 학습 서버·추론 API·벡터 DB 등 AI 워크로드 보안 등급에 맞는 서브넷 격리 설계	망분리/폐쇄망	AG_SD42-1-15

점검항목	검토항목	카테고리	관련 가이드
10. AI 특화 시나리오에 대해 정상·예외 상황 포함 여부 및 테스트 자동화 도구 활용 계획이 구체적으로 수립 여부	- AI 특유의 테스트 환경(데이터 유사성, 모델 버전 관리, LLM 연계 환경 등)이 구체적으로 계획되었는가? • AI 테스트 환경 구성 요소 정의·운영 환경 유사성 요건 명시 및 가상 환경 활용 계획·한계 • AI 특화 테스트 항목 정의, 테스트 데이터 대표성·품질 기준 정량 명시 및 LLM·라이브러리 버전 관리 계획 - 테스트 자동화 도구 활용 계획이 구체적으로 반영되어 있는가? • AI 기능·품질 요구사항의 측정 가능한 수치 테스트 항목화·요구사항 1:1 대응 및 자동화 대상 선정 기준 • 성능 스크립트 구체성, AI 특화 보안 항목 포함, 도구 AI 적합성 및 수작업/자동화 역할 분리	테스트시나리오	AG_SD42-1-16 AG_SD42-1-17
11. 온프레미스/로컬 환경에서 sLLM 경량 모델 활용 시 선정 요건, 운영계획 수립 적정성	- 온프레미스·로컬 환경 sLLM 경량 모델 선정·성능검증·보안·운영 요건이 수립되었는가? • 망분리·데이터 외부 이전 제약을 고려한 sLLM 활용 적합성 및 예상 사용량·동시성·응답시간 요구 대비 모델 성능·파인튜닝 계획 • 보유 GPU 인프라·예산 대비 운영 요건 및 보안 요건 충족	모델경량화	AG_SD42-1-18
12. AI 특화 요소에 근거한 Auto-Scaling 설계 및 서버 자원 병목해소를 위한 아키텍처 설계 여부	- AI 추론 서버 Auto-scaling 방안이 적정하게 설계되었는가? • Auto-scaling 정책이 최소/최대 인스턴스·임계값·쿨다운 수치 정의·GPU 용량 산정에 근거 • 스케일링 실패 시 폴백 전략과 성능·확장성 요구사항의 설계 매핑	추론인프라	AG_SD42-1-19

3.4. 구현단계

가. 응용시스템

점검항목	검토항목	카테고리	관련 가이드
1. AI 실체 표시 및 주의사항 고지 문구 구현의 적정성	- 튜토리얼·안내 팝업·AI 특성 고지 내용 등이 사용자 환경별 작동이 시험되었는가? • 안내 팝업 노출 조건의 설계-구현 일치·재노출 통제 및 AI 특성 고지 내용·UI 가독성·접근성 여부	UI/UX, 접근성	AG_SD43-2-01
2. 마스킹·비식별화 View의 권한별 차등 여부	- 권한별 개인정보·민감정보 마스킹·비식별화 뷰가 규칙대로 차등 적용되고 우회가 불가능한가? • 마스킹 대상 항목의 실제 적용과 마스킹 규칙 일치 및 권한별 차등 마스킹 적용 • 학습 데이터·추론 결과 화면 마스킹 적용과 URL·API 등 마스킹 우회 불가 여부	비식별화	AG_SD43-2-02
3. 프롬프트 엔지니어링 관리체계 구축의 적정성 여부	- 프롬프트 설계·버전관리·보안·성능 검증이 체계적으로 운영되고 있는가? • 프롬프트 인젝션·민감정보 노출·시스템 프롬프트 유출 방어 등 보안 통제 • 골든셋 기반 정량 평가 및 변경 빈도에 따른 버전 관리 운영	프롬프트설계	AG_SD43-2-03
4. 동일하거나 빈출되는 사용자 질의에 대해 캐시 서버(Redis 등)를 통한 응답 구현 여부	- 캐싱 히트율(Cache Hit Rate) 목표치 달성 여부와 캐싱 적용으로 인한 응답 속도 개선 효과 여부가 점검되었는가? • 캐싱 대상 질의 선정 기준·질의 유형별 TTL 정책의 명확한 설계 및 캐시 히트율 목표값 정의·달성·적중/미적중 응답 시간 비교 측정 • 캐시 서버 구현·연계·캐시 키·무효화 정책·개인정보 포함 응답의 캐싱 제외 처리	RAG파이프라인	AG_SD43-2-04

점검항목	검토항목	카테고리	관련 가이드
5. PDF 등 비정형 문서를 인식·추출·분류·요약하는 기능의 보안성, 정확성 여부	- 비정형 문서 인식·추출·분류·요약 기능의 정확성·보안·접근성 요건이 충족되었는가? • 대상 문서 유형·품질 분포 분석 및 OCR·레이아웃 분석·개체명 인식·분류·요약 처리의 정확성과 처리량·SLA 정의 • 추출 과정의 보안·개인정보 보호 및 접근성 요건 충족 여부	멀티모달/에이전트	AG_SD43-2-05
6. 모호·불완전 질의에 대한 재확인 기능 시험과 챗봇 상담 정확성 검증	- AI 민원·상담 챗봇의 정확한 처리와, 공공서비스 품질 수준이 충족되었는가? • AI 민원 유형·상담 정확성 처리 점검과 상담 내용 내 개인정보 처리·보호 • 민감 민원의 사람 응대 연계 및 취약계층 지원	챗봇	AG_SD43-2-06

나. 데이터베이스

점검항목	검토항목	카테고리	관련 가이드
1. 편견·금칙어 사전 DB 구축 및 관리의 적정성	- 편견/금칙어 사전 DB 구축과 업데이트 시 필터링 등 관리 절차가 구현되었는가? • 사전 DB 구축과 금칙어의 입력·출력 실제 차단 동작 및 사전 업데이트 즉각 반영·관리 화면·변경 이력 관리 • 금칙어 필터링의 오차단·미탐지율 목표 충족 여부	금칙어/가드레일	AG_SD43-3-01
2. 공통코드 무결성·ETL 파이프라인 단계별 데이터 계보 추적 체계 구현 여부	- 공통 코드 테이블, 공통 코드값 무결성이 확보되었는가? • 코드-마스터 간 참조 무결성 위배 및 고아 코드 존재 여부와 공통 코드 체계의 DB 일치 구현 및 참조 제약 • AI 학습 범주형 특성 컬럼 코드 값의 코드 테이블 일치 여부	거버넌스/데이터계보	AG_SD43-3-02

점검항목	검토항목	카테고리	관련 가이드
	- ETL/전처리 파이프라인 데이터 계보 추적이 가능한가? • 파이프라인 이력 테이블의 물리 DB 생성과 단계별 필수 이력 항목 누락 없는 적재 및 입출력 건수 정합성 • 오류·실패 이력 기록과 이력 보존·아카이빙 정책		AG_SD43-3-03
3. 데이터 이행 결과에 대한 검증과, 이상 데이터 탐지/제거 적정성	- 레거시 데이터 이행 원본-타겟 건수 일치 및 변환 오류, 무결성·정밀도 손실 검증이 되었는가? • 이행 대상 전체 테이블의 원본-타겟 건수 100% 일치, 오류 데이터 0건, 변환 규칙 정확 적용 및 이행 범위 누락 없는 완료 • 무결성 검증과 수치형 데이터 정밀도 손실 여부	정제(클렌징)/변환	AG_SD43-3-04
	- 이상 데이터 실제 제거 및 전후 건수·제거율 정량 기록과 허용 범위 부합이 이루어졌는가? • 이상치 탐지 기준 적용과 전처리 전후 건수 정량 기록 및 제거율의 허용 범위 부합·제거 사유 기록 • 전후 데이터 분포 변화 비교·문서화와 샘플 기반 제거 확인 가능성		AG_SD43-3-05
4. 데이터 정합성을 주기적으로 검증하는 배치(Batch)가 정상적으로 작동하는지 시험 여부	- 데이터 정합성 검증이 배치 파이프라인으로 구현되어 정상 작동하는가? • 정합성 검증 규칙의 배치 구현과 컬럼별 위반 건수 기록 및 위반 데이터 차단 처리 • 검증 배치의 자동 정기 실행과 위반율의 허용 기준 통제	데이터 품질	AG_SD43-3-06
5. 외부 크롤링으로 수집한 데이터가 누락 없이 적재되었는지 여부	- 외부 크롤링으로 수집한 데이터가 누락 없이 적재되었는가? • 수집 대상 대비 적재 건수, 수집 실패 원인 건별 기록 및 라이선스·출처 URL·수집 일시 메타 컬럼 관리 • 중복 수집 탐지·처리와 robots.txt 준수	학습데이터	AG_SD43-3-07

점검항목	검토항목	카테고리	관련 가이드
6. 학습·추론 저장을 위한 EDB/벡터 DB 설계를 반영한 구현 여부	- AI 추론 프롬프트·응답이 로그 테이블에 누락 없이 적재되고 필수 항목·개인정보 마스킹·보존 정책이 반영되었는가? • 응용서버 추론 건수와 DB 로그 적재 건수의 일치와 필수 로그 항목 적재 및 테이블 구조 일치 • 비동기 로깅 유실 방지, 로그 내 개인정보 마스킹, 보존·삭제 정책 구현 - 정제 전 원본과 정제 후 가공 데이터가 논리적/물리적으로 분리되고, 접근 권한이 통제되고 있는가? • 논리/물리적 분리 구조의 실제 DB 구현 및 접근 권한 분리·통제	학습추론 및 저장	AG_SD43-3-08 AG_SD43-3-09
7. 데이터 소스 상태 모니터링 메커니즘 구현·시험의 적정성	- 데이터 소스 변경 자동 감지 및 관리자 알림 모니터링 메커니즘이 구현·시험되었는가? • 모든 데이터 소스의 모니터링 대상 등록과 스키마·접속 상태 변경 자동 감지 로직 구현 및 알림 채널·에스컬레이션·변경 이력 로그 기록 • 다양한 변경 유형을 포괄하는 시험 시나리오와 합격 기준 충족 여부	데이터드리프트	AG_SD43-3-10

다. 시스템구조·보안

점검항목	검토항목	카테고리	관련 가이드
1. 적대적 사례(Evasion) 공격에 대한 방어 기능 구현 적정성	- 적대적 사례 (Evasion) 공격에 대한 방어 기능이 적정하게 구축되고 시험을 통해 확인되었는가? • AI 보안시험 시나리오에 변형된 이미지·텍스트·프롬프트를 이용한 회피 공격 시도에 대한 탐지 및 차단 기능 • 탐지 실패 시 폴백 메커니즘과 취약점 조치·재점검	AI특화보안	AG_SD43-1-01

점검항목	검토항목	카테고리	관련 가이드
2. 모델 저장소에 비인가 접근 차단 구현의 적정성	- 모델 저장소에 비인가 접근이 차단되고, 접근 감사로그 기록이 구현되었는가? • RBAC/ACL 정책의 실제 적용, CI/CD 파이프라인 인가 계정 제한 및 관리 콘솔·API의 관리자 IP 화이트리스트 제한 • 모델 접근·수정·삭제 감사로그 기록 및 발견 취약점 조치	인증/인가/접근통제	AG_SD43-1-02
3. 학습/추론 관련 데이터가 저장된 파일 스토리지와 데이터베이스에 대해 물리적 암호화 구현 여부	- AI 모델 및 파일 스토리지 암호화가 설계대로 구현되었는가? • 모델 가중치 파일 스토리지 및 학습·임베딩 DB의 암호화 적용과 키 생성·보관·교체·폐기 절차 수립·이행 및 KMS/Vault 연동 • 암호화 적용 후 모델 로딩·추론 성능 영향 시험결과 및 보안점검결과 기록	암호화	AG_SD43-1-03
4. 보안약점 예방 기능 구현의 적정성	- 보안약점 예방 필터 계층이 실제 시스템에 적정하게 구현되었는가? • SQLi·XSS 등 웹 공격 차단 규칙이 설계대로 구현되고, 시험 결과 및 프롬프트 인젝션·탈옥 입력 필터링 모듈의 AI 계층 적용·동작 • 시큐어코딩 점검 결과·조치 완료 및 전개 전 보안 확인 절차 - 파일 업로드 기능에 악성코드 실행 방지 필터링이 검증되었는가? • 파일 업로드 보안 통제, 악성코드 스캔 및 이미지·문서 내 악성 스크립트·매크로 차단과 격리 실행 • 전처리 단계 간접 프롬프트 인젝션 필터링 검증과 악성파일 보안시험 수행·전개 환경 적용	시큐어코딩	AG_SD43-1-04 AG_SD43-1-05

3.5. 시험단계

가. 응용시스템

점검항목	검토항목	카테고리	관련 가이드
1. 편향 모니터링 대시보드가 편향 수치를 정확하게 표출하는지에 대한 시험 여부	- 편향 모니터링 지표가 정상적으로 표출되는가? • 편향 지표 항목의 누락 없는 화면 표출과 계산 방식·집단 분류 기준·갱신 주기 일치 여부 • 임계값 초과 경보 동작 및 편향 관련 AI 특화 시험 항목	편향/공정성	AG_SD44-2-01
2. AI 윤리·인간감독·설명가능성 기능 구현 적정 및 검증 여부	- XAI 설명가능성 기능이 화면에서 정상 동작하는가? • XAI 기능의 결과값을 사용자가 이해할 수 있는 형태로 시각화 하여 화면에 표출 • 응답 시간 내 정상 표출·오류 처리·XAI 전용 시험 항목 독립 포함·특이값·경계값 입력 처리	AI윤리/xAI	AG_SD44-2-02
3. 환각 억제 시험의 적정성 여부	-환각 억제 기능이 설계 의도에 따라 정상적으로 작동하는지 시험되었는가? • 출처 명시·근거 범위 제한·응답 거부 조건 정의와 환각 억제율·출처 명시율 수치 기재 및 기준값 충족 • 출처 불명 질의 응답 거부 동작·출처 정확성 시험·환각 발생 이력 기록 여부	할루시네이션	AG_SD44-2-03
4. 다국어 처리, 웹 접근성 기준 등을 준수 여부에 대한 시험 여부	- 비동기 추론 처리 중 프로그레스 바·상태 표출이 정상 동작하는가? • 상태 정보 정의 및 요청 직후 즉시 로딩 표출·지속성·장시간 처리 시 추가 안내·취소 수단·완료 시 정상 전환	UI/UX, 접근성	AG_SD44-2-04
5. 오작동 시 Fail-safe 자동 전환 및 복구 시험 적정성	- 시스템 오작동 시 기본값으로 응답하고, 안전 모드로 전환이 점검되었는가? • 오작동 유형별(타임아웃·AI 추론 실패·외부 API 오류) 시나리오 시험·디폴트 반환 전환	가용성/이중화	AG_SD44-2-05

점검항목	검토항목	카테고리	관련 가이드
	- 임계값 초과·장애 시 안전모드 자동 전환 여부가 시험되었는가? • 안전모드 전환 시험·전환 응답 시간 기준값 충족·이용자 안내 표출 • 안전모드 해제·복구 절차에 따른 시험		AG_SD44-2-06
6. AI 서비스의 평균 응답시간이 목표 수준 달성 여부에 대한 테스트 적정성	- 실시간 동시 접속 부하 조건에서 기능 유형별 응답시간·오류율 목표를 달성하고 최대 부하·목표 미달 원인 분석·모니터링이 점검되었는가? • 기능 유형별 응답시간 목표값·동시 접속자 수 기준의 수치 정의·승인 및 부하 조건별 평균·P95·P99 응답시간·오류율 수치 기재·목표값 • AI 특화 기능 응답시간 독립 시험·최대 부하 조건 시험·목표 미달 원인 분석·개선·재시험·운영 모니터링 기준	성능	AG_SD44-2-07
7. 프롬프트 인젝션 방어 필터 기능 점검의 적정성	- 프롬프트 인젝션 방어 필터의 차단율·오탐률이 수치로 검증되었는가? • 차단 대상 공격 유형·시험 시나리오 대응과 차단율·오탐률 수치 기준 정의 및 기준 충족 여부 • 차단 시 에러 메시지 내부 정보 미노출 및 탈옥·오탐 시험과 시도 이력 로그 포함 여부	AI특화보안	AG_SD44-2-08
8. AI 전 구간 Trace ID 연결 등 모니터링·재학습 로깅 점검 여부	- 긍정 피드백 데이터를 재학습용 데이터 마트로 자동 선별·이관하는 파이프라인이 정상 동작하는가? • 긍정 피드백 선별 기준·이관 항목·주기·데이터 마트 스키마 정의 및 추론 결과 생성부터 이관까지 전 구간 시험·경계값·이상 데이터 제외 시나리오 • 이관 데이터 스키마 일치·이관 오류 재처리·중복 방지·이관 이력 추적성·데이터 품질·비식별화 정의	모니터링/재학습	AG_SD44-2-09

점검항목	검토항목	카테고리	관련 가이드
	- 실시간 이용자 행동·AI 추론 성능 지표 로깅이 점검되었는가? • 수집 이벤트 유형·필드·형식·적재 지연 기준·개인정보 처리 방침 정의 및 실제 적재 로그의 설계 일치·적재 지연 허용 기준 이내·중복·누락 방지 • 로그 내 개인정보 비식별화·AI 추론 로그와 행동 로그의 이벤트 ID 연결·주요 지표 집계·조회 기능		AG_SD44-2-10
9. AI 특화 항목을 포함한 연계시험 여부와 연계 데이터 정합성 확인 여부	- 외부 연계 시스템 통합시험에 AI 특화 항목을 포함하여 점검되었는가? • 외부 AI API 연계 구간의 AI 특화 예외 시나리오·폴백 처리, 데이터 정합성·민감정보 비의도 송신 방지 • 결함 관리·재시험 이력 및 배포·전개 방안 수립	외부연계	AG_SD44-2-11
	- AI 서비스-연계시스템 간 API 송수신 데이터 정합성이 점검되었는가? • API 연계 명세의 시스템별 정의·AI 특화 데이터 필드 포함 및 정상·오류 케이스 독립 시험·통과율·포맷 오류율·필드 누락률의 허용 기준 • 연계 오류 시 재시도 로직·이용자 안내 메시지·오류 이력 기록 정의 및 시험 확인		AG_SD44-2-12

나. 데이터베이스

점검항목	검토항목	카테고리	관련 가이드
1. 보호변수의 분리 보관·접근 통제 및 데이터 특성 추출 시 편향·공정성 확보의 적정성	- 보호변수의 별도 스키마 분리 보관·접근 통제가 구현되었는가? • 보호변수가 별도 테이블·스키마 분리 보관 구현 및 훈련 프로세스 접근 통제 • 훈련용 뷰의 보호변수 배제 정의와 대리변수 식별·문서화	편향/공정성	AG_SD44-3-01

점검항목	검토항목	카테고리	관련 가이드
2. 학습 데이터 버전 관리 적정성 및 모델-데이터 리비전 연동 시험의 적정성 여부	- 학습 데이터 버전관리 체계가 수립되고, 모델-데이터 변경 통제가 적정하게 동작하는지 검증이 이루어졌는가? • 데이터 리비전 식별 체계의 메타 테이블 구현과 데이터 변경 시 리비전 생성·기록·모델-데이터 리비전 연동 매핑·복구 동작 • 리비전별 메타정보 기록과 버전관리 운영 절차 매뉴얼 문서화	거버넌스/데이터계보	AG_SD44-3-02
3. 개인정보 비식별화 이행 및 동의 철회한 데이터 처리의 적정성	- 학습·시험에 사용된 개인정보의 비식별화·파기가 적정하게 이행되고 근거가 관리되고 있는가? • 개인정보가 복구 불가 수준의 완전 파기 이행 및 담당자·개인정보보호 담당자·PM의 서명·승인 • 마스킹 대상 항목 누락 없는 처리, 산출물 간 처리 결과 일치성 및 생성형 AI 파인튜닝 학습 데이터의 개인정보 처리·잔류 위험 대응	비식별화	AG_SD44-3-03
	- 데이터 활용 동의 철회 회원 데이터가 학습 파이프라인에서 제외되었는가? • 옵트아웃 신청 화면 구현, 동의 상태 관리 테이블에 철회 상태 정확한 기록 및 학습 데이터 추출 시 옵트아웃 회원의 실제 제외 • 옵트아웃 처리 이력 관리와 재동의 시 학습 데이터 재포함 프로세스		AG_SD44-3-04
4. 대용량 데이터 백업 스케줄러가 지정된 시간에 자동으로 실행되고, 백업이 성공적으로 완료 여부에 대한 시험	- 대용량 데이터 백업 스케줄러가 지정된 시간에 자동 실행·성공적 완료, 복구 시험이 실시되었는가? • 백업 스케줄러의 지정 시간 자동 실행, 백업 목표 시간 이내 성공 완료, 백업 대상 일치 및 복구 시험 • 백업 실패 시 알림·재시도 메커니즘 동작과 실제 백업 파일의 물리적 생성·보존	데이터 백업/복구	AG_SD44-3-05

점검항목	검토항목	카테고리	관련 가이드
5. 악성 노이즈·데이터 중독 필터링 스크립트 시험 및 결과 적정성	- 악성 노이즈·데이터 중독 필터링 스크립트 시험과 파이프라인 자동 통합이 이루어졌는가? • 악성 데이터 탐지 규칙의 필터 스크립트 구현과 악성 라벨·이상 패턴 차단 시험·차단율 정량 기록 • 정상 데이터 오차단율 검증과 필터의 파이프라인 자동 통합·실행	AI특화보안	AG_SD44-3-06
6. 관리자/일반 사용자 DB 권한 통제 적정성 여부	- 역할별 DB 계정 분리와 최소 권한 원칙이 적용되었는가? • 역할별 DB 계정 분리 생성과 일반 계정의 DML 권한 제한 및 학습 데이터 변경 권한의 전담 계정 제한 • 공용 계정 금지·1인 1계정 원칙과 비밀번호 정책 적용	인증/인가/접근통제	AG_SD44-3-07
7. 파일 스토리지 및 데이터베이스 민감 컬럼에 대한 암호화·복호화 적정성	- 스토리지 및 데이터베이스 민감 컬럼에 대한 암호화·복호화가 정상적으로 작동하는가? • 민감 컬럼의 암호화 적용과 NIST 승인 표준 알고리즘 사용 및 암·복호화 정상 동작·파일 스토리지 암호화·암호키 관리 체계 수립 • 암호화 후 성능 저하의 SLA 허용 범위 이내 여부와 DB 직접 조회 시 암호문 확인	암호화	AG_SD44-3-08
8. 성능 평가용 마트 집계 데이터의 원천 대비 정합성·교차 검증 및 모니터링 대시보드 표출 통계 수치 정확성 확보의 적정성 여부	- 성능 평가용 마트 집계 데이터의 원천 대비 정합성과 모니터링 대시보드 통계 수치 정확성이 확보되었는가? • 마트 스키마의 차원·측정값 정의의 성능 KPI 일치, ETL 집계 로직 정합 및 원천-마트 건수 일치·주요 집계 수치 교차 검증·시간대 처리 일관성 • 대시보드 표출 수치와 마트 집계 결과 일치 및 갱신 주기·예외 데이터 처리 적정성	모니터링/재학습	AG_SD44-3-09

점검항목	검토항목	카테고리	관련 가이드
9. 학습데이터 샘플링 적용 후 클래스 비율이 설계 목표에 부합하는 수준으로 균형을 달성 여부	- 클래스 불균형 해소 샘플링 기법 적용 및 전후 클래스별 건수·비율 목표 부합과 데이터 누수 방지가 이루어졌는가? • 샘플링 기법 실제 적용, 전후 클래스별 건수 기록 및 샘플링 후 클래스 비율 목표 부합·주요 파라미터 기록 • 합성 데이터 품질 검증과 데이터 분리 전 샘플링	학습데이터	AG_SD44-3-10
10. 인덱스/파티셔닝 효율성 및 실시간 스트리밍 적재 무유실·중복 방지 동작의 적정성	- 설계 ERD·테이블 정의가 물리 DB에 100% 일치 구현되고 AI 특화 테이블 누락이 없는가? • ERD 엔터티의 물리 테이블 생성 일치, AI 특화 테이블 구현 완전성 및 컬럼 정의·제약조건의 DDL 일치 • 인덱스 설계 반영, 파이프라인 소스-타겟 매핑 정합성, 무결성	학습추론 및 저장	AG_SD44-3-11
	- 실시간 스트리밍 적재 파이프라인 동작이 시험되었는가? • 메시지 발행 건수와 DB 적재 건수의 100% 일치, 적재 지연 SLA 이내 및 브로커 장애 시 무유실 복구·중복 적재 방지 동작 • 토픽 구조 구현, 목표 처리량 달성, Event Time-Load Time 비교 교차 검증		AG_SD44-3-12
	- 대용량 쿼리 인덱스·파티셔닝 효율성이 확보되었는가? • 주요 대용량 쿼리의 Full Table Scan 미발생, 파티션 히트 정상 동작 및 응답 시간 SLA 달성 여부 • 인덱스·파티셔닝 키의 실제 DB 적용과 AI 학습 배치 쿼리의 파티션 히트 동작		AG_SD44-3-13

다. 시스템구조·보안

점검항목	검토항목	카테고리	관련 가이드
1. 프롬프트 및 모델 응답에 포함된 개인정보 유출 방지 기능 검증의 적정성	- 프롬프트 입력·AI 응답에 포함된 개인정보가 유출되지 않도록 DLP 연동 구현과 시험이 진행되었는가? • DLP 연동 아키텍처의 실제 구현, 주민번호·여권번호 등 탐지 패턴 적용 • 개인정보 포함 프롬프트 전송 시 차단·경고 동작과 AI 응답 출력 필터링·마스킹 • DLP 차단·경고 이벤트 로그 기록 및 관리자 알림 체계 구현	개인정보보호	AG_SD44-1-01
2. 성능 지표별 목표 임계값 달성 여부	- 성능평가 지표의 실제 측정값 대비 목표 임계값 달성 여부가 확인되었는가? • 핵심 성능 지표 실제 측정값이 목표 충족 • 테스트 데이터셋의 대표성·학습 데이터와의 독립성 • 미달 항목 조치계획·미달 배포 시 검수 확인, 지표 선택 적정성 및 결과의 재현성·검증 독립성	모델평가	AG_SD44-1-02
	- AI 추론 파이프라인의 성능 특성을 반영한 용량 충분성이 확인되었는가? • 추론 지연·TTFT·TPOT 등 AI 특화 지표 포함한 운영 환경 동등 시험 환경 • 실제 사용 패턴 반영 부하 시나리오 현실성 • 목표 100% 달성 증빙, 한계 부하 시 안정적 degradation 및 테스트 도구·방법론		AG_SD44-1-03
3. 안전모드(Fail-safe) 이중화 전환 가동 및 안전모드 안내·정상 복구 검증 여부	- RTO 전환, 폴백 모델 응답 품질, AI 장애유형별 전환, 사용자 안내·서비스 제한, 정상 복구가 검증되었는가? • Failover 시험 시나리오·결과 기록, 장애 시 이중화 서버 3초 이내 전환 검증 및 폴백 모델·서버의 응답 품질 시험과 AI 장애 유형별 전환 절차 • 안전모드 사용자 안내·서비스 제한 적용, Failback 정상 복귀 검증 및 전환 임계값	가용성/이중화	AG_SD44-1-04

점검항목	검토항목	카테고리	관련 가이드
4. 접근통제 및 인증 체계 시험 여부	- API 게이트웨이의 인증 및 인가 통제 메커니즘이 시험되었는가? • JWT 기반 인증과 무효 토큰 호출 시 HTTP 401 반환, 토큰 만료·갱신 정책 적정성 및 관리 기능의 RBAC 분리 구현과 권한 미보유 시 403 처리 • 서비스 간 통신 토큰 최소 권한 적용과 토큰 탈취 대응 구현	인증/인가/접근통제	AG_SD44-1-05
	- 내부 서버 간 제로트러스트 인증 (mTLS/토큰)이 되고 있는지 시험되었는가? • WAS와 추론 서버 간 mTLS 상호 인증·추론 API 토큰 기반 인증·권한 검증 적용 및 인증 없는 직접 접근 차단 시험과 인증서·토큰 만료·폐기 처리 절차 • 내부 통신 TLS 1.2 이상 암호화와 비정상 접근 차단·감사로그 동작		AG_SD44-1-06
	- AI 서비스 관리자 페이지 접근통제가 시험되었는가? • 허용 IP 대역의 실제 반영·비인가 IP 접근 차단 시험 및 프록시·VPN·포트 직접 접근 등 우회 경로 차단 검증과 접근 시도 로그·경보 설정 • MFA 우회 불가성 테스트, 관리자 변경 이력 감사로그 동작 및 전개 계획		AG_SD44-1-07
5. 전송 구간 암호화 검증과 대용량 데이터 저장 시 암호화에 대한 검증 여부	- AI 추론 API·모델 서빙 구간 TLS 1.2 이상 적용·취약 프로토콜 비활성화, 인증서 유효성·체인, HTTPS 강제, 내부 구간 mTLS, 암호 스위트·PFS, 모델 다운로드 구간 암호화·무결성이 확인되었는가? • 추론 API·모델 서빙 엔드포인트의 TLS 1.2 이상 적용·취약 프로토콜 비활성화 및 인증서 유효성·체인 완전성, HTTPS 강제, 내부 구간 mTLS 적용 • 취약 암호 스위트 비활성화·PFS 지원과 모델 다운로드 구간 암호화·무결성 검증	암호화	AG_SD44-1-08

점검항목	검토항목	카테고리	관련 가이드
6. 전체 AI 스택에 대한 취약점, 보안패치, 의존성 점검 수행 여부	- DDoS 등 대용량 트래픽 방어 모의훈련 및 WAF/L7 차단 동작이 시험되었는가? • AI 추론 API를 명시적 대상으로 한 DDoS 모의훈련 시나리오·L7 공격 유형 포함 및 임계값 초과 시 WAF/L7 차단 동작 • 정상·공격 트래픽 분리, 실 운영 트래픽 부하 반영 - 서버에 보안 패치 적용 및 CVE 취약점 잔존 점검되었는가? • AI 서비스 구성 서버 목록과 서버별 OS·미들웨어 버전·소프트웨어 스택 파악 및 GPU 드라이버·CUDA·컨테이너 런타임 등 AI 고유 스택의 CVE 스캐닝 도구·패치 절차 • 발주기관 패치 정책 대비 점검 범위·중점 사항 확정	시큐어코딩	AG_SD44-1-09 AG_SD44-1-10
7. 클라우드 환경 보안체계 및 연계서버·방화벽 통제에 대한 시험 여부	- 망분리 환경의 AI 모델 훈련 인프라 구축 여부 및 외부 인터넷으로의 아웃바운드 통신을 차단하는가? • 훈련 서버의 망분리 폐쇄 구간 실제 배치, 외부 인터넷 아웃바운드 기본 차단 및 예외 룰의 승인·아웃바운드 통신 차단 시험 • 학습 데이터 반입 경로 제한과 모델 가중치 반출 승인·기록 절차 - 클라우드 보안 설정(방화벽, VPC 등)이 실제 시스템에 적정하게 작동 여부가 시험되었는가? • 추론 서버·GPU 클러스터·벡터 DB의 서브넷 분리 배치, 불필요 포트 100% 차단 및 추론 API 엔드포인트·GPU 노드 간 통신 포트의 외부 노출 차단 검증 • NACL 설정 적정성, 설계 대비 구현 일치성 및 모니터링 포트 외부 접근 제한	망분리/폐쇄망	AG_SD44-1-11 AG_SD44-1-12

점검항목	검토항목	카테고리	관련 가이드
8. AI 서비스 신뢰성 테스트 시행 및 결과 관리 적정성 여부	- AI 모델 신뢰성 테스트 시나리오가 100% 수행되고 결과가 기록/관리되며 절차가 준수되었는가? • 신뢰성 시나리오 목록·수량 확인 및 계획 대비 결과 제출·형식적 완결성 • 테스트 수행 방식·담당자·역할 및 결과 승인	AI QA체계	AG_SD44-1-13
9. 테스트 결과가 사전에 정의된 성능 지표 충족 여부	- 테스트 결과가 사전 설계된 성능 지표(KPI) 대비 동등 이상의 성능이 확인되었는가? • AI 전용 시험 시나리오·악의적 입력 케이스 반영과 KPI 기준치 • 자동화 스크립트 커버리지·편향성 시험 결과	테스트시나리오	AG_SD44-1-14
10. AI 추론 서버 등에 대한 지속적 모니터링 체계 수립 여부	- AI 추론 서버 성능 저하 탐지 및 알람 발송 기능이 시험되었는가? • AI 추론 지표 모니터링 포함, 임계값 초과 시 이메일·SMS 등 알림 채널 설정·발송 테스트 수행 • 알림 수신자 담당자 지정과 모델 버전 변경 후 성능 비교 체계 - 지속적 모니터링 관리 계획이 구체적으로 수립되었는가? • AI 고유 KPI와 기준값·임계값의 수치화 및 모니터링 주기·이상 감지 알림·에스컬레이션 절차 구체성 • 재학습·재배포 트리거 조건과 모니터링 요건	모니터링/재학습	AG_SD44-1-15 AG_SD44-1-16
11. AI 모델 백업 및 버전 복구(롤백) 테스트 적정성	- AI 모델에 대한 백업 및 버전 복구(롤백) 테스트가 계획대로 수행되었는가? • 구성요소별 백업 계획·버전 관리 체계와 롤백 트리거 조건 정량화·테스트 시나리오·정상 동작 • 복구 목표 시간 달성 여부와 구성요소 버전 정합성	모델관리	AG_SD44-1-17

점검항목	검토항목	카테고리	관련 가이드
12. 서버 리소스 쿼터 통제 및 특정 프로세스가 GPU 메모리 독점을 차단하기 위한 설정 여부	- 리소스 쿼터 설정·초과 시 제한 동작, 단일 프로세스 GPU 메모리 독점 차단, 리소스 독점 방지 메커니즘 작동이 시험되었는가? • 컨테이너·Pod별 CPU·메모리·GPU 쿼터 설계대로 적용·초과 시 제한 동작 및 단일 프로세스 GPU 메모리 독점 차단 상한 설정·동작 검증 • 자원 독점 시나리오 부하시험 및 자원 통제 정책 대비 구현·시험 결과 일치성	GPU/NPU연산	AG_SD44-1-18
13. 운영자매뉴얼 작성 계획에 AI 특화 항목 (모델 재학습 절차, AI 장애 대응 등) 포함 여부	- AI 특화 항목을 포함한 운영자매뉴얼 작성 계획이 수립되었는가? • 모델 재학습 절차·AI 장애 대응·출력 모니터링·버전 관리/롤백 목차의 독립적 존재 및 일반 IT와의 구분	안정화/하자보수	AG_SD44-1-19
14. AI 구성요소 중 오픈소스 라이선스 분석 여부	- 오픈소스 소프트웨어(OSS)의 라이선스 유형이 상용 서비스 배포 조건과 충돌하지 않는지 검증되었는가? • SBOM에 전체 구성요소가 등재되고, 상업적 이용 제약·저작권 법무 검토 • 라이선스 의무사항 이행, SCA 도구 점검 범위 및 충돌 발견 시 해소 조치	저작권/라이선스	AG_SD44-1-20

라. 품질보증

점검항목	검토항목	카테고리	관련 가이드
1. 편향 발생 시 즉각 조치·차별 발언 차단 훈련 여부	- 편향 발생 시 즉각 조치 및 차별 발언 차단·모델 업데이트 훈련이 계획 이행되었는가? • AI 편향·차별 대응 관련 요구사항 명시 및 대응 훈련 일정·담당자 • 편향·차별 대응 훈련 계획 존재 및 이행 현황	편향/공정성	AG_SD44-4-01

점검항목	검토항목	카테고리	관련 가이드
2. AI 고유 결함 조치 및 독립된 주체에 의한 교차검증 적정성과 지속적인 신뢰성 평가 파이프라인 체계 수립 여부	- AI 편향·강건성·적대적 입력 신뢰성 테스트 및 자동화 평가 파이프라인 지속 검증 체계가 구축되었는가? • 서비스 위험 등급에 따른 신뢰성 평가 영역 설정 및 골든셋 등 평가 데이터셋 기반 적대적 공격·편향 테스트 수행 • 지속 평가 인프라 보유 및 자동화 검증 체계 구축	AI QA체계	AG_SD44-4-02

3.6. 전개단계

가. 시스템구조·보안

점검항목	검토항목	카테고리	관련 가이드
1. AI 특유의 파이프라인 보안 위험과 모델 배포 환경 보안 점검 여부	- CI/CD 배포 파이프라인이 실제 시스템으로 구현되어 정상 동작하는지, 빌드-테스트-배포 전 과정에 AI 모델 보안 점검이 자동으로 연동되는가? • 빌드-테스트-배포 전 과정 자동화 구현·모델 버전 관리·무결성 검증 자동화 및 탈옥·편향성·컨테이너 취약점 등 배포 전 보안 게이트 연동 • 배포 승인 절차·롤백 계획 수립과 파이프라인 접근통제·감사추적성	CI/CD파이프라인	AG_SD45-1-01
2. AI 서비스 전체 컨테이너 이미지에 대한 보안 점검 및 조치 확인 여부	- AI 서비스 전체 컨테이너 이미지 보안 점검이 수행되었는가? • AI 서비스 구성 전체 컨테이너 이미지의 악성코드 미검출·런타임 이미지 Critical/High CVE 조치 완료 및 AI 프레임워크·모델 서빙 엔진·GPU 드라이버 등 AI 특화 컴포넌트 취약점 점검 • CI/CD 파이프라인 내 보안 스캔 통합·미통과 시 배포 차단과 베이스 이미지 출처·무결성 검증	시큐어코딩	AG_SD45-1-02

점검항목	검토항목	카테고리	관련 가이드
3. 모델 드리프트 감지·재학습 조건 정의와, 감사추적 체계 구축 여부	- 모델 드리프트 감지 기준 및 재학습 트리거 조건이 측정 가능한 수준으로 정의되었는가? • 운영·유지보수 단계 책임 주체 식별 및 운영 SLA의 AI 품질 지표 임계값 정의 • 평가 데이터셋 보유 여부 및 갱신 계획의 적정성	모니터링/재학습	AG_SD45-1-03
	- 운영 로그 수집·보관 및 LLM 프롬프트·응답 이력 감사 추적 체계가 구축되었는가? • 법령별 로그 보존 의무 기간 충족 및 고위험 AI 자동 로그 기록 요건 충족 • 프롬프트·응답에 포함된 개인정보 보호 조치 및 프롬프트·토큰 등 AI 특화 로그 항목 수집 체계		AG_SD45-1-04
4. 인프라 용량의 적정성, 부하 시험 결과의 신뢰성, 오토스케일링 정책의 실효성 점검	- 인프라의 리소스 가용성과 부하 대응 체계의 적정성이 확인되었는가? • 예상 동시 사용자 수·피크 트래픽 데이터·인프라 환경과 AI 모델 운영 방식·SLO/SLA 목표 정의 • 전개 환경의 인프라 동일성	추론인프라	AG_SD45-1-05
5. AI 사고 유형별 대응방안이 포함된 AI 특화 운영자매뉴얼 완성도	- 운영단계에 발생할 수 있는 AI 사고 유형별 대응 절차가 수립되었는가? • 환각·편향·인젝션 등 AI 고유 사고 유형별 대응 절차 독립 기술과 Human-in-the-Loop 전환 기준·Kill Switch 절차·권한 체계 구체성 • 사고 탐지 임계값 정량화, 사고 기록·보고 체계 및 추적성	안정화/하자보수	AG_SD44-1-21

3.7. 실행통제단계

가. 사업관리

점검항목	검토항목	카테고리	관련 가이드
1. 라벨링 가이드 관리, 라벨링 품질 및 이력 추적관리 방안 적정성	- 라벨링 교차검수 규칙, 절차 등이 설계되었는가? • 교차 검수 상태 코드 체계·상태 전이 규칙과 검수 이력 테이블·반려 사유 코드·학습 데이터 필터링 연계·다중 검수자 IAA 구조 • LLM-as-Judge 결과 구분 기록·검수자 불일치 해결 흐름·라벨 품질 지표 저장 설계	라벨링	AG_PM04-2-01
2. 라벨링 교차검수 관리와 검수완료 데이터만 학습에 편입되는 필터링 체계	- 검수완료 데이터만 학습 데이터로 편입되는 필터링이 동작하는가? • 라벨링 상태 코드 체계의 구현과 검수완료 상태만 추출되는 필터링 동작 및 교차 검수 일치율 산출·기록 • 라벨링 이력 관리와 반려 데이터 재라벨링·재검수 프로세스 및 역할별 상태 변경 권한 분리	라벨링	AG_PM04-2-02
3. 라벨러 구성 다양성 및 라벨러 교육 실행 여부	- 라벨러 다양성 목표 달성 및 교육으로 라벨링 편향이 방지되었는가? • 라벨러 다양성 목표 달성과 연령대·성별 분포 균형성 및 전문 도메인 라벨러의 자격·경력 요건 충족 • 라벨러 간 일치도 검증과 라벨링 가이드라인·교육 이수 문서화 여부	라벨링	AG_PM04-2-03
4. 외부 데이터 출처 명시·라이선스 기록 및 저작권 위반 방지 요건 정의 여부	- 외부 데이터 출처·라이선스 기록과 저작권 검토·라이선스 위반 방지 요건이 정의되었는가? • 외부 데이터셋의 명칭·출처·버전·라이선스·귀속 의무 기록과 AI 학습 허용 범위·이용약관·robots.txt·계약 조항 검토 계획	저작권/라이선스	AG_PM04-2-04

점검항목	검토항목	카테고리	관련 가이드
5. 사용자매뉴얼에 AI 특화 요소가 반영 여부 및 인수인계 절차 구체화 여부	- 사용자매뉴얼에 AI 특화 항목이 포함되도록 설계되었는가? • AI 한계·안전 사용 관련 항목의 요구사항 식별 및 매뉴얼 목차 설계 초안	교육훈련기술이전	AG_PM04-2-05
6. 하도급 업체 보안점검 및 데이터 반출입 통제 적정성	- 하도급 업체 보안점검 절차·데이터 반출입 통제·NDA·최소권한·보안사고 대응이 구체화되고 추적되는가? • 하도급 보안점검 절차·NDA·제재 절차, 데이터 유형별 반출입 승인·암호화·최소권한 접근통제 및 로그·이상행위 탐지 체계 • 보안사고 대응 절차 및 요구사항-설계 간 추적 매핑	보안관리	AG_PM04-2-06

3.8. 종료단계

가. 사업관리

점검항목	검토항목	카테고리	관련 가이드
1. XAI 기능의 유용성·수용성에 대한 평가단 평가 결과 적정성	- XAI 평가단 평가 결과가 완결성 있게 작성·확보되었는가? • XAI 기능의 유용성·사용자 수용성 평가 결과 완결성 및 평가단 평가 계획 • 평가단 구성의 사전 계획 부합	AI윤리/xAI	AG_PM04-3-01
2. AI 윤리기준 최종 검토가 절차에 따라 완료 여부	- AI 윤리기준 최종 검토가 진행되었는가? • AI 윤리기준 최종 검토 절차 이행·결과 반영 및 권한 있는 서명자의 최종 서명 • 편향성·공정성 최종 평가 결과의 허용 기준 충족, AI 윤리원칙 체크리스트 완결성, 저작권·법규 준수 검토 및 HITL 설계와의 일치성	법령/AI윤리	AG_PM04-3-02

점검항목	검토항목	카테고리	관련 가이드
3. AI 하도급 업체 보안점검 및 보안관리 적정성	- AI 관련 하도급 업체 보안관리가 적정하게 수행되었는가? • AI 하도급 업체 보안점검 수행 및 데이터·모델 취급자의 보안 서약 징구 • 데이터·모델 반출입 통제 절차·이력 대장 관리 적정성 및 학습데이터 내 개인정보 사전 비식별화 처리	보안관리	AG_PM04-3-03
4. AI 특화 요소를 반영한 인수테스트 시행과 인수 조건 충족 여부	- AI 특화 인수기준에 따라 인수테스트가 적정하게 수행되고, 결함 처리가 완료되었는가? • AI 특화 인수기준의 인수테스트 시나리오 수치 반영·현업 참여 및 정량·정성 평가·결함 처리·재테스트 • 발주기관·현업 부서장의 최종 서명, 모델 버전-배포 형상 일치 및 운영 이관 문서	시범운영/인수테스트	AG_PM04-3-04

CHAPTER

4

개별 감리 점검 가이드

4. 개별 감리 점검 가이드

본 장은 195개 점검 가이드를 감리영역·생명주기 단계별로 수록한 것으로, 각 가이드는 목적·필요성·검토방법·점검결과 판정기준·감리 시 유의사항 및 실무 Tip의 5부로 구성된다.

각 가이드의 첫머리에는 식별용 머리글이 두 줄의 띠 형태로 제시된다. 윗줄(제목 띠)은 「관리코드 | 가이드명」으로 구성되고, 아랫줄(속성 띠)은 「감리영역 | 감리단계 | AI 생명주기 | 카테고리」의 네 가지 속성을 차례로 표기한다.

속성 띠에서 감리영역은 사업관리, 품질보증, 응용시스템, 데이터베이스, 시스템구조·보안의 5개 영역 중 하나를, 감리단계는 요구정의·분석부터 전개에 이르는 감리 단계(사업관리는 착수·실행통제·종료)를, AI 생명주기는 해당 점검이 대응하는 생성형 AI 생명주기 단계(생명주기 관리, 데이터 수집·처리, 인공지능 모델 개발, 시스템 구현, 운영·모니터링)를, 카테고리는 편향/공정성, AI특화보안 등 세부 점검 주제를 나타낸다.

※ 산출물 표기 규약: 본 장에서 {중괄호}로 표기한 산출물은 사업에서 공식적으로 정의된 산출물(공식 산출물)을, '작은따옴표'로 표기한 산출물은 사업 수행 과정에서 작성·활용되는 작업(Working) 산출물을 의미한다.

※ 각 가이드에 부여된 코드(SD4·PM04)의 구성 체계는 2장의 '감리 점검 가이드 코드 체계' 절을 참조한다.

▣ 수록 현황

감리영역	건수	단계 구성
사업관리	20	착수 10 · 실행통제 6 · 종료 4
품질보증	7	요구정의 5 · 시험 2
응용시스템	49	요구정의 15 · 설계 16 · 구현 6 · 시험 12
데이터베이스	45	요구정의 11 · 설계 11 · 구현 10 · 시험 13
시스템구조·보안	74	요구정의 24 · 설계 19 · 구현 5 · 시험 21 · 전개 5
합계	**195**	**5개 영역 · 생명주기 전 단계**

4.1. 사업관리

착수부터 종료까지 사업 전반의 위험관리, 하도급·보안통제, 인수 점검 가이드를 수록한다. (총 20건)

착수단계 (10건)

AG_PM04-1-01 KPI 점검가이드라인

AG_PM04-1-02 이해관계자 매핑

AG_PM04-1-03 AI 제약사항점검

AG_PM04-1-04 위험관리계획수립

AG_PM04-1-05 AI 리스크 분류 체계 점검요건

AG_PM04-1-06 데이터 라벨링 도급관리

AG_PM04-1-07 데이터 라벨링 요구정의

AG_PM04-1-08 라벨링데이터 교차검수 요건 및 다수 검수자 검증 프로세스 반영

AG_PM04-1-09 라벨링 편향방지 및 주관적 판단 배제 라벨링 기준 요건

AG_PM04-1-10 데이터 수집 편향배제 및 특정 출처 편중방지 수집 요건

실행통제단계 (6건)

AG_PM04-2-01 라벨링 교차검수 상태값 설계점검

AG_PM04-2-02 라벨링 교차검수 상태코드 매핑점검

AG_PM04-2-03 라벨러 분포 인적 다양성 확보결과 점검

AG_PM04-2-04 외부 데이터 출처 명시·라이선스 기록 및 저작권 위반 방지

AG_PM04-2-05 사용자매뉴얼 목차 설계

AG_PM04-2-06 하도급 보안통제

종료단계 (4건)

AG_PM04-3-01 평가단 XAI 평가결과

AG_PM04-3-02 윤리기준 최종검토

AG_PM04-3-03 하도급업체 보안점검

AG_PM04-3-04 이해관계자 인수테스트

AG_PM04-1-01 | KPI 점검가이드라인

사업관리 | 착수단계 | 생명주기 관리 | 사업타당성/ROI

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · NIST AI RMF MEASURE · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, {사업수행계획서}에 기술된 사업목표 달성 지표(KPI) 적정성과 AI 시스템 도입 효과 지표 명확성 여부를 감리원이 제3자 관점에서 점검하기 위한 구체적인 절차와 기준을 제공하는 것을 목적으로 한다.

이를 통해 요구정의 단계에서 측정 가능한 KPI와 도입 효과 지표가 사전에 수립되었는지를 확인함으로써, 향후 검사 단계에서의 분쟁을 예방하고, 사업 목표의 달성 여부를 객관적으로 검증할 수 있는 기반이 확립되었는지를 점검하는 것이 핵심이다.

2. 필요성

▶ AI 시스템은 일반 소프트웨어와 달리 동일 입력에도 출력 결과가 달라질 수 있는 확률적 특성을 가지며, 모델 성능의 가변성으로 인해 비전문가가 직관적으로 품질을 판단하기 어렵다. 이로 인해 KPI 없이 사업이 추진될 경우, 검수 단계에서 "AI가 잘 동작한다"는 모호한 기준으로 합격 여부를 다투는 분쟁이 발생할 수 있다.

▶ 도입 효과 지표가 명시되지 않으면 AI 투자의 타당성을 사후에 검증할 근거가 없어지며, 과도한 예산 집행과 사업 실패에 대한 책임 소재가 불분명해지는 리스크가 발생한다.

▶ 특히 생성형 AI의 경우, 초기 단계에서 "무엇을 얼마나" 달성해야 하는지가 정의되지 않으면 검사 기준 자체가 성립되지 않는다는 점에서, 요구정의 단계에서 측정 가능한 KPI와 도입 효과 지표를 사전에 수립하는 것이 필수적이다.

3. 검토방법

No	점검 항목	확인 방법
1	KPI 수치 목표값(Target) 정의 여부	{사업수행계획서}의 KPI 항목별 목표값이 수치와 단위로 정의되었는지 확인. '향상'·'개선' 등 정성 표현만 있으면 미흡으로 판정
2	KPI 기준값(Baseline)의 현황 데이터 근거 설정 여부	각 KPI의 기준값이 현황 측정 데이터에 근거하여 산출되었는지 확인하고 산출 근거 자료의 출처·측정 시점을 대조
3	KPI 측정 방법 및 측정 주기 명시 여부	KPI별 측정 방법(산식·데이터 원천)과 측정 주기가 {사업수행계획서}에 명시되었는지 확인

No	점검 항목	확인 방법
4	AI 특화 KPI 포함 여부	정확도·오탐률·응답시간·환각 발생률 등 AI 특화 KPI가 KPI 목록에 포함되었는지 확인
5	비기능 요구사항 내 AI 성능 기준과 품질목표 일치 여부	비기능 요구사항의 AI 성능 기준과 {사업수행계획서}의 품질목표를 대조하여 수치가 일치하는지 교차 확인
6	KPI와 사업 목적 간 인과 관계 명확성	각 KPI가 사업 목적·기대효과와 논리적으로 연결되는지 목적 기술문과 대조하여 확인
7	AI 도입 효과 지표(AS-IS/TO-BE)의 정량 제시 여부	AI 도입 효과 지표가 AS-IS와 TO-BE 수치로 각각 제시되고 산출 근거가 명시되었는지 확인
8	도입 효과 지표의 단계별 구분 여부	도입 효과 지표가 구축 완료 시점·운영 3개월 등 단계별로 구분되어 목표값이 제시되었는지 확인
9	정성적 효과 항목의 측정 방법 제시 여부	만족도 등 정성적 효과 항목에 대해 설문 등 측정 방법과 표본 기준이 제시되었는지 확인
10	KPI 목표값의 현실적 달성 가능성	유사 공공 AI 사업의 성과·벤치마킹 자료와 대조하여 KPI 목표값이 현실적으로 달성 가능한 수준인지 확인
11	AI 특화 KPI의 수치 목표 2개 이상 정의 여부	AI 정확도·자동화율·응답 품질 만족도·인간 검토 감소율 등 AI 특화 KPI가 수치 목표와 함께 2개 이상 정의되었는지 확인
12	ROI 산정 공식 및 측정 일정 명시 여부	ROI 산정 공식과 측정 일정이 명시되었는지 확인하고 인건비 절감·처리 비용 감소·오류 감소 비용의 정량화 방법론을 사업자 인터뷰로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	KPI 목표값이 수치와 단위로 정의되고, 기준값(Baseline)의 산출 근거와 측정 방법·주기가 명시되며, AI 특화 KPI(정확도·오탐률·응답시간·환각 발생률 등)가 2개 이상 수치 목표와 함께 포함되고, 사업 목적과의 인과 관계 및 유사 사업 대비 달성 가능성이 확인되며, AS-IS/TO-BE 도입 효과 지표가 단계별로 정량 제시되고 정성적 효과의 측정 방법과 ROI 산정 공식·측정 일정까지 {사업수행계획서}에 명시된 경우	현행 유지 (수용 기준으로 활용 가능 여부 재확인)
보완권고 (△)	KPI가 추상적으로 기술되어 있거나, 측정 방법·기준값이 불명확하거나, 도입 효과 지표가 누락된 경우	해당 항목 보완 요청 후 재확인 (구체적인 수치화 방안 및 보완 기한 명시)

판정	판정 기준	주요 사례 / 조치
부적합 (X)	KPI 또는 도입 효과 지표가 전혀 정의되지 않아 사업 목표 달성 여부를 객관적으로 검증할 기반이 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) KPI 수치의 근거를 반드시 확인하라: KPI 목표값이 있더라도 그것이 현황 데이터(AS-IS)에 근거한 것인지를 반드시 확인해야 한다. "정확도 95% 이상"이라고 기재되어 있어도 현재 정확도가 40%라면 사실상 달성 불가능한 목표이며, 이는 사업 전체의 리스크 요인이 된다. 기준값(Baseline) 없는 목표값은 미흡으로 판정하고 조치를 요구한다.

2) AI 특화 KPI 누락 여부에 주목하라: AI 시스템에는 일반 IT KPI(가용성, 응답시간 등)와 함께 AI 고유의 KPI(모델 정확도, 오탐률, 편향성 지표, 생성형 AI의 경우 환각 발생 비율, 유해 표현 비율 등)가 반드시 포함되어야 한다. AI 특화 KPI가 누락된 경우, 이는 향후 검사 단계에서 AI 품질 판단 기준이 없음을 의미하므로 반드시 조치 요구 사항으로 기록한다.

3) 모호한 표현을 수치화하도록 유도하라: 추상적으로 기술된 KPI는 측정 가능한 수치 기준으로 구체화하도록 유도하고, 점검 시 형식 충족 여부에 그치지 말고 실질적 이행 근거와 책임자 서명 여부까지 확인한다.

AG_PM04-1-02 | 이해관계자 매핑

사업관리 | 착수단계 | 생명주기 관리 | 의사소통

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · 「장애인차별금지 및 권리구제 등에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템을 실제로 사용·개발·운영할 다양한 이해관계자의 요구사항을 체계적으로 식별하고, 이를 {요구사항정의서}에 구체적이고 완전하게 반영하였는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

특히, 이해관계자 분석이 부실하거나 특정 집단(취약계층 사용자, 운영자 등)의 요구사항이 누락될 경우, 이후 설계·개발 단계에서 요건 변경에 따른 비용 증가, 서비스 오류에 대한 책임 공백, AI 편향성 문제 등이 발생할 위험이 있다.

본 가이드라인은 이러한 위험을 요구정의 단계에서 사전 확인하는 기준을 제공한다.

2. 필요성

▶ AI 서비스는 동일한 시스템이라 하더라도 사용자·개발자·운영자의 역할과 요구가 명확히 구분되어야 한다. 생성형 AI 서비스의 경우 학습 데이터 편향, 환각(Hallucination) 등 AI 고유의 오류 특성으로 인해 이해관계자별 요구사항을 세밀하게 정의하지 않으면 검수 기준의 모호성, 운영 중 책임 소재 불명확 등의 분쟁이 발생할 수 있다.

▶ 사용자 요구의 다양성: 동일한 AI 서비스라도 일반 시민, 내부 담당자, 고령자·장애인 등 취약계층의 기대 수준과 오류 허용 범위가 상이하다. 이를 구분하지 않고 일괄 정의하면 특정 집단의 요구가 반영되지 않아 서비스 차별 문제가 발생한다.

▶ 개발자 AI 특화 요건 부재 위험: AI 모델 학습·검증 요건, 편향성 점검 기준, 성능 지표가 요구사항에 정량적으로 명시되지 않으면 검수 단계에서 분쟁이 발생한다.

▶ 운영자 요구 누락 위험: AI 시스템은 배포 후에도 모델 성능 저하(드리프트), 비정상 출력 발생 등 지속적인 관리가 필요하다. 운영자의 모니터링·개입 요건이 요구정의 단계에서 정의되지 않으면 운영 단계에서 대응 체계가 공백 상태가 된다.

▶ 이해관계자 참여 구조 미비: TTA 신뢰할 수 있는 인공지능 개발안내서 및 NIST AI RMF는 AI 시스템 기획 단계에서 다양한 이해관계자의 요구사항을 수렴하도록 명시하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	직접 사용자 유형의 요구사항정의서 명시 여부	일반 시민·내부 직원·의사결정자 등 AI 서비스의 직접 사용자 유형이 {요구사항정의서}에 구체적으로 명시되었는지 확인
2	사용자별 특성 분석 및 요구사항 반영 여부	이용 목적·디지털 리터러시 수준·취약계층 여부 등 사용자별 특성이 분석되어 {요구사항정의서} 및 '이해관계자 분석서'에 반영되었는지 확인
3	AI 출력 결과 신뢰성 요건 정의 여부	결과 설명 방식·오류 시 고지 방안 등 AI 출력 결과 신뢰성 요건이 {요구사항정의서}에 정의되었는지 확인하고 사업자 면담으로 보완
4	이의제기·정정 요청 절차 포함 여부	AI 출력에 대한 사용자의 이의제기·정정 요청 절차가 {요구사항정의서}에 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 서비스의 직접 사용자 유형이 {요구사항정의서}에 구체적으로 명시되고, 사용자별 특성(이용 목적, 디지털 리터러시 수준, 취약계층 여부 등) 분석 결과가 요구사항에 반영되었으며, AI 출력 결과 신뢰성 요건과 이의제기·정정 요청 절차가 모두 정의된 경우	현행 유지
보완권고 (△)	이해관계자 유형은 명시되었으나 사용자별 특성 분석이 미흡하거나, AI 출력 신뢰성 요건·이의제기 절차 등 일부 항목이 누락되거나 형식적으로만 작성된 경우	해당 항목 보완 요청 후 재확인 (담당자 면담 후 보완 권고 의견 작성)
부적합 (X)	이해관계자 식별·분석 내용이 {요구사항정의서}에 전혀 반영되지 않았거나, '이해관계자 분석서' 등 관련 산출물이 미제출되어 이해관계자별 요구사항 반영 여부를 확인할 수 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 이해관계자 분석의 형식적 작성 경계: 사업자가 '이해관계자 분석서'를 제출하더라도, 이해관계자가 단순히 나열만 되고 각 집단의 AI 서비스 관련 특수 요구(예: 취약계층 접근성, 운영자 긴급 개입 조건 등)가 요구사항으로 연결되지 않은 경우에는 형식적 작성으로 판단하여 보완 권고한다. 핵심 확인 기준은 이해관계자 분석 결과가 {요구사항정의서}의 특정 요구사항 항목에 실질적으로 반영되었는지 여부이다.

2) AI 성능 요건의 수치화 유도: 성능 지표는 전체 평균이 아닌 소수 클래스·엣지 케이스 기준도 반드시 포함하여 점검한다. "AI가 정확하게 답변해야 한다", "적절한 수준의 응답을 제공해야 한다" 등 모호한 표현이 발견된 경우, 측정 가능한 기준(예: 정확도 90% 이상, 집단별 성능 편차 5% 이하, 응답 거부 시 사유 고지율 95% 이상)으로 구체화하도록 보완 권고한다.

3) 취약계층 사용자 누락 여부 집중 확인: 고령자·장애인 등 취약계층 사용자의 요구사항이 누락되면 특정 집단에 대한 서비스 차별 문제로 이어질 수 있으므로 누락 여부를 집중 확인하고, 발견 사항은 구체적 근거와 함께 감리보고서에 기록한다.

AG_PM04-1-03 | AI 제약사항점검

사업관리 | 착수단계 | 생명주기 관리 | AI 위험관리

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」·「개인정보 보호법」·「전자정부법」·「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 감리원이 예비조사를 수행할 때 두 가지를 체계적으로 점검하는 것을 목적으로 한다.

첫째, {감리계획서} 수립 과정에서 AI 시스템 특유의 점검 제약 사항(학습 데이터 비공개, 외부 기반모델 검증 한계, 법적 규제 범위의 불확실성 등)이 식별되어 명시되었는지, 그리고 이에 대한 대체 점검 방법이 계획에 반영되었는지를 확인한다.

둘째, 발주기관이 제시한 요건(RFP)과 수행기관이 정의한 {요구사항정의서} 간의 AI 관련 요건 정합성을 검토하여, AI 성능·안전성·신뢰성 요구사항이 측정 가능한 기준으로 구체화되었는지를 확인한다.

이를 통해 본 감리 수행 시 발생할 수 있는 점검 불가 항목 관련 분쟁, 산출물 미확보에 따른 점검 공백, 모호한 요구사항으로 인한 검사 단계 분쟁 등의 위험을 사전에 예방하는 데 최종 목적이 있다.

2. 필요성

▶ 일반 정보시스템과 달리 AI 서비스는 감리 수행 과정에서 고유한 제약이 존재하며, 이를 예비조사 단계에서 사전 파악하지 않으면 감리의 실효성이 크게 저하된다.

▶ 학습 데이터 접근 제한: AI 모델의 성능은 훈련 데이터 품질에 직접 의존하나, 학습 데이터가 영업 비밀·개인정보 등의 사유로 비공개되어 감리원이 직접 검토하지 못하는 경우가 빈번하다.

▶ 외부 기반모델(LLM) 검증 한계: OpenAI, Anthropic 등 외부 클라우드 기반 LLM을 활용하는 사업의 경우, 모델 내부 구조가 블랙박스(Black-box)로 제공되어 감리원의 직접 검증이 불가하다.

▶ 요구사항 모호성으로 인한 검사 단계 분쟁: "AI가 적절한 답변을 제공한다"와 같이 검증 기준이 불명확한 요구사항이 감리 착수 전 수치화되지 않으면, 준공 검사 단계에서 이행 여부에 대한 분쟁이 발생할 위험이 높다.

▶ 법적·윤리적 요건의 불완전한 반영: 인공지능법, 개인정보 보호법 등 AI 특유의 법적 요건이 요구사항 정의 단계에서 누락될 경우, 후속 설계 및 개발 단계에서의 수정 비용이 크게 증가한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 관련 법적·기술적 제약사항 식별 여부	개인정보 보호법·저작권법·외부 AI API 이용약관 등 법적·기술적 제약이 RFP 또는 {요구사항정의서}에 식별·기재되었는지 확인
2	외부 기반모델(LLM) 데이터 보안 정책 반영 여부	LLM·SaaS형 AI API 사용 시 제공자의 데이터 보안 정책(입력 데이터 저장·학습 여부 등)이 요구사항 및 계획서에 반영되었는지 확인
3	점검 불가 항목의 {감리계획서} 명시 및 협의 여부	학습 데이터 원본 비공개·모델 내부구조 검증 불가 등 기술적 점검 한계 항목이 {감리계획서}에 명시되고 발주기관과 협의된 근거(회의록·협의서 등)가 있는지 확인
4	점검 불가 항목 대체 검증 방법 명시 여부	직접 점검이 불가한 항목에 대해 인터뷰·확인서 징구·테스트 결과 검토·외부 인증서 확인 등 대체 검증 방법이 {감리계획서}에 구체적으로 명시되었는지 확인
5	AI 성능 요구사항의 측정 가능성	'적절한 답변 제공' 등 모호한 표현이 '정확도 X% 이상'·'응답 시간 X초 이내'·'금칙어 미발생률 99% 이상' 등 측정 가능한 기준으로 정의되었는지 확인
6	AI 비기능(신뢰성) 요구사항 수용 기준 명시 여부	공정성(편향 방지)·투명성(설명 가능성)·개인정보 보호 등 AI 신뢰성 요구사항의 수용 기준이 {요구사항정의서}에 구체적으로 기재되었는지 확인
7	학습·테스트 데이터 요건 검토 여부	학습 데이터의 출처·수집 방법·품질 기준·개인정보 포함 여부 및 처리 방안이 검토되고 계획서에 반영되었는지 확인
8	인간 감독(HITL) 절차 명시 여부	AI 산출물에 대한 인간 검토 절차·의사결정 개입 방식·오류 발생 시 처리 프로세스 등이 {요구사항정의서}에 명시되었는지 확인
9	RFP와 {요구사항정의서} 간 AI 요건 정합성	RFP에 명시된 AI 요건이 {요구사항정의서}에 누락 없이 분석·상세화되었는지 두 문서를 대조하고, 생성형 AI의 경우 환각 방지·프롬프트 인젝션 대응 요건 포함 여부를 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	① AI 제약 사항(기술적·법적)과 외부 기반모델의 데이터 보안 정책이 {감리계획서} 및 요구사항에 명시되고, ② 점검 불가 항목이 {감리계획서}에 기재되어 발주기관과 협의된 근거 문서(회의록, 협의서 등)와 대체 검증 방법이 존재하며, ③ AI 성능 요구사항이 측정 가능한 기준으로 정의되고, ④ 공정성·투명성·개인정보 보호 등 신뢰성 요구사항의 수용 기준, 학습·테스트 데이터 요건, 인간 감독(HITL) 절차가 {요구사항정의서}에 기재되며 RFP와의 AI 요건 정합성이 확인된 경우	현행 유지
보완권고 (△)	① AI 제약 사항이 {감리계획서}에 일부만 언급되거나, ② 협의 근거 문서가 불명확하거나, ③ AI 성능 요구사항이 일부 모호한 표현으로 기재된 경우	시정·보완 권고 후 차회 감리 전 보완 여부 재확인
부적합 (X)	① AI 제약 사항 관련 내용이 {감리계획서}에 전혀 없거나, ② 협의 절차가 수행되지 않았거나, ③ RFP와 요구사항 정의서 간 AI 요건의 중요한 누락이 발생한 경우	시정 요구(즉시 조치 필요) 후 이행 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 AI 요건의 수치화 유도: "AI는 적절한 답변을 해야 한다"와 같이 모호한 문구는 향후 이행 여부를 둘러싼 분쟁의 원인이 된다. 감리원은 이러한 모호한 요건을 발견하면 "금칙어 목록 1,000개 이상 반영", "유해 콘텐츠 차단율 99% 이상", "정상 응답률 95% 이상" 등 측정 가능한 검사 기준으로 구체화할 것을 발주기관과 PM에게 서면으로 권고한다. 이때 감리원은 수치 자체를 결정하는 것이 아니라, 수치화의 필요성과 기준 마련을 촉구하는 역할임을 명심한다.

2) 점검 불가 항목의 협의 및 문서화 선행: AI 사업에서는 학습 데이터 원본 미열람, 외부 LLM 내부 구조 미검증 등 일반 IT 감리와 달리 점검 자체가 불가한 항목이 반드시 존재한다. 이를 {감리계획서} 수립 시 발주기관과 명확히 협의하고 그 근거를 회의록 또는 협의서로 남겨야 한다. 사전 협의 없이 감리 수행 중 점검 불가 항목이 발견될 경우 발주기관의 불신을 초래하거나 감리 결과의 신뢰도를 훼손할 수 있다.

3) 외부 AI API 이용 시 계약 관계 및 데이터 주권 확인: OpenAI, Claude, Google Gemini 등 외부 API를 활용하는 사업에서 사용자 입력 데이터가 외국 서버에 저장·학습되는 경우, 개인정보의 국외 이전 요건과 데이터 보안 정책 준수 여부가 요구사항에 반영되었는지 확인이 필요하다. 감리원은 기술적 판단 전에 해당 AI API 서비스의 이용약관 또는 데이터 처리 계약서(DPA) 존재 여부를 먼저 확인한다.

4) RFP 원본과 요구사항 정의서를 반드시 병행 검토: {요구사항정의서}만 단독으로 검토하면 RFP에서 요구된 AI 요건이 누락되었는지 파악하기 어렵다. '제안요청서(RFP)'와 {요구사항정의서}를

병행 대조하여 AI 관련 요건의 추적 가능성(traceability)을 확인하는 것이 실무적으로 효과적이며, 요건 추적 행렬(RTM)이 작성된 경우 이를 우선 활용한다.

5) 감리원 역할 경계 준수 – 기술 자문과 감리의 구분: 감리원은 AI 요구사항의 적정성을 점검하는 역할이며, AI 모델 선정이나 아키텍처 설계를 직접 지시하는 것은 감리 역할 범위를 초과한다. "A 모델을 사용하라"가 아닌 "성능 요구사항에 대한 검증 기준이 명시되어 있어야 한다"는 방식으로 점검 의견을 제시하는 것이 적절하다.

AG_PM04-1-04 | 위험관리계획수립

사업관리 | 착수단계 | 생명주기 관리 | AI 위험관리

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」제32조 · NIST AI RMF MANAGE · 「개인정보 보호법」 · 「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템의 생명주기 전반에 걸쳐 발생할 수 있는 위험을 사전에 식별·평가하고, 이에 대한 대응 방안을 수립하였는지 여부를 감리원 관점에서 제3자적으로 점검하는 것을 목적으로 한다.

특히, AI 시스템 고유의 위험(환각, 편향, 개인정보 침해, 보안 취약점, 저작권 문제 등)에 대해 위험 식별(Identification) → 위험 분석·평가(Analysis/Evaluation) → 위험 대응(Treatment)의 체계적 3단계 프로세스가 {요구사항정의서} 및 'AI위험관리계획서'에 구체적으로 반영되었는지를 확인한다.

단순한 형식적 기술이 아니라, 각 단계가 실질적 내용을 갖추고 있는지를 중점적으로 점검한다.

2. 필요성

▶ AI 시스템은 데이터 기반 의사결정 특성상 일반 소프트웨어와 다른 고유한 위험을 내포하고 있으며, 이를 기획·분석 단계부터 체계적으로 관리하지 않으면 시스템 운영 중 심각한 문제가 발생할 수 있다. 감리원 관점에서 위험관리 계획의 사전 수립이 필요한 이유는 다음과 같다.

▶ AI 고유 위험의 조기 식별 필요성: 환각(Hallucination), 데이터 편향, 개인정보 유출, 적대적 공격(Adversarial Attack), 저작권 침해 등 AI 특유의 위험은 개발 완료 후 발견 시 수정 비용이 기하급수적으로 증가한다. 감리원은 요구정의 단계에서 이러한 위험이 반드시 식별되었는지를 확인해야 한다.

▶ 사업 전체 품질 보증 기반 마련: 위험관리 계획이 사전에 수립되지 않으면, 이후 설계·개발·시험 단계에서 동일한 위험 요소가 반복 재발하여 사업 일정과 비용에 직접적 영향을 미친다. 감리원은 위험관리 계획이 사업 수행 계획과 연계되어 있는지를 확인한다.

▶ 법적·규제 준수 의무 이행 확인: 공공 AI 사업은 개인정보 보호법, AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법) 등 관련 법규 준수가 요구되며, 위험관리 계획에 법령 준수 항목이 포함되지 않을 경우 감리 지적 대상이 된다.

▶ 발주기관 의사결정 지원: 위험 파급효과가 매우 큰 경우 AI 기술 적용 자체를 재검토해야 하며, 이를 위한 근거가 요구정의 단계에서 확보되어야 발주기관이 적시에 의사결정을 할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	심각도(Severity) 등급 기준 정의 여부	높음/중간/낮음 또는 1~5단계 등 위험 심각도 등급 기준이 'AI위험관리계획서'에 정의되었는지 확인
2	발생 가능성(Likelihood) 기준 정의 여부	위험 발생 가능성 판정 기준이 'AI위험관리계획서'에 정의되었는지 확인
3	위험 크기(Risk Level) 산출 방식 명시 여부	심각도×발생가능성 매트릭스 등 위험 크기 산출 방식이 'AI위험관리계획서'에 명시되었는지 확인하고 담당자 면담으로 보완
4	AI 특성 고려 추가 척도 반영 여부	편향성 영향 집단 규모·설명가능성 수준 등 AI 특성을 고려한 추가 척도가 'AI위험관리계획서'에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'AI위험관리계획서'에 심각도·발생가능성 등급 기준과 위험 크기 산출 방식(심각도×발생가능성 매트릭스), AI 특성을 고려한 추가 척도(편향성 영향 집단 규모, 설명가능성 수준 등)가 모두 구체적 수치와 함께 정의된 경우	현행 유지
보완권고 (△)	위험 평가 기준은 정의되었으나 AI 특화 척도 누락, 위험 매트릭스 수치 미기재, 대응 방안의 구체성 부족 등 일부 요소가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	위험관리 계획이 수립되지 않았거나, 'AI 위험관리를 수행한다'는 선언적 문구만 있고 심각도·발생가능성 등급 기준과 위험 크기 산출 방식이 전혀 정의되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'AI위험관리계획서' 부재 시 형식보다 내용 실질성 우선: 독립 문서로 제출되지 않더라도 {요구사항정의서} 또는 {사업수행계획서}에 위험 식별·평가·대응 방안이 실질적으로 기술되어 있다면 형식적 문서명 불일치를 이유로 부적합 처리하지 않는다. 반대로 문서가 존재하더라도 선언적 문구만 있고 구체적 위험 목록·평가 결과·대응 방안이 없으면 부적합으로 판정하고 즉시 보완을 요청한다.

2) AI 고유 위험과 일반 IT 위험을 반드시 구분하여 점검: 위험관리 계획에 일반 소프트웨어 위험(일정 지연, 인력 이탈, 서버 장애 등)만 나열되고 AI 고유 위험(환각, 편향, 개인정보 유출, 프롬프트 인젝션, 저작권 침해 등)이 누락된 경우는 보완권고 또는 부적합으로 처리한다. 위험 목록에서 AI 관련 항목의 존재 여부를 먼저 확인하는 것이 실무적으로 효율적이다.

3) 위험 매트릭스에 수치가 채워져 있는지 반드시 확인: 심각도×발생가능성 매트릭스 양식만 존재하고 개별 위험 항목에 대한 평가 수치가 비어 있으면 실질적 위험 평가가 수행되지 않은 것이므로, 평가 결과 기재를 요구하고 이행 여부를 재확인한다.

AG_PM04-1-05 | AI 리스크 분류 체계 점검요건

사업관리 | 착수단계 | 생명주기 관리 | AI 위험관리

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」제33조·제35조 ·「인공지능 및 데이터 기반 행정 활성화에 관한 법률」· EU AI Act

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 구축하고자 하는 AI 시스템의 위험 수준을 분류하고 위험 등급에 따른 관리·규제 요건을 사전에 식별하여 {사업수행계획서} 및 과업 범위에 반영하였는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 '리스크 분류 매트릭스', '법무 검토 의견서' 또는 '전문가 자문서', {사업수행계획서}를 근거로 EU AI Act 부속서Ⅲ 고위험 분야와 AI 기본법 고영향 AI의 해당 여부, 금지 AI 적용 가능성, 외부 검토의 독립성, 분류 결과의 과업·예산 반영을 확인한다.

2. 필요성

▶ EU AI Act(2024.8 발효, 고위험 의무 단계적 적용)는 AI 시스템을 위험도에 따라 분류하고 고위험 AI에 데이터 거버넌스·기술문서·인간 감독 등 추가 요건을 부과한다.

▶ AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 2026.1.22 시행)은 고영향 AI에 대한 사업자 의무를 명시하며, 그 판단 기준과 적용 범위는 EU AI Act의 고위험 분류와 동일하지 않으므로 각각 검토해야 한다.

▶ 기획 단계에서 적절한 위험 분류가 이루어지지 않으면 이후 설계·구현 단계에서 법적 컴플라이언스 문제가 드러나며, 이 시점에는 과업 범위와 예산의 재조정이 불가피하다.

▶ 위험 등급별로 데이터 거버넌스·모니터링·인간 감독 의무가 다르므로 분류 결과는 사업 범위·산출물 목록·예산 산정에 직접 영향을 미친다. 분류만 하고 과업에 반영하지 않으면 형식적 절차에 그친다.

▶ 위험 등급이 올라가면 의무와 비용이 함께 증가하므로 사업자에게는 낮게 분류할 유인이 있다. 외부의 독립적 검토 없이는 분류 결과의 객관성을 담보할 수 없다.

3. 검토방법

No	점검 항목	확인 방법
1	EU AI Act 부속서Ⅲ 고위험 분야 적용 여부	'리스크 분류 매트릭스'에서 EU AI Act 부속서Ⅲ의 각 분야와 대상 AI 시스템을 대조하여 해당 여부와 판단 근거 조항이 기재되었는지 확인

No	점검 항목	확인 방법
2	AI 기본법 '고영향 AI' 해당 여부	'리스크 분류 매트릭스'에서 AI 기본법의 고영향 AI 판단 기준과 대조한 결과가 EU AI Act 검토와 분리되어 기재되었는지 확인
3	수용 불가(금지) AI 적용 가능성	EU AI Act 제5조 각 호와 대조한 표가 '리스크 분류 매트릭스'에 작성되고 '해당 없음' 판단의 사유가 기재되었는지 확인
4	위험 등급별 관리·규제 의무의 식별	분류 결과에 따라 요구되는 데이터 거버넌스·기술문서·인간 감독·모니터링 의무가 목록으로 식별되었는지 '리스크 분류 매트릭스'에서 확인
5	분류 결과의 과업 범위·예산 반영	식별된 의무가 {사업수행계획서}의 과업·산출물 목록과 예산에 반영되었는지 WBS와 대조
6	위험 분류 결과의 외부 검토	'법무 검토 의견서' 또는 '전문가 자문서'의 작성 주체가 사업 수행 조직과 분리되어 있는지, 검토 시점이 분류 확정 이전인지 확인하고 담당자 면담으로 보완

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	리스크 분류 매트릭스를 통해 EU AI Act 부속서Ⅲ 고위험 분야 및 AI 기본법 고영향 AI 해당 여부, 금지 AI 적용 가능성이 법령 조항 근거와 함께 검토되고, 분류 결과에 따른 데이터 거버넌스·기술문서·인간 감독·모니터링 의무가 목록으로 식별되며, 외부(법무·전문가) 검토를 거쳐 {사업수행계획서}의 과업·산출물·예산에 반영된 경우	현행 유지
보완권고 (△)	위험 분류는 수행되었으나 법령 조항 인용·판단 근거가 미흡하거나, 외부 전문가·법무 검토가 수행되지 않은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 위험 분류가 수행되지 않았거나, 고위험·고영향 해당 여부에 대한 검토 없이 사업 범위가 확정된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 위험 등급의 '낮음' 편향(under-classification)을 가장 먼저 확인한다: 등급이 올라가면 의무·예산·일정이 함께 늘어나므로 사업자는 낮게 분류할 유인을 갖는다. 감리원은 분류 결과 자체보다 '낮음' 판단의 근거 조항과 배제 사유가 문서로 제시되었는지를 확인한다.

2) 고영향 AI와 고위험 AI는 별개의 기준이다: AI 기본법의 고영향 AI와 EU AI Act의 고위험 AI는 판단 기준과 적용 범위가 다르므로 하나만 검토하고 나머지를 생략해서는 안 된다. 감리원은 두 기준의 해당 여부가 리스크 분류 매트릭스에 분리 기재되었는지 확인한다.

3) 분류 결과가 과업 범위·예산에 실제 반영되었는지 추적한다: 매트릭스만 작성하고 {사업수행계획서}에 반영되지 않으면 형식적 분류에 그친다. 감리원은 고위험·고영향 판정 시 요구되는 데이터 거버넌스·인간 감독·모니터링 과업이 WBS와 산출물 목록에 존재하는지 대조한다.

4) 금지(수용 불가) AI 검토는 반드시 명시적으로 기록한다: 해당 없음이 자명해 보여도 검토 기록이 없으면 감리 근거가 남지 않는다. 감리원은 EU AI Act 제5조 각 호에 대한 대조표와 '해당 없음' 판단 사유가 기재되었는지 확인한다.

5) 외부 검토는 수행 여부가 아니라 독립성을 확인한다: 사업 수행 조직 내부 인력의 자체 검토를 외부 검토로 제시하는 사례가 있다. 감리원은 법무 검토 의견서·전문가 자문서의 작성 주체가 사업 수행 조직과 분리되어 있는지, 검토 시점이 분류 확정 이전인지를 확인한다.

AG_PM04-1-06 | 데이터 라벨링 도급관리

사업관리 | 착수단계 | 데이터 수집·처리 | 조직체계

근거 | NIST AI RMF GOVERN · 「개인정보 보호법」 · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025. 12.)

1. 목적

{사업수행계획서} 내 데이터 라벨링 도급 관리 요건(하수급인 자격, 품질 기준, 검수 절차) 및 하도급 작업자 보안 요건(신원 확인, 접근 통제, 보안서약)이 구체적이고 실행 가능한 수준으로 명시되었는지 여부를 제3자적 관점에서 점검한다.

특히, 라벨링 작업 중 발생할 수 있는 개인정보 유출·데이터 오염·품질 저하 등의 위험에 대한 사전 통제 방안이 사업 기획 단계부터 요구사항으로 구체화되었는지를 확인한다.

2. 필요성

▶ AI 모델의 성능과 신뢰성은 학습 데이터의 품질에 직결되며, 라벨링 작업은 대규모 인력이 투입되는 도급·하도급 구조로 수행되는 경우가 많다. 이 과정에서 다음과 같은 위험이 발생할 수 있어 요구정의 단계부터 관리 요건을 명확히 수립할 필요가 있다.

▶ 데이터 품질 위험: 하도급 작업자의 자격 미달, 불명확한 라벨링 기준으로 인해 학습 데이터에 오류·편향이 유입되어 AI 모델의 신뢰성이 저하될 수 있다.

▶ 개인정보 유출 위험: 라벨링 대상 데이터에 개인정보가 포함된 경우, 하도급 작업자에 의한 유출·유용이 발생할 수 있으며, 이는 개인정보 보호법 위반으로 이어질 수 있다.

▶ 책임 소재 불명확 위험: 도급·하도급 구조에서 라벨링 오류나 보안사고 발생 시 원도급사와 하수급인 간 책임 경계가 불분명하여 사후 분쟁이 발생할 수 있다.

▶ 공급망 보안 위험(NIST AI RMF GOVERN 6.2): 외부 라벨링 인력을 통한 데이터 유출, 악의적 데이터 오염(Data Poisoning) 등 공급망 단계의 보안 위협이 존재한다.

3. 검토방법

No	점검 항목	확인 방법
1	라벨링 도급 업체 선정 기준 명시 여부	자격요건·경험·보안인증 등 라벨링 도급 업체(하수급인) 선정 기준이 {사업수행계획서}에 명시되었는지 확인
2	라벨링 작업 범위·수량·일정·품질 기준 정의 여부	라벨링 작업 범위·수량·일정과 정확도 목표치 등 품질 기준이 {사업수행계획서}에 정의되었는지 확인하고 사업자 면담으로 보완
3	라벨링 결과 검수·품질 확인 절차 및 책임자 명시 여부	라벨링 작업 결과에 대한 검수·품질 확인 절차와 책임자가 {사업수행계획서}에 명시되었는지 확인

No	점검 항목	확인 방법
4	개인정보 처리 위탁 조항·보안서약서 징구 요건 명시 여부	하도급 계약서에 개인정보 처리 위탁 조항과 보안서약서 징구 요건이 포함될 것임이 {사업수행계획서} 및 '하도급(도급) 계약서(안)'에 명시되었는지 확인
5	저작권 귀속 및 결과물 반납·파기 절차 반영 여부	라벨링 데이터의 저작권·지식재산권 귀속 관계와 작업 결과물 반납·파기 절차가 {사업수행계획서} 및 '하도급(도급) 계약서(안)'에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	하수급인 선정 기준, 작업 범위·수량·일정·품질 기준(정확도 목표치), 검수 절차·책임자, 개인정보 처리 위탁 조항·보안서약서 징구 요건, 저작권 귀속·결과물 반납·파기 절차가 {사업수행계획서}에 모두 구체적으로 명시됨	현행 유지
보완권고 (△)	도급 관리 계획은 수립되어 있으나 품질 기준 수치 미정의, 검수 절차·책임자 불명확, 개인정보 위탁 조항·보안서약 요건 또는 파기 절차 등 일부 요건이 미흡함	해당 항목 보완 요청 후 재확인
부적합 (X)	라벨링 도급 관리 요건과 하도급 작업자 보안 요건이 {사업수행계획서}에 전혀 반영되지 않았거나 '적절히 관리한다' 수준의 선언적 문구만 존재함	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선언적 문구를 측정 가능한 요건으로 구체화 유도: "하도급 업체를 적절히 관리한다", "보안교육을 실시한다"와 같은 문구는 이행 여부 판단이 불가능하다. 감리원은 {사업수행계획서}에 측정 가능한 기준(예: 작업 착수 전 100% 보안서약서 징구, 라벨링 결과물 10% 이상 샘플 검수)으로 명시할 것을 권고해야 한다.

2) 라벨링 대상 데이터의 민감도에 따른 차등 요건 적용 확인: 라벨링 대상이 일반 이미지인지, 개인 식별 정보가 포함된 데이터인지에 따라 요구되는 보안 수준이 다르다. 감리원은 사업자가 데이터 민감도 등급(예: 공개/내부/기밀)을 사전 분류하고, 등급별 관리 요건을 차등 적용하는 계획을 수립하였는지 확인해야 한다.

3) 하도급 재위탁 여부 및 승인 요건 확인: 하수급인이 라벨링 작업을 제3자에게 재위탁하는 경우 발주기관·원도급사의 사전 승인 절차와 동일 수준의 보안·품질 요건 적용 여부를 확인하고, 계약서(안)에 재위탁 제한·승인 조항이 포함되었는지 내용의 충실성까지 검토해야 한다.

AG_PM04-1-07 | 데이터 라벨링 요구정의

사업관리 | 착수단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 모델 학습에 사용되는 데이터 라벨링(어노테이션) 작업의 품질을 확보하기 위한 '라벨링 가이드라인' 문서를 구체적이고 상세하게 작성하였는지, 일관된 라벨링을 위한 규칙 작성 요건이 요구사항에 적절하게 정의되었는지, 그리고 라벨링 작업자(라벨러)를 체계적으로 교육하기 위한 계획과 편향(Bias) 방지 교육 계획이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 라벨링 가이드라인이 단순한 작업 매뉴얼 수준에 머물지 않고 AI 학습 데이터의 품질·일관성·편향 방지를 위한 구체적인 규칙·기준을 포함하고 있는지, 이를 뒷받침하는 교육 계획과 검수 체계가 요구사항정의서에 반영되었는지를 확인한다.

2. 필요성

▶ AI 모델의 성능은 학습 데이터의 품질에 직접적으로 좌우되며, 학습 데이터의 품질은 라벨링 작업의 일관성과 정확성에 의해 결정된다. 특히 생성형 AI 서비스에서는 파인튜닝(Fine-tuning), RLHF(강화 학습 기반 인간 피드백) 등의 과정에서 라벨링된 데이터가 모델 출력 품질을 직접 좌우하므로, 라벨링 가이드라인의 구체성과 일관성 확보가 사업 초기에 반드시 정의되어야 한다.

▶ 데이터 품질 확보: 라벨링 기준이 모호하거나 부재할 경우, 작업자 간 판단 차이로 인해 동일 데이터에 대해 상이한 라벨이 부여되어 학습 데이터의 일관성이 저하되고, 결과적으로 AI 모델의 정확도와 신뢰성이 하락한다.

▶ 편향 방지: 라벨링 작업자의 개인적 편향(성별, 인종, 문화적 선입견 등)이 데이터에 반영될 수 있으므로, 이를 사전에 차단하기 위한 명확한 작업 표준과 교육 체계가 착수 단계에서 정의되어야 한다.

▶ 감리 기준 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야) 요구사항 05-1d 및 07-3a에서는 라벨링 작업자 교육 시행과 가이드 문서 마련, 라벨링 기준의 명확한 수립을 요구하고 있으며, 이는 감리 시 필수 점검 대상이다.

▶ 재작업 비용 절감: 라벨링 가이드라인과 교육 계획이 초기에 명확하게 수립되지 않으면, 이후 라벨링 오류 발견 시 대규모 재작업이 발생하여 사업 일정과 비용에 심각한 영향을 미칠 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	라벨링 대상 데이터 유형별 라벨링 규칙 명시 여부	텍스트, 이미지, 음성 등 유형별 라벨링 규칙이 요구사항에 명시되었는지 확인
2	라벨 분류 체계(택소노미)와 경계 조건 명시 여부	라벨 분류 체계와 각 라벨의 정의·경계 조건이 요구사항에 명시되었는지 확인
3	정답/오답 예시 포함 요건	라벨링 작업의 입력-출력 예시(정답/오답 예시)가 포함되도록 요구되었는지 확인
4	경계 사례(Edge Case) 판단 기준 명시 여부	경계 사례 및 애매한 경우의 판단 기준이 요구사항에 명시되었는지 확인
5	WBS 산출물 반영 여부	'라벨링 가이드라인' 문서 작성이 WBS에 산출물로 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	데이터 유형별 라벨링 규칙, 라벨 분류 체계와 경계 조건, 정답/오답 예시 포함 요건, 경계 사례 판단 기준이 {요구사항정의서}에 구체적으로 정의되어 있고, '라벨링 가이드라인' 작성이 WBS 산출물로 반영됨	현행 유지
보완권고 (△)	라벨링 규칙·분류 체계는 정의되어 있으나 경계 사례 판단 기준, 정답/오답 예시, WBS 산출물 반영 등 일부 요건이 누락되거나 모호하게 기술됨	해당 항목 보완 요청 후 재확인
부적합 (X)	라벨링 가이드라인 작성 요건이 {요구사항정의서}에 전혀 정의되어 있지 않거나 도구 사용법 수준의 매뉴얼만 계획됨	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 요건의 수치화 유도: '라벨링 품질을 확보한다'와 같은 모호한 문구는 향후 분쟁의 원인이 된다. '라벨링 일치도(Kappa) 0.8 이상', '검수 비율 전체 라벨링 건수의 10% 이상', '편향 방지 교육 이수율 100%' 등 측정 가능한 기준으로 구체화하도록 유도한다.

2) 라벨링 가이드라인와 작업 매뉴얼의 구분: '라벨링 가이드라인'는 단순한 도구 사용법(작업 매뉴얼)이 아니라 라벨 분류 체계, 판단 기준, 경계 사례 처리 원칙, 편향 방지 규칙 등을 포함하는 품질 관리 문서이다. 사업자가 도구 매뉴얼만 제출한 경우 보완을 권고한다.

3) 데이터 유형별 편향 검토 차별화: 라벨링 대상이 텍스트·이미지·음성 등 어떤 유형인지에 따라 발생 가능한 편향 유형이 다르므로, 유형별로 편향 검토 기준과 교육 내용을 달리 적용하도록 점검한다.

4) 발견 사항은 구체적 근거와 함께 기록: 라벨링 규칙 누락, 경계 사례 기준 부재 등 발견 사항은 해당 요구사항 ID와 문서 근거를 명시하여 감리보고서에 기록한다.

AG_PM04-1-08 | 라벨링데이터 교차검수 요건 및 다수 검수자 검증 프로세스 반영

사업관리 | 착수단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용하는 라벨링 데이터의 품질을 보장하기 위한 교차 검수 프로세스(다수 검수자 검증·불일치 해소 절차·검수자 간 일치율 기준)이 {사업수행계획서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

단일 검수자 의존 구조에서 발생하는 주관적 편향과 오류 누적을 방지하기 위한 객관적 품질 보증 체계의 사전 설계 여부를 확인한다.

2. 필요성

▶ AI 모델의 예측 품질은 학습 데이터에 부여된 라벨의 정확성에 직접 의존한다. 단일 검수자가 라벨링 결과를 검수하면 개인의 주관적 판단 기준 차이·피로도에 의한 집중력 저하·특정 패턴에 대한 인지 편향이 누적되어 학습 데이터 전체의 라벨 품질이 저하될 수 있다. 다수 검수자 교차 검수 프로세스와 검수자 간 일치율 기준을 사전에 정의해야 한다.

▶ 단일 검수자 주관 편향에 의한 라벨 오류 누적 방지: 한 명의 검수자가 모든 라벨을 검수하면 그 검수자의 주관적 해석 기준이 전체 학습 데이터에 반영된다. 검수자가 특정 유형의 경계 사례를 일관되게 잘못 판단하면 모델이 편향된 패턴을 학습하여 서비스 운영에서 체계적 오류를 반복 발생시킨다.

▶ 라벨링 작업자와 검수자 간 기준 불일치 조기 탐지: 다수 검수자가 동일 데이터를 독립적으로 검수하면 검수자 간 판단 불일치 구간이 명확히 드러나며, 이는 라벨링 가이드라인 자체의 모호성을 탐지하는 신호가 된다. 교차 검수 없이는 가이드라인 모호성이 표면화되지 않고 오류가 누적된다.

▶ 검수자 간 일치율(Inter-Annotator Agreement) 기반 객관적 품질 지표 확보: Cohen's Kappa·Fleiss' Kappa 등 통계적 일치율 지표를 사전 정의하고 기준 수치를 충족하지 못할 경우 라벨링 가이드라인 재검토 절차를 {사업수행계획서}에 요건으로 포함해야 AI 학습 데이터 품질을 객관적으로 관리할 수 있다.

▶ TTA 신뢰 AI 및 NIST AI RMF 데이터 품질 투명성 요건 충족: TTA 신뢰할 수 있는 인공지능 개발안내서는 라벨링 품질 보증 프로세스의 투명성·문서화를 신뢰 AI 요건으로 명시하며, NIST AI RMF MAP 2.3은 학습 데이터 품질 위험 통제 방안의 설계 단계 사전 정의를 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	교차 검수 대상 범위 및 검수자 수·자격 요건 정의	전체 라벨링 데이터 중 교차 검수 적용 대상 범위(전수 검수 또는 샘플링 비율), 교차 검수에 참여하는 검수자 최소 인원(예: 동일 항목당 검수자 2인 이상), 검수자 자격 요건(도메인 경험·교육 이수 등)이 {사업수행계획서}에 정의되었는지 확인
2	검수자 간 일치율 목표 기준 수치 정의	교차 검수 후 검수자 간 라벨 일치율을 측정하는 방법(Cohen's Kappa·Fleiss' Kappa·단순 일치율 등)과 목표 기준 수치(예: 'Cohen's Kappa 0.8 이상')가 정의되었는지 확인
3	불일치 라벨 해소 절차 및 최종 라벨 결정 기준 정의	다수 검수자 간 라벨 판단이 불일치할 경우 최종 라벨을 결정하는 절차(다수결·도메인 전문가 중재·재검수 등)와 불일치 비율이 허용 임계값(예: 20%)을 초과할 경우 라벨링 가이드라인 재검토를 수행하는 절차가 정의되었는지 확인
4	교차 검수 결과 기록·이력 관리 산출물 요건 정의	교차 검수 수행 결과(검수자별 판정 내역·일치율 측정값·불일치 해소 내역·최종 라벨 승인 이력)를 기록하는 산출물('라벨링 품질 검수 결과 보고서')이 {사업수행계획서}에 정의되었는지 확인
5	RFP 대비 교차 검수 요건 완전성 추적	RFP에 명시된 라벨링 교차 검수·다수 검수자 검증·일치율 기준·불일치 해소 절차 관련 요구사항이 {사업수행계획서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	교차 검수 대상 범위·비율, 검수자 최소 인원·자격 요건, 검수자 간 일치율 측정 방법·목표 기준 수치, 불일치 해소 절차·가이드라인 재검토 트리거, 교차 검수 결과 기록 산출물 요건이 {사업수행계획서}에 모두 정의됨	현행 유지
보완권고 (△)	교차 검수 수행 계획은 있으나 검수자 자격 요건 미정의, 일치율 목표 수치 미정의, 불일치 해소 절차 또는 결과 기록 산출물 요건 중 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	교차 검수 프로세스 자체가 {사업수행계획서}에 정의되지 않거나 '검수를 철저히 수행한다'는 선언적 서술만 존재	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '검수를 철저히 한다'는 사업수행계획은 가장 흔한 미정의 패턴: '라벨링 품질 관리를 철저히 수행한다', '전문가가 검수한다'는 서술은 교차 검수 프로세스가 전혀 정의되지 않은 선언에 불과하다. 감리원은 검수자 최소 인원·자격 요건, 일치율 측정 방법·목표 수치, 불일치 해소 절차, 결과 기록 산출물이 모두 {사업수행계획서}에 구체적으로 포함되도록 유도해야 한다.

2) 라벨링 작업자가 본인 작업을 검수하는 구조는 교차 검수로 인정 불가: 일부 사업자는 라벨링 작업자가 스스로 자신의 작업 결과를 재검토하는 것을 '검수 완료'로 처리하는 경우가 있다. 감리원은 라벨링 작업자와 검수자가 역할 분리된 별도 인원인지, 동일인의 자체 재검토와 제3자 교차 검수가 명확히 구분되는지 {사업수행계획서}에서 확인해야 한다.

3) Cohen's Kappa 0.6 미만이면 라벨링 가이드라인 자체를 의심해야: 교차 검수 후 Cohen's Kappa가 0.6 미만으로 측정되면 검수자 능력 부족보다 라벨링 가이드라인 자체의 모호성이 원인일 가능성이 높다. 감리원은 Kappa 기준 미달 시 가이드라인 재검토 트리거 절차가 {사업수행계획서}에 정의되었는지 확인하고, Kappa 목표 수치가 0.8 이상 수준으로 설정되었는지 점검해야 한다.

4) 교차 검수 결과 기록 없는 라벨링은 감리 시 검증 불가: 교차 검수를 수행했더라도 검수자별 판정 결과·일치율 측정값·불일치 해소 내역이 기록된 산출물이 없으면 감리원이 교차 검수 수행 여부와 품질을 사후 검증할 수 없다. '라벨링 품질 검수 결과 보고서'가 개발 과정의 이관 산출물로 {사업수행계획서}에 정의되도록 반드시 유도해야 한다.

5) 골든 데이터셋 혼입은 검수자 일관성 모니터링의 가장 효과적인 방법: 정답이 확정된 골든 데이터를 검수 대상에 무작위 혼입하여 검수자가 인식하지 못하게 하면 검수자의 실제 판단 정확도를 객관적으로 모니터링할 수 있다. 감리원은 이 골든 데이터셋 혼입 모니터링 방법과 기준 정확도 미달 시 검수자 교체 절차가 {사업수행계획서}에 포함되었는지 확인해야 한다.

AG_PM04-1-09 | 라벨링 편향방지 및 주관적 판단 배제 라벨링 기준 요건

사업관리 | 착수단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용되는 데이터의 라벨링(Labeling) 작업 시 어노테이터의 주관적 판단에 의한 편향을 배제하는 명확한 라벨링 기준과 작업자 간 일치도(IAA) 검증 요건이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 모델의 예측 품질은 학습 데이터 라벨의 정확성과 일관성에 직접 의존한다. 라벨링 기준이 모호하거나 어노테이터의 주관적 판단에 맡겨지면 동일 데이터에 대해 서로 다른 라벨이 부여되는 레이블 노이즈가 발생하고, 이는 모델 학습 과정에서 오류 패턴으로 내재화된다. 라벨링 기준·편향 방지 절차·작업자 간 일치도(IAA) 검증 요건을 명확히 정의해야 한다.

▶ 라벨 노이즈에 의한 모델 성능 저하 방지: 어노테이터마다 다른 기준으로 라벨을 부여하면 학습 데이터에 모순된 패턴이 생성되어 모델이 잘못된 결정 경계를 학습한다. 라벨링 기준이 객관적 수치 기준으로 정의되어야 라벨 노이즈를 구조적으로 차단할 수 있다.

▶ 어노테이터 편향의 모델 내재화 예방: 어노테이터가 특정 문화권·성별·연령 배경에서 비롯된 주관적 판단으로 라벨을 부여하면, 그 편향이 학습 데이터 전체에 확산되어 모델이 특정 집단에 불리한 결과를 반복 생성하는 구조적 편향이 발생한다. 다수 어노테이터 검토·다수결 합의 방식 등 편향 방지 절차가 요건으로 정의되어야 한다.

▶ TTA 신뢰 AI 및 NIST AI RMF 요건 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 학습 데이터 라벨링 작업의 일관성·정확성·편향 배제를 신뢰 AI 확보 요건으로 명시하며, NIST AI RMF MAP 2.3은 데이터 품질 위험 요소로 라벨링 오류와 주관적 판단 편향을 명시적으로 포함한다.

▶ 감리원의 라벨링 품질 독립 검증 기반 확보: 작업자 간 일치도(IAA) 기준과 라벨링 가이드라인이 {요구사항정의서}에 명시되어야 감리원이 이후 설계·구현 단계에서 라벨링 품질 결과의 적정성을 독립적으로 검증할 수 있다. IAA 기준 없이는 라벨링 품질 적정성 판단이 불가능하다.

3. 검토방법

No	점검 항목	확인 방법
1	라벨 유형별 정의 및 객관적 판단 기준 명세	라벨 유형(클래스명·레이블 값 등)별 정의와 라벨 부여 판단 기준이 어노테이터의 주관적 해석 없이 적용 가능하도록 객관적·수치적·예시 기반으로 {요구사항정의서} 및 '라벨링 가이드라인'에 명세되었는지 확인
2	경계 사례(Edge Case) 처리 기준 정의	라벨 부여가 명확하지 않은 경계 사례 유형과 처리 방법(특정 라벨 우선 부여·보류·상위 검토자 결정 등)이 '라벨링 가이드라인'에 사전 정의되었는지 확인. 경계 사례 정의 없이 어노테이터 개인 판단에 맡겨지는 구조인지 함께 점검
3	작업자 간 일치도(IAA) 목표치 및 측정 방법 정의	복수 어노테이터가 동일 데이터에 라벨을 부여할 때 결과의 일치 정도를 나타내는 IAA 목표치(Cohen's Kappa 0.7 이상, Fleiss' Kappa 0.6 이상 등)와 측정 방법이 {요구사항정의서}에 정의되었는지 확인
4	라벨 불일치 해소 절차 및 편향 방지 검수 요건 정의	어노테이터 간 라벨 불일치 발생 시 해소 절차(다수결 합의·상위 검토자 판정·재검토 기준 등)와, 특정 어노테이터의 라벨링 패턴이 지속적으로 편향될 경우 탐지·조치하는 품질 검수 요건이 정의되었는지 확인
5	RFP 대비 라벨링 기준·편향 방지 요건 완전성 추적	RFP에 명시된 라벨링 기준·편향 방지·IAA 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	라벨 유형별 객관적 기준(수치·예시·체크리스트)이 정의되고, 경계 사례 처리 기준·IAA 목표치 수치·측정 방법·불일치 해소 절차·어노테이터 편향 탐지 요건이 {요구사항정의서}에 명시됨	현행 유지
보완권고 (△)	라벨 유형 정의와 가이드라인은 있으나, 주관적 표현 포함 또는 IAA 목표치 수치 미정의, 불일치 해소 절차·어노테이터 편향 탐지 요건 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	라벨링 편향 방지 요건 및 주관적 판단 배제 라벨링 기준이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 라벨링 가이드라인에서 주관적 표현을 직접 찾아 지적해야 효과적: 감리원이 '라벨링 가이드라인'을 직접 검토할 때 '긍정적인', '적절한', '어느 정도', '전반적으로' 등의 주관적 표현을 직접 발췌하여 감리 의견에 인용하는 것이 가장 설득력 있다. '주관적 표현이 다수 포함되어 어노테이터마다 다른 기준으로 라벨링할 위험이 있다'는 구체적 사례 기반 지적이 가장 효과적이다.

2) IAA 목표치는 도메인별 적정 수준 사전 파악 후 점검: Cohen's Kappa 0.7은 일반적인 텍스트 분류에서의 목표치이지만, 의료 이미지 판독·법률 문서 분류 등 고위험 도메인에서는 0.8 이상이 요구될 수 있고, 감성 분석처럼 주관성이 본질적으로 높은 작업에서는 0.6 수준도 현실적이다. 감리원은 도메인별 적정 IAA 수준을 사전 파악하여 요건의 적정성을 점검해야 한다.

3) 경계 사례 처리 기준 누락이 실무에서 가장 빈번한 문제: 라벨 유형 정의는 있지만 경계 사례 처리 기준이 없는 가이드라인이 실무에서 가장 자주 발견된다. 어노테이터들은 경계 사례 발생 시 각자의 판단으로 처리하게 되어 대량의 레이블 노이즈가 발생한다. 감리원은 '경계 사례 유형을 몇 가지 사전 정의했는지, 처리 방법이 명시되어 있는지'를 별도로 상세 검토해야 한다.

4) 어노테이터 편향 탐지 요건은 사후 품질 관리의 핵심: 특정 어노테이터가 지속적으로 특정 클래스를 과다 부여하는 패턴은 개별 라벨 불일치보다 더 큰 편향을 발생시킨다. 어노테이터별 라벨링 패턴 모니터링 요건(예: 어노테이터별 클래스 분포 분석, 특이 패턴 발견 시 재작업 조치)이 {요구사항정의서}에 포함되도록 유도해야 한다.

5) 생성형 AI 선호도 라벨링의 루브릭 정의 누락 주의: RLHF 기반 생성형 AI 학습의 경우 '어떤 응답이 더 좋은가'를 판단하는 선호도 라벨링이 필요한데, '더 좋다'는 기준 자체가 매우 주관적이다. 유해성·정확성·유창성·유용성 등 평가 항목별 루브릭(세부 채점 기준)이 구체적으로 정의되었는지 확인해야 하며, 루브릭 없이 '전반적 선호도'로만 평가하는 경우 보완을 권고해야 한다.

AG_PM04-1-10 | 데이터 수집 편향배제 및 특정 출처 편중방지 수집 요건

사업관리 | 착수단계 | 데이터 수집·처리 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF GOVERN · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용되는 데이터를 수집할 때 특정 출처·집단·시점에 편중되지 않도록 다양성을 확보하는 편향 배제 방안과 출처 다양성 수집 요건이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 모델은 학습 데이터의 편향을 그대로 학습하고 증폭시킨다. 데이터 수집 단계에서 특정 출처·집단·시점·지역에 편중된 데이터가 수집되면 모델이 특정 그룹에 불리한 결과를 반복적으로 생성하는 구조적 편향이 발생한다. 이를 방지하려면 출처 다양성 기준, 인구통계적 균형 기준, 수집 출처별 비율 상한 등 편향 배제 기준을 수치로 정의해야 한다.

▶ 모델 편향에 의한 차별적 결과 방지: 특정 성별·연령·지역·언어에 편중된 학습 데이터는 해당 그룹에 유리하거나 불리한 예측 결과를 반복 생성한다. 수집 데이터의 인구통계적·출처별 균형 기준을 비기능 요건으로 정의해야 모델 공정성을 구조적으로 확보할 수 있다.

▶ 특정 출처 과의존에 따른 단일 관점 고착화 위험 차단: 학습 데이터가 단일 플랫폼(예: 특정 포털 뉴스·소셜미디어)이나 단일 기관 자료에 80% 이상 의존하면, 해당 출처의 시각·편견·오류가 모델 전체에 고착화된다. 출처별 수집 비율 상한(예: 단일 출처 30% 이내)을 요건으로 정의해야 특정 출처 과의존 위험을 원천 차단할 수 있다.

▶ TTA 신뢰 AI 요건 및 NIST AI RMF 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 학습 데이터의 다양성·대표성·편향 배제를 신뢰 AI 확보 필수 요건으로 명시하며, NIST AI RMF GOVERN 6.2는 편향 위험을 사전 식별하고 수집 단계부터 통제 방안을 마련할 것을 요구한다.

▶ 감리원의 데이터 수집 적정성 판단 기준 확보: 편향 배제 방안이 {요구사항정의서}에 수치로 정의되지 않으면 감리원이 이후 설계·구현 단계에서 데이터 수집 계획의 적정성을 판단할 기준이 없다. 편향 배제 요건 명시가 전체 감리 체계의 출발점이 된다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 수집 출처 다양성 기준 및 출처별 비율 상한 정의	학습 데이터를 수집할 출처 목록(웹 크롤링·공공 데이터·사내 데이터·오픈소스 데이터셋 등)과 단일 출처 의존 방지를 위한 출처별 수집 비율 상한(예: 단일 출처 30% 이내)이 {요구사항정의서}에 정의되었는지 확인
2	인구통계적·사회적 다양성 확보 기준 정의	성별·연령·지역·언어·문화권 등 인구통계적 속성의 균형 기준이 학습 데이터 수집 요건으로 정의되었는지 확인. 특정 집단 과대표현·과소표현 방지를 위한 수치 기준(예: 특정 성별 데이터 비율 40~60% 범위)이 함께 정의되었는지 확인
3	수집 시점 편향 배제 및 시간적 다양성 확보 요건 정의	특정 시기(예: 특정 연도·계절·이벤트 기간)에 편중된 데이터 수집을 방지하기 위한 수집 시점 분산 기준과, 시간적 다양성 확보 요건(예: 최근 3년 데이터 균등 수집)이 {요구사항정의서}에 정의되었는지 확인
4	편향 탐지 및 사후 보정 방안 요건 정의	수집 완료 후 편향 탐지(데이터 분포 분석, 인구통계 분포 확인 등) 및 탐지된 편향 보정 방법(오버샘플링·언더샘플링·데이터 증강 등)이 비기능 요건으로 정의되었는지 확인. 편향 탐지·보정 수행 이력 관리 요건 포함 여부도 함께 확인
5	RFP 대비 편향 배제·출처 다양성 요건 완전성 추적	RFP에 명시된 데이터 편향 배제·출처 다양성 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	출처 다양성 기준·출처별 비율 상한, 인구통계적 균형 기준, 수집 시점 분산 기준, 편향 탐지 방법·탐지 기준, 보정 방법·적용 조건, 탐지·보정 이력 관리 요건이 수치로 {요구사항정의서}에 정의됨	현행 유지
보완권고 (△)	출처 목록과 편향 배제 방침은 있으나, 출처별 비율 상한·인구통계 균형 수치 미정의 또는 편향 탐지·보정 결과 산출물 관리 요건 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	편향 배제 방안 및 출처 다양성 수집 요건이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '다양한 데이터 수집'은 선언이고, '출처별 비율 상한'이 요건이다: '다양한 출처에서 균형 있게 수집한다'는 서술은 비기능 요건이 아니다. 감리원은 '단일 출처 비율 30% 이내', '3개 이상 언어권 균등 포함', '특정 성별 40~60% 범위' 등 각 편향 유형별로 측정 가능한 수치 기준이 {요구사항 정의서}에 별도 항목으로 정의되도록 유도해야 한다.

2) 고위험 도메인은 강화된 편향 배제 요건 적용 여부 별도 확인: 채용·금융 심사·의료 진단 등 고위험 도메인 AI에서 데이터 편향은 법적 차별로 이어질 수 있다. 감리원은 AI 서비스 도메인이 고위험에 해당하는지 먼저 파악하고, 고위험 도메인인 경우 일반 서비스 대비 강화된 편향 배제 기준(외부 편향 감사, 제3자 검증 계획 포함 등)이 요건에 반영되었는지 확인해야 한다.

3) 크롤링 데이터의 구조적 편향 위험 인식: 검색 엔진 상위 노출 기반 크롤링은 플랫폼 추천 알고리즘의 편향을 그대로 흡수한다. '웹 크롤링으로 수집한다'는 요건만 있고 크롤링 대상 도메인·URL 다양성 기준이 없다면 출처 편향 위험이 높다. 크롤링 데이터는 반드시 수집 후 출처 분포 분석 및 편향 탐지 절차를 요건에 포함하도록 유도해야 한다.

4) 사내 데이터의 역사적 편향 탐지 요건 누락 주의: 기업 내부 데이터는 '이미 보유한 데이터'라는 이유로 편향 탐지 없이 그대로 학습 데이터로 활용하는 경우가 많다. 사내 데이터 활용 시 과거 의사결정 패턴에서 비롯된 역사적 편향 탐지 절차를 별도 요건으로 포함하도록 유도해야 한다.

5) 편향 보정 방법을 수치 조건과 함께 정의 유도: 오버샘플링·언더샘플링·데이터 증강 등 보정 방법만 열거하고 '어떤 조건에서 적용하는지'가 없으면 실제 수행 여부를 감리로 검증할 수 없다. '소수 집단 비율이 전체의 10% 미만인 경우 SMOTE 오버샘플링 적용' 등 적용 조건과 보정 방법을 연계하여 정의하도록 유도해야 한다.

AG_PM04-2-01 | 라벨링 교차검수 상태값 설계점검

사업관리 | 실행통제단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습 데이터의 라벨링 작업에 대한 교차 검수 상태 값(작업중·완료·검수완료·반려 등)이 {테이블정의서}에 구조화된 코드 체계로 적정하게 설계되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 학습 데이터의 품질은 라벨링 정확도에 의해 결정되며 라벨링 교차 검수는 품질을 보증하는 핵심 프로세스이므로, 교차 검수 상태 값이 DB 수준에서 체계적으로 관리되지 않으면 검수 미완료 데이터가 학습셋에 혼입되거나 반려 사유 추적이 불가능하여 라벨 품질 관리 체계가 무력화된다.

따라서 상태 코드의 정의·전이 규칙·이력 관리·학습 데이터 연계 필터링 구조 등을 종합적으로 검토한다.

2. 필요성

▶ AI 학습 데이터의 라벨링 교차 검수 상태를 DB 수준에서 구조적으로 관리하는 것은 학습 데이터 품질 보증의 기반이며, 상태 관리 없이 라벨링을 수행하면 어떤 데이터가 검수 완료되었는지, 반려된 데이터가 재작업되었는지를 체계적으로 파악할 수 없다.

▶ 학습 데이터 품질 게이트(Quality Gate) 구현: 교차 검수가 완료된 데이터만 AI 학습에 사용해야 하며, 상태 값이 DB에 없으면 'WHERE status = 검수완료' 조건으로 학습 데이터를 필터링할 수 없어 미검수·반려 데이터가 학습셋에 혼입될 위험이 있다.

▶ 라벨링 진행 현황 관리: 대규모 라벨링 프로젝트에서 수만~수십만 건의 데이터별 진행 상태(작업중, 완료, 검수중, 검수완료, 반려)를 DB에서 집계하여 진행률을 모니터링해야 하며, 상태 코드 없이는 진행 현황 대시보드 구성이 불가능하다.

▶ 반려 사유 추적 및 품질 개선: 교차 검수에서 반려된 데이터의 반려 사유(라벨 오류, 경계 모호, 가이드라인 위반 등)가 DB에 기록되어야 반복 오류 패턴을 분석하고 라벨링 가이드라인을 개선할 수 있다.

▶ 감사 추적 및 책임 소재 확인: 누가 라벨링했고 누가 검수했으며 언제 반려/승인되었는지가 이력으로 관리되어야 라벨 품질에 대한 책임 소재를 추적할 수 있으며, TTA 개발안내서에서도 데이터 품질 관리 체계를 요구하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	상태 코드 체계 설계	라벨링 교차 검수 상태(작업중·완료·검수중·검수완료·반려 등)가 코드값으로 구조화 설계되었는지 확인
2	상태 전이 규칙 정의	상태 간 허용되는 전이(예: 완료→검수중, 검수중→검수완료/반려)가 정의되고, 비허용 전이가 제약되었는지 확인
3	교차 검수 이력 관리 테이블 설계	검수자·검수일시·판정결과·반려사유 등 교차 검수 이력이 시계열로 관리되는 테이블이 설계되었는지 확인
4	반려 사유 코드화 설계	반려 사유가 자유 텍스트가 아닌 분류 코드(라벨오류, 경계모호, 가이드위반 등)로 구조화 설계되었는지 확인
5	학습 데이터 필터링 구조 설계	검수완료 상태 데이터만 학습 데이터셋에 포함되도록 필터링 가능한 테이블 구조가 설계되었는지 확인
6	다중 검수자 판정·일치도 산출 구조 설계	동일 데이터에 대한 복수 검수자의 판정 결과를 저장하고, 검수자 간 일치도(IAA)를 산출 가능한 구조가 설계되었는지 확인
7	모델 기반 교차검수 결과 구분 기록 (AI 특화)	LLM-as-Judge 등 모델 기반 교차검수 결과가 사람 검수 결과와 별도 상태값으로 구분되어 기록되도록 설계되었는지 확인. 심사자 유형(model/human)과 모델 정보가 구분 저장되도록 설계된 경우 적합
8	검수자 간 의견 불일치 처리 상태 설계 (AI 특화)	다중 검수자 간 의견 불일치(Disagreement) 발생 시 분쟁 상태와 해결 절차(예: 시니어 리뷰, 다수결)가 상태값 흐름에 포함되었는지 확인. 분쟁·해결·종결 상태가 정의된 경우 적합
9	라벨 품질 지표 저장 구조 설계 (AI 특화)	라벨 품질 지표(Cohen's κ, F1 등)가 상태값 변화와 함께 저장되어 검수 품질 추이 분석이 가능한지 확인. 검수 품질 지표 저장 컬럼/뷰가 정의된 경우 적합

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	상태 코드 체계·전이 규칙·검수 이력 테이블·반려 사유 코드·학습 필터링·다중 검수자 구조가 모두 구체적으로 설계되고, 모델 기반 교차검수 결과가 심사자 유형(model/human)과 함께 사람 검수 결과와 구분 기록되며, 검수자 간 의견 불일치의 분쟁·해결·종결 상태와 라벨 품질 지표(Cohen's κ, F1 등) 저장 구조가 설계된 경우(예: 7단계 상태 코드 정의, 전이 다이어그램 포함, 이력 테이블 설계, IAA 산출 가능)	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	상태 컬럼은 존재하나 상태 코드가 단순(Y/N 등)하거나, 전이 규칙 미정의, 이력 테이블 미설계, 반려 사유 미코드화 등 일부 미흡한 경우(예: 상태 2개만 존재, 검수 이력 없음, 반려 사유 자유 텍스트)	해당 항목 보완 요청 후 재확인
부적합 (X)	라벨링 교차 검수 상태 컬럼이 {테이블정의서}에 전혀 설계되지 않은 경우(상태 컬럼 부재, 검수 관리 구조 전무, 품질 게이트 불가)	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '완료/미완료' 2단계 상태는 AI 라벨링에서 불충분: 일반 업무 시스템의 '처리완료 Y/N'과 달리 AI 라벨링은 작업중→라벨링완료→검수중→검수완료/반려→재작업 등 다단계 워크플로우를 거치므로, 최소 5~7단계 상태 코드가 정의되었는지 확인하고 2단계만 있는 경우 세분화를 권고한다.

2) 검수 이력 없이 상태만 관리하면 품질 추적 불가: 라벨링 데이터 테이블에 현재 상태만 저장하고 이력 테이블이 없으면 '이 데이터가 몇 번 반려되었고 왜 반려되었는지'를 추적할 수 없으므로, 검수 이력 테이블의 존재를 반드시 확인한다.

3) 반려 사유의 코드화가 품질 개선의 핵심: 반려 사유가 자유 텍스트로만 기록되면 '라벨 오류가 전체 반려의 몇 %인지'를 분석할 수 없으므로, 코드+텍스트 병행 구조(코드로 분류, 텍스트로 상세 기술)가 설계되었는지 확인한다.

4) 학습 데이터 품질 게이트 확인: 가장 중요한 점검은 '검수완료(APPROVED) 상태만 학습에 사용되는 구조인가'이며, 학습 데이터셋 구성 설계에서 상태 필터링이 필수 조건으로 포함되었는지, 반려 데이터의 학습 혼입 방지 설계가 있는지를 확인한다.

5) 모델 기반 검수 결과만으로 최종 라벨 확정은 부적합 가능성: LLM-as-Judge 등 모델 기반 검수의 결과만으로 최종 라벨이 확정되도록 설계되었다면 부적합 가능성이 높으므로, 최소 일부 샘플에 대한 사람 검수 또는 사후 감사 절차가 포함되었는지 확인한다.

AG_PM04-2-02 | 라벨링 교차검수 상태코드 매핑점검

사업관리 | 실행통제단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 라벨링 작업의 교차 검수(Cross-Review) 상태 코드가 {데이터베이스}에 체계적으로 관리되고, '검수완료' 상태의 데이터만 AI 모델 훈련(학습) 데이터로 편입되는 필터링 메커니즘이 정상 동작하는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

라벨링 품질은 AI 모델 성능의 근본적 기반이다.

교차 검수가 완료되지 않은 데이터(미검수, 검수 반려 등)가 학습에 포함되면 잘못된 라벨로 인한 모델 정확도 저하, 편향 발생 등 심각한 품질 문제를 초래한다.

이에 따라 상태 코드 관리 체계와 '검수완료' 필터링의 설계·구현·동작 적정성을 검증한다.

2. 필요성

▶ AI 지도학습(Supervised Learning) 모델의 성능은 라벨링 품질에 전적으로 의존한다. 라벨링 작업은 다수의 작업자가 수행하므로 품질 편차가 불가피하며, 교차 검수를 통해 라벨의 정확성을 확인하는 과정이 필수적이다. 교차 검수가 완료되지 않은 데이터가 학습에 사용되면, '쓰레기가 들어가면 쓰레기가 나온다(GIGO)'는 원칙 그대로 모델 품질이 저하된다.

▶ 라벨 품질 보증: 교차 검수를 통과하지 않은 라벨은 오류 가능성이 높다. 검수 미완료 데이터가 학습에 포함되면 모델이 잘못된 패턴을 학습하여 예측 정확도가 저하된다.

▶ 상태 코드 기반 데이터 흐름 통제: 라벨링 → 1차 검수 → 교차 검수 → 검수완료의 단계별 상태 코드가 DB에서 체계적으로 관리되어야, 학습 파이프라인에서 '검수완료' 데이터만 정확히 필터링할 수 있다.

▶ 감사 추적성 확보: 특정 학습 데이터가 어떤 라벨러에 의해 라벨링되었고, 어떤 검수자에 의해 교차 검수되었는지 추적할 수 있어야, 모델 품질 문제 발생 시 원인을 라벨 수준까지 역추적할 수 있다.

▶ TTA·NIST 데이터 품질 요건 충족: TTA 개발안내서는 학습 데이터의 라벨링 품질 관리를 요구하며, NIST AI RMF는 데이터 수집·선택의 과학적 무결성을 강조하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	라벨링 상태 코드 체계 DB 구현 여부	대기→완료→검수→검수완료/반려 라벨링 상태 코드 체계가 '라벨링 관리 설계서'대로 {데이터베이스} 상태 코드 테이블에 구현되었는지 대조
2	검수완료 상태 학습 데이터 필터링 여부	학습 데이터 추출 시 '검수완료(APPROVED)' 상태만 필터링되는지 {테스트결과서} 필터링 동작 시험 결과에서 확인
3	교차 검수 라벨 일치율 기록 여부	교차 검수 수행 결과(라벨 일치율)가 '라벨링 품질 보고서'에 기록되었는지 확인
4	라벨링 이력 DB 기록·관리 여부	라벨러·검수자·라벨 값·상태 변경 일시 등 라벨링 이력이 '데이터베이스 설계서' 이력 테이블에 기록·관리되는지 확인
5	반려 데이터 재라벨링·재검수 프로세스 구현 여부	'반려(REJECTED)' 상태 데이터의 재라벨링·재검수 프로세스가 구현되었는지 {테스트결과서} 반려 재처리 시험 또는 '라벨링 관리 설계서'에서 확인
6	상태 코드 변경 권한 역할별 분리 여부	상태 코드 변경 권한이 라벨러·검수자·관리자 역할별로 분리되었는지 '라벨링 관리 설계서' 권한 정의에서 확인
7	라벨링 검수 요구사항 추적 가능 여부	{요구사항추적표}의 라벨링 검수 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	라벨링 상태 코드 체계가 설계서대로 DB에 구현되고, '검수완료' 데이터만 학습에 편입되는 필터링과 라벨링 이력 관리·역할별 권한 분리가 시험으로 확인되며, 교차 검수 결과(라벨 일치율)가 기록되고 반려 데이터의 재라벨링·재검수 프로세스가 구현된 경우	현행 유지
보완권고 (△)	상태 코드는 구현되었으나 필터링 동작 시험 미수행, 라벨 일치율(IAA) 미산출, 반려 데이터 재처리 프로세스 미비 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	상태 코드 관리 체계가 없거나 검수 미완료 데이터가 학습에 포함되는 등 라벨 품질 통제가 실질적으로 미구현된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '검수완료 필터링 시험'의 수행 여부를 최우선으로 확인하라: 상태 코드가 DB에 구현되어 있더라도, 학습 데이터 추출 시 '검수완료' 필터링이 실제로 동작하는지 시험으로 확인하는 것이 핵심이다. {테스트결과서}에 해당 시험이 포함되어 있는지를 최우선으로 점검한다.

2) '자기 검수 방지(Self-Review Prevention)'를 반드시 확인하라: 라벨러가 자신이 라벨링한 데이터를 직접 '검수완료'로 변경할 수 있으면, 교차 검수의 의미가 사라진다. 라벨러 ID와 검수자 ID가 동일한 건이 차단되는 로직이 구현되었는지를 사업자 면담으로 확인한다.

3) '반려(REJECTED)' 데이터의 재처리 경로를 확인하라: 반려된 데이터가 재라벨링·재검수를 거쳐 '검수완료'로 전이되는 프로세스가 구현되지 않으면, 반려 데이터가 영구 방치되어 학습 데이터 규모가 부족해질 수 있다. 반려 후 재처리 흐름이 시험으로 확인되었는지 점검한다.

4) 라벨링 품질 보고서의 '일치율(IAA)' 수치를 확인하라: 교차 검수의 실질적 품질은 라벨 일치율(IAA)로 측정된다. '라벨링 품질 보고서'에 IAA(Cohen's Kappa 또는 단순 일치율)가 산출·기록되어 있는지, 목표 기준(예: 일치율 90% 이상)을 충족하는지 확인한다. IAA가 미산출된 경우 보완 권고한다.

5) 교차 검수 없이 단순 검수만 수행하는 사업의 대응: 모든 AI 사업에서 '교차 검수'가 필수적인 것은 아니다. 소규모 사업이나 라벨링 난이도가 낮은 경우, 단일 검수(1인 검수)만 수행할 수 있다. 이 경우 '검수완료' 필터링 동작 자체가 핵심 상세 검토 항목이며, 교차 검수 미수행 자체를 부적합으로 판정하지 않되, '라벨링 관리 설계서'에 검수 방식과 품질 기준이 명시되어 있는지 확인한다.

AG_PM04-2-03 | 라벨러 분포 인적 다양성 확보결과 점검

사업관리 | 실행통제단계 | 데이터 수집·처리 | 라벨링

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용된 라벨링 데이터를 생성한 라벨러(어노테이터) 풀의 인적 구성이 사업 계획 시 정의된 다양성 기준(연령, 성별, 전문성, 지역 등)을 충족하였는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

AI 학습 데이터의 라벨링 품질은 라벨러의 인적 구성에 직접적인 영향을 받으며, 특정 연령대나 성별, 특정 전문 분야의 라벨러에 편중된 라벨링은 데이터에 무의식적 편향(Unconscious Bias)을 내포시켜 AI 모델이 특정 집단에 불공정한 결과를 산출할 위험이 있다.

이에 따라 {결과보고서}를 중심으로 라벨러 풀의 인적 다양성 확보 결과를 종합적으로 점검한다.

2. 필요성

▶ 라벨링 편향(Annotation Bias) 방지: 라벨러의 인적 구성이 동질적이면, 특정 관점이나 판단 기준이 라벨링에 과도하게 반영되어 데이터에 체계적 편향이 발생한다. 예를 들어, 감성 분석 데이터를 특정 연령대 라벨러만 작업하면, 해당 연령대의 언어 사용 패턴과 감성 판단 기준이 과대 대표되어 모델의 공정성이 저하된다.

▶ AI 공정성(Fairness) 확보: TTA 신뢰할 수 있는 인공지능 개발안내서에서는 AI 시스템의 공정성 확보를 핵심 요구사항으로 제시하고 있으며, 라벨러 다양성은 학습 데이터 수준에서 공정성을 확보하기 위한 전제 조건이다. 라벨러 풀의 편중은 데이터 편향을 거쳐 모델 편향으로 이어지는 연쇄 구조를 형성한다.

▶ 도메인 전문성 품질 보장: 의료, 법률, 금융 등 전문 도메인의 AI 서비스에서는 해당 분야 전문 지식을 갖춘 라벨러의 참여가 필수적이다. 전문성이 부족한 라벨러가 작업한 데이터는 라벨링 정확도가 저하되어 모델의 신뢰성에 직접적인 영향을 미친다.

▶ 사업 계약 이행 및 감사 추적성: 공공 AI 사업에서는 라벨러 풀의 구성 기준이 RFP 또는 '데이터 구축 계획서'에 명시되는 경우가 많으며, 이는 사업 이행의 정량적 평가 기준이 된다. NIST AI RMF에서도 데이터 수집·라벨링 과정의 인적 구성 문서화를 권고하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	라벨러 다양성 목표 정의 확인	'데이터 구축 계획서'에 연령·성별·전문성·지역 등 라벨러 풀의 다양성 목표(인원·비율)가 구체적으로 정의되어 있는지 확인
2	실제 라벨러 구성 달성 여부	'라벨링 작업 결과 보고서' 및 {결과보고서}에 보고된 실제 라벨러 분포가 목표 기준을 충족하는지 확인
3	연령대별 분포 적정성	라벨러의 연령대별 분포가 계획 대비 균형적으로 구성되었는지, 특정 연령대 편중이 없는지 확인
4	성별 분포 적정성	라벨러의 성별 분포가 계획 기준에 부합하며, 극단적 편중(예: 한쪽 성별 90% 이상)이 없는지 확인
5	도메인 전문성 확보 여부	전문 도메인(의료·법률·금융 등) 라벨링에 참여한 라벨러가 해당 분야 자격·경력 요건을 충족하는지 확인
6	라벨러 간 일치도(IAA) 검증	라벨러 간 일치도(Inter-Annotator Agreement) 시험이 수행되었고, 합격 기준을 충족하는지 확인
7	라벨러 교육·가이드라인 적용	라벨링 가이드라인이 작성되고, 라벨러 교육이 실시되었으며, 교육 이수 결과가 문서화되어 있는지 확인
8	미달 시 사유 및 대응 방안	다양성 목표에 미달한 항목이 있는 경우, 미달 사유와 보완 대책이 문서화되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	라벨러 풀의 연령·성별·전문성 분포가 계획 기준을 충족하고, IAA 시험 결과가 합격 기준에 도달하며, 라벨러 교육 및 가이드라인 적용이 문서화되고, 다양성 목표 미달 항목이 있는 경우 그 사유와 보완 대책이 문서화된 경우 (예: 연령대 3개 이상 분포, 성별 비율 계획 충족, 전문가 참여 목표 달성, Kappa 0.7 이상, 교육 이수 기록 완비)	현행 유지
보완권고 (△)	라벨러 다양성이 대부분 달성되었으나, 일부 인적 속성에서 미달이 있거나, IAA 시험·교육 기록이 일부 미흡한 경우 (예: 연령대 2개에 편중, IAA 일부 항목 기준 미달, 교육 기록 일부 누락, 작업량 편중 경미)	해당 항목 보완 요청 후 재확인
부적합 (X)	라벨러 다양성 목표가 정의되지 않았거나, 실제 라벨러 구성이 극단적으로 편중되었거나, IAA 시험이 수행되지 않은 경우 (예: 다양성 목표 미정의, 단일 연령대·성별 90% 이상 편중, IAA 미측정, 라벨링 가이드라인 부재)	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 라벨러 '인원 수'보다 '참여 비율과 작업량 분포'에 주목하라: 30명의 라벨러가 참여했더라도 상위 5명이 전체 작업량의 70%를 담당했다면, 실질적으로 5명의 판단이 데이터를 지배하는 것이다. 인원 수 달성뿐 아니라 작업량 분포의 균형성을 함께 점검하면 실효성 있는 다양성 확보 여부를 판단할 수 있다.

2) 개인정보 보호 원칙을 철저히 준수하라: 라벨러의 연령·성별·전문성 분포를 확인할 때, 개별 라벨러의 이름이나 신상 정보를 요구해서는 안 된다. 익명화된 통계 분포표(예: 20대 8명/30대 12명/40대 10명)만으로 충분하며, 이는 개인정보 보호법의 최소 수집·이용 원칙에 부합한다.

3) IAA 합격 기준의 '사전 정의 여부'를 반드시 확인하라: 라벨러 간 일치도 측정은 수행했으나 합격 기준이 사전에 정의되지 않은 경우, 사후에 결과에 맞추어 기준을 설정하는 '역방향 정의' 우려가 있다. '라벨링 가이드라인' 또는 '데이터 구축 계획서'에 IAA 합격 기준이 사전에 명시되어 있는지를 점검한다.

4) 도메인 전문성은 자격 증빙으로 확인하라: 의료·법률 등 전문 도메인에서 '전문가 라벨러 5인 참여'를 보고한 경우, 해당 라벨러의 관련 자격증, 학위, 경력 사항이 증빙되어 있는지 확인한다. 단순히 '전문가'라고 기술된 것만으로는 전문성 확보의 객관적 근거가 부족하다.

5) 라벨러 다양성과 라벨링 품질의 상관관계를 종합적으로 판단하라: 라벨러 다양성 확보 자체가 목적이 아니라, 다양성을 통해 '편향 없는 고품질 라벨링'을 달성하는 것이 궁극적 목적이다. 다양성은 달성하였으나 IAA가 현저히 낮다면, 다양한 관점이 통합되지 못한 것이므로 라벨링 가이드라인의 명확성 보완을 권고한다.

AG_PM04-2-04 | 외부 데이터 출처 명시·라이선스 기록 및 저작권 위반 방지

사업관리 | 실행통제단계 | 데이터 수집·처리 | 저작권/라이선스

근거 | 「저작권법」 · EU AI Act · NIST AI RMF GOVERN

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 활용되는 외부 데이터(오픈소스 데이터셋·크롤링 데이터·구매 데이터·제3자 제공 데이터)에 대해 다음 두 가지 요건이 핵심 산출물에 적절히 반영되었는지를 감리원의 제3자적 관점에서 통합 점검하는 것을 목적으로 한다.

첫째, 출처 명시·라이선스 기록 요건으로서 {요구사항정의서}에 오픈소스 데이터셋의 출처, 버전, 라이선스 종류, 귀속(Attribution) 의무 이행 방안이 데이터셋별로 구체적으로 정의되었는지 확인하고, 둘째, 저작권 검토·라이선스 위반방지 요건으로서 {사업수행계획서}에 모든 외부 데이터 소스의 저작권 귀속 확인, 라이선스 유형별 AI 학습 허용 범위 검토, 라이선스 위반 시 대응 절차, 데이터 출처·라이선스 이력 관리 계획이 수립되었는지 확인한다.

두 요건을 요구정의 단계에서 사전 점검함으로써 저작권 침해 소송, 데이터 사용 중단 명령, AI 서비스 강제 중단 위험을 원천 차단한다.

2. 필요성

▶ AI 모델 학습에 활용하는 외부 데이터는 출처와 라이선스 유형에 따라 상업적 AI 학습 허용 여부가 근본적으로 다르다. 요구정의 단계에서 출처·라이선스 요건을 명확히 정의하지 않으면 법적·운영적 위험이 발생한다.

▶ 공개 데이터셋의 비상업적(NC) 조건 미검토: AI Hub·Kaggle·공공데이터포털 등의 공개 데이터셋은 라이선스 조건이 데이터셋마다 다르며 일부는 상업적 이용 금지(CC-BY-NC) 또는 동일 라이선스 재배포(SA) 조건을 부과한다. 소스별 라이선스 검토 계획이 없으면 상업적 서비스에 비상업적 데이터셋을 사용하는 위반이 발생한다.

▶ 웹 크롤링 데이터의 저작권·이용약관 위반: 공개된 텍스트·이미지·영상을 AI 학습 목적으로 복제·가공하는 것은 저작권 침해에 해당할 수 있으며, 이용약관이 크롤링·상업적 이용을 금지하는 경우 손해배상·서비스 차단 위험에 노출된다.

▶ 데이터 구매 계약상 AI 학습 허용 조항 미확인: 전문 데이터 제공 업체로부터 구매한 데이터라도 계약서에 AI 학습 목적 사용 허용 조항이 없으면 계약 위반 분쟁이 발생할 수 있다.

▶ 라이선스 변경·만료 미대응: 외부 데이터 소스의 라이선스 조건은 운영 기간 중 변경될 수 있어, 모니터링 계획이 없으면 적법하게 수집한 데이터가 위반 상태로 전환되어도 인지하지 못하는 상황이 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	오픈소스 데이터셋 목록 및 출처 명시 요건	학습에 활용되는 오픈소스 데이터셋의 명칭, 출처 URL·저장소, 수집 버전, 수집 일시가 {요구사항정의서}에 데이터셋별로 명시되도록 요건이 정의되었는지 확인한다. '공개 데이터셋을 활용한다'는 단순 서술은 부적합으로 판단한다.
2	데이터셋별 라이선스 종류, 허용 활용 범위 및 귀속(Attribution) 의무 기록 요건	각 데이터셋에 적용된 라이선스(CC BY, CC BY-NC, MIT, Apache 2.0 등)와 상업적 활용 가능 여부, 파생물 배포 허용 여부, SA 조건 여부, 귀속 의무 이행 방안(서비스 화면 출처 표기, 라이선스 사본 첨부 등)이 요건으로 정의되었는지 확인
3	외부 데이터 소스별 라이선스 유형 및 AI 학습 허용 범위 검토 계획 (전체 소스)	오픈소스 데이터셋 외 크롤링·구매·제3자 제공 데이터 포함, 모든 외부 소스의 라이선스 유형(CC 계열·오픈 라이선스·전용 계약 등)과 AI 학습 목적 상업적 이용 허용 여부를 소스별로 검토하는 계획이 {사업수행계획서}에 정의되었는지 확인
4	크롤링 데이터의 저작권·이용약관 준수 및 robots.txt 확인 절차 계획	크롤링 대상 사이트의 이용약관 내 크롤링 금지 조항·상업적 이용 제한 조항, robots.txt의 크롤링 허용 범위 확인 절차가 {사업수행계획서}에 정의되었는지 확인한다. '크롤링 허용'과 'AI 학습 목적 이용 허용'을 구분하여 검토하는 절차 포함 여부도 확인
5	데이터 구매 계약 시 AI 학습 허용 조항 포함 여부 검토 계획	데이터 구매 계약 체결 전 AI 학습 목적 이용 허용·2차 가공 허용·상업적 AI 서비스 배포 허용 조항을 계약서에 포함시키는 법적 검토 절차가 {사업수행계획서}에 계획되었는지 확인한다. 구두 합의나 이메일만으로 진행하는 계획은 위험으로 기록한다.
6	데이터 출처·라이선스 이력 관리 대장 작성 및 라이선스 변경·만료 대응 계획	모든 외부 데이터 항목의 출처 URL·수집 일자·라이선스 유형·AI 학습 허용 여부·버전·검토 담당자·검토 일자를 '데이터 출처 및 라이선스 관리 대장'으로 기록·관리하는 계획과 라이선스 변경 정기 모니터링 계획이 {사업수행계획서} 및 {요구사항정의서}에 정의되었는지 확인
7	라이선스 위반 발견 시 즉시 대응 절차 및 모델 재검토 계획	라이선스 위반 데이터가 AI 모델 학습에 사용된 사실이 발견된 경우 해당 데이터 즉시 학습 제외 절차, 모델 재학습 여부 판단 기준, 법적 대응(라이선스 소급 취득 협의·저작권자 통보) 절차가 {사업수행계획서}에 계획되었는지 확인

No	점검 항목	확인 방법
8	RFP 대비 데이터 출처·저작권 검토 요건 완전성 추적	RFP에 명시된 데이터 출처 투명성·저작권 준수·라이선스 검토 관련 요구사항이 {요구사항정의서} 및 {사업수행계획서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인한다. '저작권', '라이선스', '크롤링 허용', '데이터 출처', '귀속', '공개 데이터' 등 키워드로 항목을 식별한다.

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	① 데이터셋별 출처·버전·수집일시 목록화, ② 라이선스 종류·활용 허용 범위·귀속 의무 이행 방안 정의, ③ 모든 외부 소스 라이선스 유형·AI 학습 허용 범위 소스별 검토 계획, ④ 크롤링 이용약관·robots.txt 확인 절차, ⑤ 데이터 구매 계약 AI 학습 허용 조항 법적 검토 계획, ⑥ 출처·라이선스 이력 관리 대장 작성 계획, ⑦ 라이선스 위반 시 즉시 대응 절차가 모두 계획됨	현행 유지
보완권고 (△)	소스 목록은 있으나 라이선스 유형 미명시, 크롤링 이용약관 검토 절차 누락, 귀속 의무 이행 방안 미정의, 데이터 구매 계약 AI 학습 허용 조항 검토 미계획, 출처·라이선스 이력 관리 대장 계획 미수립, 라이선스 위반 시 대응 절차 미정의 등 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	외부 데이터 출처·저작권·라이선스 검토 계획이 {요구사항정의서} 및 {사업수행계획서}에 전혀 없거나 '공개 데이터는 자유롭게 사용 가능하다'는 잘못된 전제하에 검토 계획이 전무한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '공개 데이터 = 자유 이용 가능'은 잘못된 전제 — 반드시 라이선스 조건 확인: AI Hub·공공데이터포털·Kaggle 등에 공개된 데이터셋도 라이선스 조건이 데이터셋마다 다르며, 일부는 비상업적 이용만 허용(CC-BY-NC)하거나 동일 라이선스 재배포 의무(CC-BY-SA)가 부과된다. 사업자가 '공개 데이터이므로 자유롭게 사용 가능하다'는 전제로 라이선스 검토 계획을 수립하지 않은 경우 이를 라이선스 검토 미수행으로 명확히 판단하고 부적합으로 기록해야 한다.

2) 감리원은 법적 판단이 아닌 '검토 계획의 존재·체계성'을 점검: 감리원은 법률 전문가가 아니므로 특정 데이터 사용이 저작권법 위반인지 여부를 직접 판단해서는 안 된다. 감리원의 역할은 사업자가 소스별 라이선스 유형을 확인했는지, 법적 검토(법무팀·외부 법률 자문)를 수행하는 절차가

계획되었는지를 점검하는 것이며, 구체적 법적 해석은 사업자의 법무 검토 결과를 확인하는 수준에서 마무리한다.

3) 데이터 구매 계약서의 'AI 학습 허용' 조항 명시 여부 특별 주의: 전문 데이터 업체로부터 데이터를 구매했더라도 계약서에 'AI 학습 목적 이용 허용', '2차 가공 허용', '상업적 AI 서비스 배포 허용' 조항이 명시되지 않으면 데이터 공급 업체가 추후 계약 위반을 주장할 수 있다. {사업수행계획서}에 데이터 구매 계약 시 이 조항들을 계약서에 포함시키는 검토 절차가 계획되었는지 반드시 확인해야 한다.

4) 크롤링 데이터는 '수집 가능'과 'AI 학습 활용 가능'이 다름: robots.txt가 크롤링을 허용해도 해당 사이트 이용약관이 상업적 AI 학습 목적 이용을 금지하면 크롤링 자체는 가능하지만 AI 학습에 활용하는 것은 이용약관 위반이 된다. {사업수행계획서}에 크롤링 허용 여부(robots.txt)와 AI 학습 목적 이용 허용 여부(이용약관)를 구분하여 검토하는 절차가 계획되었는지 확인해야 한다.

5) 데이터 출처·라이선스 관리 대장은 분쟁 발생 시 핵심 방어 증거: 라이선스 분쟁이 발생했을 때 사업자가 적법한 검토를 통해 해당 데이터를 활용했음을 증명하려면 수집 시점의 라이선스 조건·검토 담당자·검토 일자 기록이 필요하다. 이 기록이 없으면 적법한 검토를 수행했더라도 입증이 불가능하므로, '데이터 출처 및 라이선스 관리 대장' 작성 계획이 없으면 강도 높은 보완 권고를 제시해야 한다.

AG_PM04-2-05 | 사용자매뉴얼 목차 설계

사업관리 | 실행통제단계 | 운영·모니터링 | 교육훈련기술이전

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 설계(상세설계) 단계에서 요구사항 정의 단계의 분석 내용이 {사용자매뉴얼}의 목차 설계에 빠짐없이 상세하고 구체적으로 반영되었는지를 확인하는 것을 목적으로 한다.

특히, AI 서비스의 경우 일반 SW와 달리 시스템이 출력하는 결과에 오류 가능성·한계가 내재되어 있어, 사용자가 AI의 특성과 한계를 인지하고 안전하게 활용할 수 있도록 안내하는 항목이 매뉴얼에 사전에 설계되었는지 여부를 제3자 관점에서 점검한다.

감리원은 {사용자매뉴얼} 목차가 단순 기능 안내에 그치지 않고, ①AI 한계점, ②안전한 사용법, ③사용자 책임 범위 등 AI 특화 항목을 포함하도록 설계되었는지를 문서 완결성 관점에서 검토한다.

2. 필요성

▶ AI 서비스는 동일한 입력에 대해 항상 일정한 출력을 보장하는 결정론적 소프트웨어와 달리, 확률적 특성으로 인해 사용 맥락·표현 방식에 따라 결과가 달라지거나 부정확한 정보를 사실처럼 제시하는 현상(환각)이 발생할 수 있다.

▶ 사용자가 AI 응답을 무비판적으로 수용하거나 시스템의 한계를 인지하지 못한 채 활용할 경우, 잘못된 의사결정·민원·법적 분쟁으로 이어질 수 있다. 따라서 {사용자매뉴얼}에는 AI의 작동 원리와 한계, 안전한 활용 방법이 명확히 포함되어야 한다.

▶ 설계 단계에서 매뉴얼 목차 구성을 점검함으로써, 구현 완료 후 누락 사항에 대한 재작업 비용과 민원 발생 리스크를 사전에 차단할 수 있다. 감리원은 목차 구성이 요구사항 분석서에 정의된 내용과 일치하는지를 추적(Traceability) 관점에서 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	{사용자매뉴얼} 목차 설계본의 제출 여부	산출물 목록과 {사용자매뉴얼} 목차 설계본이 제출되었는지 확인하고, 요구정의 단계 산출물과 동일 버전인지 대조
2	AI 한계점 안내 항목의 목차 포함 여부	목차에 AI의 확률적 특성, 환각 발생 가능성, 학습 데이터 기준 시점, 적용 제외 업무 등 한계점 안내 항목이 독립절로 포함되었는지 확인

No	점검 항목	확인 방법
3	안전한 사용법 안내 항목의 목차 포함 여부	목차에 프롬프트 작성 방법, 결과 검증 절차, 개인정보 입력 금지, 오류 발견 시 신고 방법 등 안전한 사용법 항목이 포함되었는지 확인
4	면책 안내 항목의 목차 포함 여부	목차에 AI 결과가 참고용이며 최종 판단과 책임은 이용자에게 있음을 고지하는 면책 안내 항목이 포함되었는지 확인
5	AI 특화 항목의 세부 구체화 수준	AI 특화 항목이 단순 목차 나열에 그치지 않고 2단계 이상 하위 항목으로 구체화되었는지 {사용자매뉴얼} 목차 계층 구조를 확인
6	{요구사항정의서}와의 추적성	도출된 사용자 안내 관련 요구사항이 목차 항목과 1:1로 매핑되는지 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{사용자매뉴얼} 목차에 AI 한계점·안전한 사용법·면책 안내 항목이 모두 포함되고, {요구사항정의서}와의 추적성이 확인된 경우	현행 유지
보완권고 (△)	AI 특화 항목이 포함되어 있으나 세부 항목 구분이 없거나, 요구사항과 매핑되지 않은 항목이 일부 존재하는 경우	해당 항목 보완 요청 후 재확인 (사업자에게 보완 계획 제출 요구)
부적합 (X)	AI 한계점 또는 안전한 사용법 항목이 목차에 전혀 없는 등 핵심 항목이 2개 이상 누락되었거나, {사용자매뉴얼} 설계본 자체가 미제출된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 사용자매뉴얼 목차가 기능 목록 나열에 그치는지, 실제 사용자 관점의 학습 흐름을 반영하는지 구분하라: '기능명 + 설명' 형태의 목차는 형식적 산출물에 불과하다. 사용자가 AI 시스템을 처음 사용하는 시나리오 기준으로 단계적 학습이 가능한지 확인해야 한다.

2) AI 고유 기능(모델 한계·신뢰도·재시도 방법 등)에 대한 안내 목차가 포함되었는지 확인하라: 일반 SW 매뉴얼에는 없는 항목(AI 출력 불확실성 안내, 이상응답 시 대처 방법 등)이 누락되는 경우가 많다.

3) 매뉴얼 목차와 실제 산출물 구성이 일치하는지 추적성을 확인하라: 목차 설계서에는 있지만 실제 매뉴얼에 빠진 절이 있으면 감리 지적 대상이다.

4) '목차 존재'와 '내용 반영'을 혼동하지 말 것: 설계 단계 감리의 특성상 {사용자매뉴얼}이 완성본이 아닌 목차 설계본 수준으로 제출될 수 있다. 감리원은 "해당 항목이 목차에 존재하는가"를 기

준으로 점검하되, 항목 제목만 있고 세부 구조가 전혀 없는 경우에는 보완 권고로 처리하고, 다음 단계 감리에서 내용의 충실성을 재확인할 계획임을 감리보고서에 명기한다.

5) AI 한계점 서술의 구체성 수준 검토: 단순히 "AI의 한계가 있을 수 있음"과 같은 포괄적 문구는 사용자 보호 관점에서 불충분하다. "응답 정확도 한계", "특정 도메인 외 질문에 대한 응답 오류 가능성", "법적·의료적 조언 목적으로의 활용 불가" 등 구체적인 한계 유형이 항목으로 구분되도록 사업자에게 요청한다.

AG_PM04-2-06 | 하도급 보안통제

사업관리 | 실행통제단계 | 생명주기 관리 | 보안관리

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「개인정보 보호법」 · ISO/IEC 27001

1. 목적

본 가이드라인은 AI 서비스 구축 사업의 설계(상세설계) 단계에서 {사업수행계획서}에 정의된 하도급 업체 보안점검 절차와 데이터 반출입 시 하도급 보안통제 절차가 설계 산출물에 구체적이고 빠짐없이 반영되었는지를 점검하는 것을 목적으로 한다.

AI 서비스 구축 사업에서는 하도급 업체가 학습 데이터 전처리, 모델 파인튜닝, 프롬프트 엔지니어링 등 핵심 업무를 수행하는 경우가 많아, 기획/분석 단계에서 요구사항으로 정의된 보안통제 사항이 설계 단계 산출물에 명확히 구현 수준으로 구체화되었는지를 제3자 관점에서 확인하는 것이 핵심이다.

2. 필요성

▶ AI 서비스 구축 사업은 다수의 하도급 업체가 참여하여 학습 데이터, AI 모델, API 키 등 민감 자산을 처리하는 특성이 있다. 설계 단계에서 보안통제 절차가 구체적으로 설계되지 않으면 사업 완료 이후에도 보안 취약점이 잔존하고, 분쟁 발생 시 책임 소재가 불분명해지는 문제가 발생한다.

▶ 요구사항 설계 반영 확인: {사업수행계획서}에서 정의한 하도급 보안점검 절차 및 보안통제 요구사항이 설계 산출물에 구체적으로 반영되었는지 추적·확인이 필요하다.

▶ AI 특화 데이터 유출 위험 관리: AI 학습 데이터, 파인튜닝용 데이터, 추론 결과 등은 일반 소프트웨어와 달리 대규모 민감 데이터를 포함하므로, 반출입 통제 절차의 구체성 확보가 중요하다.

▶ 설계-구현 간 정합성 확보: 설계 단계에서 보안통제 절차가 명확히 설계되어야 이후 개발·테스트 단계에서 검증 기준으로 활용할 수 있다.

▶ 책임 경계 명확화: 하도급 업체와의 보안 책임 범위를 설계 산출물에서 명시함으로써 사업 완료 후 발생하는 책임 분쟁을 예방한다.

3. 검토방법

No	점검 항목	확인 방법
1	하도급 업체 보안점검 절차 설계 반영 여부	설계서에서 하도급 보안점검 주기(착수·월간·완료 시점), 담당 조직·역할 정의, 점검 결과 보고·조치 절차가 기술되었는지 확인

No	점검 항목	확인 방법
2	데이터 반출입 보안통제 절차 설계 반영 여부	데이터 반출입 절차서·아키텍처 설계서에서 AI 학습·파인튜닝·추론 데이터 유형 분류, 반출입 승인 프로세스(신청-승인-실행-기록), 암호화·마스킹 대상·방법이 정의되었는지 확인
3	하도급 업체 보안서약·NDA 체결 절차 설계 반영 여부	설계 산출물에서 보안서약서·NDA 체결 시점·방법, 계약 외 추가 보안 의무사항, 위반 시 제재 절차가 반영되었는지 검토
4	AI 데이터·모델 접근권한 관리 설계 반영 여부	아키텍처·접근권한 관리 설계서에서 하도급 업체별 접근 범위 정의, 최소권한 원칙 적용, 접근 로그 수집·이상행위 탐지 설계를 확인
5	보안사고 발생 시 대응 절차 설계 반영 여부	설계 산출물에서 하도급 관련 보안사고 유형 정의, 인지→보고→조치→사후분석 흐름, AI 데이터 유출 시 특수 대응 절차가 포함되었는지 확인
6	보안통제 절차와 {사업수행계획서} 간 요구사항 추적성 확보 여부	요구사항 추적 매트릭스(RTM) 또는 동등 문서에서 각 보안 요구사항과 설계 항목 간 1:1 매핑 여부를 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	하도급 보안점검 절차(주기·담당 조직·보고 체계), 데이터 반출입 승인·암호화 통제, 보안서약·NDA 체결 절차, 접근권한 관리·로그 수집, 보안사고 대응 절차가 설계 산출물에 구체적으로 반영되고 {사업수행계획서} 요구사항과 1:1 추적이 가능함	현행 유지
보완권고 (△)	보안통제 설계는 존재하나 점검 주기·반출입 승인 프로세스·위반 시 제재 절차·이상행위 탐지 설계 등 일부 요소가 누락되었거나 추상적으로 기술되어 구현 검증 기준으로 활용하기 어려움	해당 항목 보완 요청 후 재확인
부적합 (X)	하도급 업체 보안점검 절차와 데이터 반출입 보안통제 절차가 설계 산출물에 전혀 반영되지 않았거나 {사업수행계획서} 요구사항과의 추적이 불가능함	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항과 설계 산출물의 대조 순서: {사업수행계획서}에서 하도급 보안 관련 항목을 먼저 목록화한 후 설계 산출물을 검토하는 순서로 진행하면 누락 항목을 빠짐없이 확인할 수 있다. 반대로 설계 산출물만 먼저 읽으면 무엇이 누락되었는지 파악하기 어렵다.

2) AI 특화 보안 항목의 추상적 표현 포착: '데이터를 안전하게 반출한다', '접근을 적절히 통제한다' 등의 추상적 표현이 설계 산출물에 있으면 승인 주체·암호화 방식·통제 대상 데이터 유형 등 구현 수준으로 구체화하도록 권고해야 한다. AI 프로젝트에서 흔히 나타나는 패턴이므로 주의 깊게 검토한다.

3) 산출물 미작성과 내용 미반영의 구분: 별도의 하도급 보안 설계서가 없더라도 아키텍처 설계서·운영 절차서 등에 해당 내용이 분산 반영되어 있을 수 있다. 감리원은 산출물 자체의 미작성과 내용의 미반영을 구분하여 판정하고, 분산 반영된 경우 해당 위치를 감리보고서에 명시한다.

AG_PM04-3-01 | 평가단 XAI 평가결과

사업관리 | 종료단계 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE

1. 목적

종료(시험/전개) 단계에서 AI 시스템이 설계 요구사항에 따라 구현되었는지 확인하는 절차 중, 사용자 평가단(설명가능성, XAI)이 실제 현업 사용자를 대상으로 수행한 평가 결과가 적정하게 작성·확보되었는지를 점검하는 것이 목적이다.

본 가이드라인은 XAI 기능(AI가 왜 그 결론을 도출했는지 사용자에게 설명하는 기능)의 실제 유용성과 사용자 수용성을 평가한 결과 보고서가 존재하는지, 그 내용이 감리원이 판단할 수 있는 수준으로 완결성 있게 작성되었는지 여부를 검토 대상으로 한다.

감리원은 XAI 기술 자체의 정확성이나 알고리즘 로직을 점검하는 것이 아니라, 요구된 평가 절차가 수행되고 그 결과가 산출물로 남겨졌는지를 문서 완결성 관점에서 검토한다.

2. 필요성

▶ AI 시스템은 일반 IT 시스템과 달리 동일한 입력에도 매번 다른 출력을 생성할 수 있으며, 그 판단 근거를 사용자가 이해하지 못하면 시스템의 신뢰성이 저하된다. TTA 2024 신뢰할 수 있는 인공지능 개발안내서(생성AI 기반 서비스 분야)에서는 사용자가 AI 결과를 신뢰하고 활용할 수 있도록 설명가능성(Explainability) 확보를 권고하고 있다.

▶ 그러나 공공 AI 서비스 사업의 경우, XAI 평가를 형식적으로만 수행하거나 결과보고서를 작성하지 않은 채 시스템을 운영 전개하는 사례가 발생할 수 있다. 이 경우 다음과 같은 리스크가 발생한다.

▶ 사용자가 AI 결과를 맹목적으로 수용하거나 반대로 전혀 활용하지 않는 양극단이 발생하여 시스템 도입 효과가 저하될 수 있다.

▶ XAI 기능에 대한 사용자 피드백이 누락되어 운영 단계에서의 지속적 개선 기반이 마련되지 않는다.

▶ 발주기관이 시스템 검수 시 XAI 기능의 실효성을 판단할 근거가 없어 분쟁이 발생할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	평가단 구성의 적정성 확인	[수작업 점검] 'AI평가단결과보고서' 내 참여자 명단, 소속, 직군 정보 검토
2	XAI 기능 평가 범위 명확성	[수작업 점검] 보고서 내 평가 대상 XAI 기능 목록 확인
3	평가 절차 및 방법론 기술 여부	[수작업 점검] 보고서 내 평가 방법 기술 내용 검토
4	설문 결과의 정량·정성 분석 포함 여부	[수작업 점검] 설문 결과표, 그래프, 의견 요약 검토
5	결과 기반 개선 사항 도출 여부	[수작업 점검] 보고서 내 개선 의견 및 조치 계획 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	현업 사용자로 구성된 평가단이 XAI 기능을 평가한 'AI평가단결과보고서'가 존재하고, 평가 범위·절차·방법론, 정량·정성 분석, 개선 사항 도출까지 완결성 있게 작성된 경우	현행 유지
보완권고 (△)	평가는 수행되었으나 평가단 구성의 대표성 부족, 평가 범위·방법론 기술 미흡, 분석 내용 불충분, 개선 사항 미도출 등 보고서의 완결성이 일부 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	XAI 사용자 평가가 수행되지 않았거나, 결과보고서 및 평가 수행을 입증할 대체 증거가 전혀 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'AI평가단결과보고서' 부재 시 대응: 사업자가 "평가단 운영은 했으나 보고서를 별도로 작성하지 않았다"고 주장하는 경우, 평가 수행을 입증할 수 있는 대체 증거(참석자 서명부, 회의록, 이메일 기록 등)를 요청하여 이행 여부를 판단한다. 증거가 전혀 없는 경우 부적합으로 처리한다.

2) 설문 결과의 통계적 신뢰성 여부를 확인한다: 평가단 인원이 지나치게 적거나 특정 직군에 편중된 경우 평가 결과의 대표성이 떨어지므로, 참여자 수·소속·직군 분포가 실제 사용자 집단을 대변할 수 있는 수준인지 확인한다.

3) 감리 범위의 한계를 준수한다: 감리원은 XAI 알고리즘 자체의 기술적 정확성을 평가하는 것이 아니라, 요구된 평가 절차의 수행 여부와 결과보고서의 완결성을 문서 관점에서 점검하는 데 집중한다.

4) 설문 문항이 XAI 기능의 유용성을 실제로 측정하는지 확인한다: '시스템에 만족하십니까' 수준의 일반 만족도 문항만으로는 설명가능성의 실효성을 판단할 수 없다. 감리원은 설명 내용의 이해도, 판단 근거로서의 활용 여부 등 XAI 특화 문항이 포함되었는지 확인한다.

5) 평가 결과가 개선 조치로 연결되었는지 추적한다: 개선 의견만 나열되고 조치 계획·담당자·기한이 없으면 지속적 개선 기반이 마련되지 않는다. 감리원은 도출된 개선 사항이 조치 계획과 함께 기재되고, 미조치 항목의 사유가 설명되었는지 확인한다.

AG_PM04-3-02 | 윤리기준 최종검토

사업관리 | 종료단계 | 생명주기 관리 | 법령/AI윤리

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「인공지능기본법 이행 가이드라인」(과기정통부, 2026.1.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

AI 서비스 구축 사업에서 법무·거버넌스 팀이 수행한 AI 윤리기준 최종 검토가 절차에 따라 완료되었는지, 그 결과가 {결과보고서}에 적정하게 반영되고 최종 서명이 이루어졌는지 확인하는 것이 목적이다.

특히 AI 시스템이 서비스 오픈 전 법적·윤리적 기준을 충족하였음을 확인하는 문서적 근거가 마련되었는지 점검한다.

2. 필요성

▶ AI 시스템은 편향성·차별·개인정보 침해·저작권 위반 등 일반 소프트웨어에는 없는 고유한 윤리적 리스크를 내포하고 있다. 시스템이 구현된 이후라도 이러한 리스크를 점검하지 않고 서비스에 투입하면 사회적 피해와 법적 분쟁이 발생할 수 있다.

▶ 따라서 개발과 테스트를 통해 구현된 AI 기능이 사전에 정의된 AI 윤리기준을 충족하는지 최종적으로 확인하고, 그 결과를 공식 문서로 남기는 절차가 필수적이다.

▶ 공공 AI 서비스의 경우, 이해관계자(국민) 보호 의무 및 공공기관 책임성 요건이 특히 강조되므로 윤리 검토의 완결성과 공식 서명 절차 이행 여부를 감리 관점에서 독립적으로 검증하는 것이 중요하다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 윤리기준 최종 검토 절차 이행 여부	법무·거버넌스 팀이 정해진 절차에 따라 AI 윤리기준 최종 검토를 수행하였는지, 검토 결과가 {결과보고서}에 공식적으로 반영되었는지 확인한다. 검토 섹션에 검토 일자, 검토자, 검토 범위가 명시되었는지 함께 확인
2	최종 서명 이행 및 서명자 적정성	윤리 검토 완료를 확인하는 최종 서명이 이루어졌는지, 서명자가 법무·거버넌스 팀 또는 AI 윤리 담당자로 적절한 권한과 역할을 보유한 자인지 확인
3	편향성·공정성 최종 평가 결과 적정성	수행된 편향성·공정성 측정 결과가 사전 정의된 허용 기준(임계값) 이내인지, 기준 초과 시 조치 내용과 재평가 결과가 기록되었는지 확인

No	점검 항목	확인 방법
4	AI 윤리원칙 준수 체크리스트 완결성	정의된 AI 윤리원칙(투명성·공정성·책임성·안전성 등)이 체크리스트에 항목별로 반영되었는지, 미충족 항목에 대한 조치 계획이 수립되었는지 확인
5	저작권·법규 준수 검토 의견서 구비	AI 학습데이터와 생성 출력물에 대한 저작권 및 개인정보보호법 적합성 검토 의견서가 구비되었는지, 의견서에 법무팀 서명이 포함되어 있는지 확인
6	조건부 승인 항목의 사전 조치 완료 여부	조건부 승인이 부여된 경우 사전 조치 이행 결과가 {결과보고서}에 반영되었는지, 미조치 항목이 있는 경우 그 사유와 향후 계획이 명시되었는지 확인
7	Human-in-the-Loop 설계와의 일치성	요구정의·설계 단계에서 계획된 Human-in-the-Loop 구조(AI 출력에 대한 인간 최종 확인 체계)가 실제 구현된 시스템 및 운영 절차에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{결과보고서}에 AI 윤리기준 검토 섹션이 존재하고 일자·검토자·범위가 명시되었으며, 법무·거버넌스 팀 서명이 완료되고, 편향성·공정성 평가 결과가 허용 기준 이내이며, AI 윤리원칙 준수 체크리스트가 항목별로 완결되고, 저작권·법규 준수 검토 의견서가 법무팀 서명과 함께 구비되며, 조건부 승인 항목의 사전 조치가 완료되고 Human-in-the-Loop 설계가 실제 시스템·운영 절차에 반영되었음이 확인된 경우	현행 유지
보완권고 (△)	서명란은 존재하나 실제 서명이 누락된 경우, 체크리스트 일부 항목이 공란인 경우, 편향성 기준 초과 항목의 조치 계획만 수립되고 실제 조치가 미완료된 경우 등 일부 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{결과보고서}에 AI 윤리기준 검토 내용이 전혀 없거나, 법무·거버넌스 팀 최종 서명이 없거나, 체크리스트 또는 편향성·공정성 평가 보고서가 미작성된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 형식 충족과 실질 이행을 구분하라: AI 윤리기준 최종 검토 결과서가 작성되었더라도 검토 항목이 형식적으로만 나열된 경우가 있다. 각 항목에 실제 검토 근거(사용 기준, 측정값, 법규 조문 등)가 구체적으로 기술되었는지 확인하고, 단순히 "해당 없음" 또는 "확인 완료"만 기재된 항목은 세부 근거 제출을 요청한다.

2) 서명자의 역할과 권한을 반드시 검증하라: 서명이 존재하더라도 서명자가 AI 윤리 검토 권한이 없는 개발 담당자이거나 단순 확인자인 경우가 있다. 사전에 프로젝트 거버넌스 계획서 또는 역할 정의서를 확보하여, 서명자가 실제로 법무·거버넌스 역할을 맡은 자인지 대조한다.

3) 편향성 허용 기준이 사전에 정의되었는지 먼저 확인하라: 편향성 측정 결과만 있고 허용 기준이 없다면 판단이 불가능하다. 허용 기준은 요구정의 또는 설계 단계 산출물에서 확인하고, 기준이 없는 경우 이를 지적하여 기준 수립 및 재평가를 요구한다.

4) 조건부 승인 항목의 추적 관리를 점검하라: 이전 감리 시점에서 조건부로 승인된 AI 윤리 관련 항목이 있다면, 해당 조치 완료 여부를 이번 감리에서 반드시 추적한다. 전 단계 감리 결과보고서(또는 감리 조치 계획서)를 확보하여 미결 항목 목록을 사전에 파악한 후 현장에 임한다.

5) Human-in-the-Loop 확인 시 운영 매뉴얼을 함께 검토하라: Human-in-the-Loop 구조는 시스템 설계 문서만으로는 실제 구현 여부를 확인하기 어렵다. 운영자 매뉴얼 또는 업무 절차서에 AI 출력물에 대한 인간 검토 단계가 명시되어 있는지 함께 확인하고, 필요한 경우 담당자 인터뷰를 통해 실제 운영 시 인간 개입이 이루어지는지 확인한다.

AG_PM04-3-03 | 하도급업체 보안점검

사업관리 | 종료단계 | 생명주기 관리 | 보안관리

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「개인정보 보호법」 · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업에 참여한 하도급 업체에 대한 보안점검이 적정하게 완료되었는지와 AI 학습데이터 및 모델의 반출입 보안 서약 및 이력 관리가 제대로 이루어졌는지를 감리원이 제3자 관점에서 점검하는 것을 목적으로 한다.

특히, 보안점검의 형식적 완료 여부에 그치지 않고, 하도급 업체가 실제로 AI 학습데이터·모델 파일 등을 무단 반출하지 않도록 절차가 수립되고 이행되었는지를 산출물 검토를 통해 확인한다.

2. 필요성

▶ AI 서비스 구축 사업에서는 학습데이터 구축, 모델 파인튜닝, 테스트 등의 업무가 여러 하도급 업체에 분산 수행되는 경우가 많다. 이 과정에서 민감한 공공 데이터나 개인정보가 포함된 학습 데이터셋 또는 완성된 AI 모델 가중치(Weight) 파일이 무단으로 반출될 경우, 국가 데이터 유출 또는 지식재산 침해로 이어질 수 있다.

▶ 종료 단계의 보안 점검은 이러한 위험을 마지막으로 통제할 수 있는 시점이므로, 단순한 보안점검 완료 확인 외에 AI 특유의 반출입 이력 관리 및 서약 이행 여부까지 함께 점검할 필요가 있다. 구체적인 필요성은 다음과 같다.

▶ AI 학습데이터에는 개인정보나 공공기관 민감 데이터가 포함될 수 있어, 하도급 업체 직원에 의한 무단 반출 시 법적 책임이 발주기관에 귀속될 수 있다.

▶ 완성된 AI 모델 파일(가중치, 파인튜닝 결과물 등)은 고부가가치 지식재산으로, 외부 유출 시 경쟁력 훼손 또는 국가 안보 침해로 이어질 수 있다.

▶ 보안 서약서 및 반출입 대장 미구비는 분쟁 발생 시 책임 소재 입증을 불가능하게 하여 발주기관이 법적·행정적 불이익을 받을 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 관련 하도급 업체 보안점검 이행 여부	AI 개발·학습 업무에 참여한 하도급 업체에 대해 {보안점검결과서}가 발급되었는지 확인. 보안점검이 시스템 전개 이전에 완료되었는지 일정 및 서명 일자를 대조 검토.

No	점검 항목	확인 방법
2	보안 서약서 징구 (AI 데이터·모델 취급자)	학습데이터·파인튜닝 데이터·모델을 취급한 하도급 임직원 및 외부인력이 '보안 서약서'를 작성하고 서명하였는지 확인. 서약 내용에 AI 산출물(데이터·가중치 파일 등) 외부 유출 금지 조항이 포함되어 있는지 검토.
3	AI 데이터·모델 반출입 절차 수립 여부	'데이터·모델 반출입 통제 절차서'가 마련되었는지 확인. 절차서에 ① 반출입 신청·승인 프로세스, ② 승인권자 지정, ③ 비승인 반출 시 조치 방안이 포함되어 있는지 검토.
4	AI 데이터·모델 반출입 이력 관리 적정성	'AI 데이터 취급 내역 대장'에 반출입 일시, 담당자, 데이터 범위(건수·용량), 목적, 승인자가 모두 기록되어 있는지 확인. 무승인 반출입 이력이 없는지 대장 내 이상 항목을 점검.
5	하도급 보안 관리 계획 수립 여부	{사업수행계획서}에 하도급 업체 보안관리 계획(정보보호 서약, 장비 반입 통제, 화면보안 적용 등)이 포함되어 있는지 확인. 실제 이행 여부를 {보안점검결과서}와 대조 검토.
6	AI 학습데이터 내 개인정보 포함 여부 사전 점검	학습·테스트용 데이터셋에 개인정보(이름·주민번호·의료정보 등)가 포함되어 있었는지, 포함 시 사전 비식별화(익명화·가명화) 처리 완료 여부를 관련 처리 결과 문서로 확인. 이 항목은 데이터베이스 감리원과 협의하여 중복 없이 점검.

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{사업수행계획서}에 하도급 보안관리 계획이 수립되고 하도급 업체 보안점검이 완료되었으며, AI 데이터·모델 취급자의 보안 서약, 반출입 절차 및 이력 대장이 모두 구비·이행되고, 학습·테스트 데이터의 개인정보 포함 여부 사전 점검과 비식별화 처리가 확인되는 경우	현행 유지
보완권고 (△)	보안점검은 수행되었으나 서약서의 AI 특화 조항 부재, 반출입 대장 기재 항목 일부 누락 등 이력 관리가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	하도급 업체 보안점검이 미이행되었거나, 보안 서약서·반출입 대장 등 핵심 통제 산출물이 존재하지 않는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 사업 하도급 보안점검은 "보안점검결과서 있음"으로 종결하지 말 것: 일반 IT 사업의 보안점검과 달리, AI 사업에서는 학습데이터 및 모델 파일이 고부가가치 정보 자산이다. 보안점검결과

서가 존재하더라도, 해당 업체가 학습데이터를 실제로 취급했는지를 추가로 확인하고, 취급한 경우 반출입 이력 대장이 있는지 반드시 확인한다.

2) 보안 서약서의 AI 특화 조항 유무 확인: 기존의 일반 보안 서약서에는 AI 학습데이터, 모델 가중치, 파인튜닝 결과물 등에 대한 명시적 조항이 없는 경우가 많다. 감리원은 서약서 내용에 "AI 학습 데이터 및 모델 파일"을 특정하는 조항이 있는지를 확인하여, 단순 출력·저장 금지 수준의 일반 조항에 그친 경우 미흡으로 판정한다.

3) 반출입 대장은 기재 완전성이 핵심: 반출입 대장의 목적은 이력 추적이다. 담당자 이름, 반출입 일시, 데이터 범위, 승인자 서명 중 하나라도 누락된 경우 이력 추적이 불가능하므로 부적합으로 판정해야 한다. 특히 "기타" 또는 "전체 데이터"와 같이 범위가 불명확한 기재는 실효성 없는 대장으로 본다.

4) 하도급 투입 인력 목록과 서약서 서명 명단의 교차확인: 사업수행계획서 또는 인력 투입 현황표에 기재된 하도급 투입 인원과 보안 서약서 서명 명단을 대조한다. 중간 투입 인력(추가 투입·교체 인력)이 서약서에서 누락된 경우가 많으므로, 투입 기간과 서약 일자도 함께 확인한다.

5) 현장 인터뷰를 통한 절차 실제 이행 여부 확인: 반출입 절차서와 대장이 서류상으로는 완비되어 있더라도, PM 또는 보안 담당자 인터뷰를 통해 실제로 절차가 이행되었는지를 확인하는 것이 바람직하다. 예를 들어 "반출 승인 요청을 어떤 방식으로 받았는가?", "비승인 반출 시도가 있었는가?" 등의 질문을 통해 절차의 실질적 이행 여부를 점검한다.

AG_PM04-3-04 | 이해관계자 인수테스트

사업관리 | 종료단계 | 생명주기 관리 | 시범운영/인수테스트

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업에서 이해관계자 인수테스트(UAT)가 체계적으로 수행되었는지, 그리고 실제 사용 부서의 최종 인수 승인이 적정한 절차에 따라 이루어졌는지를 {인수테스트결과서}를 중심으로 검토하는 것이 목적이다.

AI 서비스는 일반 정보시스템과 달리 출력 결과의 가변성·비결정론적 특성으로 인해 사전에 합의된 인수기준(Acceptance Criteria) 없이 UAT가 수행될 경우 검수 분쟁으로 이어질 위험이 높다.

따라서 UAT 시나리오가 AI 특성을 반영하여 작성되었는지, 인수 승인 주체와 절차가 명확한지를 문서 검토 관점에서 점검한다.

2. 필요성

▶ AI 서비스는 동일한 입력값에도 출력이 달라질 수 있고, 환각(Hallucination) 등 AI 고유 오류 유형이 존재하므로, 일반 소프트웨어의 기능 테스트 방식만으로는 인수 적정성을 판단하기 어렵다. 다음과 같은 이유로 별도의 점검이 필요하다.

▶ AI 특화 인수기준 부재 위험: 요구사항 단계에서 정의된 AI 성능 기준(정확도, 응답 품질 등)이 UAT 시나리오에 반영되지 않은 경우, 인수 검수 자체가 무효화될 수 있음

▶ 이해관계자 참여 불충분 위험: 실제 업무를 수행하는 현업 부서가 아닌 IT 부서만 UAT에 참여하면 AI 서비스의 업무 적합성이 검증되지 않는 문제가 발생함

▶ 인수 승인 주체 불명확 위험: AI 서비스 특성상 모델 성능, 콘텐츠 안전성, 개인정보 처리 등 다양한 영역이 포함되므로, 승인 주체 및 권한이 분산될 경우 최종 인수의 법적 효력에 문제가 생길 수 있음

▶ AI 결과물 재현성 문제: 비결정론적 출력으로 인해 테스트 결과의 재현이 어려워 합격/불합격 판정 기록이 불명확해질 위험이 있음

3. 검토방법

No	점검 항목	확인 방법
1	AI 특화 인수기준의 UAT 반영 여부 (AI 성능 기준, 안전성 기준 등)	{인수테스트결과서}와 'AI 서비스 인수기준서'를 교차 검토하여 AI 성능·안전성 기준이 UAT 시나리오에 반영되었는지 확인

No	점검 항목	확인 방법
2	UAT 참여 이해관계자의 적정성 (현업 부서 참여 여부)	{인수테스트결과서}의 참여자 명단에서 실제 업무를 수행하는 현업 부서 담당자의 소속·직위 확인
3	AI 응답 품질 평가 방식의 적정성 (정량·정성 평가 구성)	'AI 응답 품질 평가 기록'에서 수치 기반 정량 평가와 정성 평가 기준의 병행 구성 여부 확인
4	결함 처리 현황 및 재테스트 완료 여부	'결함 관리 대장'에서 결함별 조치 결과와 재테스트 완료 기록 확인
5	최종 인수 승인 주체 및 서명 적정성	{인수테스트결과서}의 최종 승인란에서 발주기관 책임자의 승인 권한과 서명 여부 확인
6	AI 모델 버전 및 배포 형상 일치 여부	'AI 서비스 인수기준서'와 모델 버전 관리 이력을 대조하여 UAT 대상 모델과 배포 예정 모델의 일치 여부 확인
7	운영 이관 준비 문서의 완비 여부 (운영 매뉴얼, 장애 대응 계획)	'운영 매뉴얼', '장애 대응 계획' 등 운영 이관 문서의 작성 완료 및 AI 특화 항목 포함 여부 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 특화 인수기준이 UAT 시나리오에 반영되고 현업 부서가 참여하였으며, AI 응답 품질이 정량·정성 평가로 병행 검증되고, 결함 처리·재테스트 기록과 발주기관 책임자의 최종 승인, 모델 버전 일치 및 운영 이관 문서가 모두 확인되는 경우	현행 유지
보완권고 (△)	UAT는 수행되었으나 인수기준의 수치화 미흡, 현업 참여 증빙 부족, 잔존 결함 합의 문서 누락 등 일부 항목이 불완전한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 특화 인수기준 없이 UAT가 수행되었거나, 인수 승인 절차·산출물이 존재하지 않아 인수의 효력을 인정하기 어려운 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 인수기준의 수치화 유도: "AI 서비스는 정확하게 답변해야 한다"와 같은 모호한 인수기준은 검수 분쟁의 주원인이다. 감리 의견으로 "답변 정확도 90% 이상", "유해 콘텐츠 차단율 99% 이상" 등 측정 가능한 기준으로 명확화하도록 권고하여야 한다.

2) 현업 부서 참여 증빙 확인: IT 부서 주도의 UAT에서는 현업 부서 담당자의 서명이 형식적으로 처리되는 경우가 있다. {인수테스트결과서} 서명자의 소속·직위를 반드시 확인하고, 실제 업무를 수행하는 현업 전문가인지 인터뷰로 확인하는 것이 효과적이다.

3) 운영 단계 AI 특화 문서 사전 점검: 운영 이관 전 단계임에도 AI 서비스의 모델 드리프트 감지 계획, 재학습 절차, Human-in-the-Loop 운영 방안이 운영 이관 문서에 명시되어 있는지 사전에 확인해야 한다. 이는 운영 단계 사항으로 구축 감리 범위 밖이지만, 개발 과정에서 반드시 산출물로 작성되어야 하는 항목이다.

4) AI 모델 버전 불일치 검토 시 유의사항: AI 모델 버전 관리는 일반 소스코드 형상관리와 별도로 수행되는 경우가 많다. MLflow, DVC 등 AI 전용 형상관리 도구의 이력 화면 캡처본이나 출력물을 제출받아 확인하면 효과적이며, 모델 파일 해시값 비교 방식으로 일치 여부를 간접 확인할 수도 있다.

5) 잔존 결함 합의 문서 존재 여부 확인: 현실적으로 모든 결함이 해소된 상태에서 인수 승인이 이루어지기 어렵다. 잔존 결함이 있는 경우, 결함 수용 합의서(또는 이행각서) 등 발주기관이 인지하고 수용하였음을 증명하는 문서가 존재하는지를 반드시 확인하여야 한다.

4.2. 품질보증

요구정의와 시험 단계에서 산출물 품질과 편향·신뢰성 검증 가이드를 수록한다. (총 7건)

요구정의·분석 (5건)

AG_SD41-4-01 편향성 정량지표 점검

AG_SD41-4-02 데이터 버전관리 요건 및 모델버전 종속 데이터 형상관리 요건

AG_SD41-4-03 예외대응 요건

AG_SD41-4-04 AI산출물 품질보증 계획

AG_SD41-4-05 산출물 작성표준

시험 (2건)

AG_SD44-4-01 편향-차별훈련 이행 점검

AG_SD44-4-02 AI 신뢰성 테스트·자동화 평가 파이프라인 점검

AG_SD41-4-01 | 편향성 정량지표 점검

품질보증 | 요구정의·분석 | 인공지능 모델 개발 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP·MEASURE · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 모델의 편향성 정량지표와 모델 공평성 평가 지표 및 목표치를 {요구사항정의서}에 구체적이고 측정 가능한 형태로 정의하였는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, '공정하게 작동해야 한다', '편향이 없어야 한다'와 같은 선언적 문구에 그치지 않고, 어떤 집단에 대해 어떤 지표로 어느 수준(수치 목표치)까지 달성해야 하는지가 요구사항으로 명확히 정의되었는지 확인하는 것이 핵심이다.

이는 이후 설계·개발·검사 단계에서 공정성 요건의 추적 가능성과 검증 가능성을 확보하기 위한 전제 조건이 된다.

2. 필요성

▶ AI 시스템은 학습 데이터에 내재된 역사적 차별, 표본 불균형, 특정 집단의 과소 대표 등으로 인해 특정 집단에 불리한 결과를 생성할 수 있는 편향성(Bias) 위험을 구조적으로 내포하고 있다. 이러한 위험은 개발 초기 단계인 요구정의 시점에 정량적 목표치로 정의되지 않으면, 이후 단계에서 '적절하다'는 주관적 판단에 의존하게 되어 통제가 불가능해진다.

▶ 분쟁 예방: 편향성 지표와 목표치가 수치로 요구사항에 명시되어야만 개발 후 합격·불합격을 객관적으로 판정할 수 있다. 수치화 없이는 '편향이 있다/없다'를 둘러싼 극심한 분쟁이 발생한다.

▶ AI 공정성 법·제도 준수: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야) 및 일반 분야는 공정성 목표를 요구사항으로 명시할 것을 권고하며, NIST AI RMF MAP·MEASURE 함수에서도 편향성 측정 방법론 수립을 요구하고 있다.

▶ 이후 단계 추적 가능성 확보: 정의된 지표와 목표치는 설계서, 시험 시나리오, 검사 결과서까지 추적되어야 한다. 초기에 정의되지 않으면 이후 단계에서 보완이 사실상 불가능하다.

3. 검토방법

No	점검 항목	확인 방법
1	편향성 정량지표 및 허용 임계값 정의 여부	{요구사항정의서}에서 인구통계 동등성(Demographic Parity) ·균등화 승산비(Equalized Odds) 등 편향성 정량지표와 허용 가능 임계값(목표치)이 요구사항으로 명시되었는지 확인. '공정하게 작동' 등 선언적 문구만 있으면 미흡으로 판정
2	집단별 공평성 평가 지표·목표치 명시 여부	{요구사항정의서}에서 등가 기회율·위양성률 격차 등 공평성 평가 지표와 목표치가 성별·연령 등 보호 집단별로 구분되어 명시되었는지 확인
3	RFP 공정성 요구사항의 반영 완전성	제안요청서(RFP)의 AI 공정성 요구사항 목록을 {과업대비표}와 대조하여 {요구사항정의서}에 누락 없이 반영되었는지 교차 확인
4	공정성·편향성 평가 계획의 구체성	'AI 공정성·편향성 평가 계획서'에서 평가 방법, 학습·검증 데이터 분리 기준, 측정 시점(개발 중·인수 시·운영 중)이 구체적으로 기술되었는지 확인
5	편향성의 AI 고유 위험 식별 및 대응 방안	'AI위험관리계획서(초안)'에서 편향성이 AI 고유 위험 항목으로 식별되고, 발생 가능성·영향도 평가와 대응 방안이 기술되었는지 확인
6	공평성 지표 모니터링 주기 및 초과 시 조치 절차	요구사항에 모니터링 주기와 임계값 초과 시 보고·조치 절차(보고 대상, 기한, 재학습 여부)가 명시되었는지 확인하고, 운영 담당 부서 지정 여부를 사업자 면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	편향성 정량지표·공평성 평가 지표와 집단별 허용 임계값(목표치)이 {요구사항정의서}에 수치로 명시되고, RFP 공정성 요건이 누락 없이 반영되었으며, 평가 계획서·위험관리 계획서·모니터링 절차까지 구체적으로 정의됨	현행 유지
보완권고 (△)	편향성 지표는 명시되었으나 목표치 수치 미정의, 일부 집단·요건 누락, 평가 방법·측정 시점 또는 모니터링·조치 절차가 미흡하게 기술됨	해당 항목 보완 요청 후 재확인
부적합 (X)	편향성·공평성 요구사항이 '공정하게 작동해야 한다' 수준의 선언적 문구뿐이거나 정량지표·목표치가 {요구사항정의서}에 전혀 정의되지 않음	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 표현의 수치화 요구: 'AI는 공정하게 동작해야 한다', '편향이 최소화되어야 한다'와 같은 문구는 극심한 분쟁의 원인이 된다. 감리원은 '집단 간 위양성률 차이 ≤ 3%p', '인구통계 동등성 지수 ≥ 0.95'와 같이 측정 가능한 형태로 요구사항을 구체화하도록 사업자에게 요구해야 한다.

2) 도메인별 적용 가능 지표를 구분하여 검토: 모든 AI 서비스에 동일한 편향성 지표를 적용할 수는 없으며, 이진 분류 모델·순위 추천 모델·자연어 생성 모델은 각각 다른 지표가 적용된다. 감리원은 {요구사항정의서}에 기재된 서비스 유형에 적합한 지표가 선택되었는지 사업자로부터 근거를 제출받아 확인한다.

3) 전체 평균이 아닌 집단별·소수 클래스 기준 포함 점검: 공평성 지표는 전체 평균 성능이 아니라 집단별 성능 편차와 소수 클래스·엣지 케이스 기준을 포함해야 실질적인 편향 통제가 가능하므로, 집단별 목표치가 각각 정의되었는지 확인한다.

4) RFP와 요구사항정의서의 공정성 요건 불일치 확인: RFP에 명시된 AI 공정성 요구사항이 {과업대비표}·{요구사항추적표}를 통해 누락이나 완화 없이 {요구사항정의서}에 반영되었는지 교차 검토하고, 불일치 발견 시 구체적 근거와 함께 감리보고서에 기록해야 한다.

AG_SD41-4-02 | 데이터 버전관리 요건 및 모델버전 종속 데이터 형상관리 요건

품질보증 | 요구정의·분석 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」·「신뢰할 수 있는 인공지능 개발안내서」(TTA)·NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 학습 데이터셋·전처리 파이프라인·하이퍼파라미터 등 데이터 관련 산출물을 모델 버전과 연계하여 관리하는 데이터 버전관리 시스템(DVC: Data Version Control) 도입 요건, 데이터-모델 버전 매핑 기준, 데이터 변경 이력 추적 요건, 특정 모델 버전 재현을 위한 데이터 복원 요건이 {요구사항정의서}에 정의되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

데이터 형상관리 체계 미수립으로 인해 AI 모델 성능 이상 발생 시 원인 추적 불가·이전 버전 모델 재현 불가·감사 대응 불가 상황을 사전에 차단한다.

2. 필요성

▶ AI 모델은 동일한 알고리즘·코드로 학습하더라도 학습 데이터셋이 다르면 완전히 다른 성능 결과를 낸다. 학습 데이터셋과 AI 모델 버전이 연계 관리되지 않으면 모델 성능 저하 원인이 '코드 변경'인지 '데이터 변경'인지 파악할 수 없어 디버깅과 감사 대응이 불가능해지므로, {요구사항정의서}에 데이터 버전관리 요건이 반드시 정의되어야 한다.

▶ 데이터 변경으로 인한 AI 모델 성능 저하 원인 추적 불가 방지: 학습 데이터가 업데이트될 때마다 버전이 기록되지 않으면 '어떤 데이터로 학습했을 때 모델 성능이 가장 좋았는가', '현재 성능 저하가 새로운 학습 데이터 품질 문제 때문인가'를 파악할 수 없어 성능 회복 조치에 수 주에서 수 개월이 소요될 수 있다.

▶ AI 모델 재현성(Reproducibility) 보장 불가로 인한 감사·검수 분쟁 방지: 감리·감사 또는 성능 이슈 분쟁 시 '특정 AI 모델 버전을 동일한 조건으로 재현하라'는 요구에 응하려면 해당 모델 학습에 사용된 데이터셋 버전·전처리 코드 버전·하이퍼파라미터를 정확히 복원할 수 있어야 한다. 데이터 버전관리 체계가 없으면 재현이 불가능하여 분쟁 대응이 곤란해진다.

▶ 데이터 파이프라인 변경 영향 추적 불가로 인한 AI 모델 품질 저하 방지: 전처리 파이프라인(결측값 처리 방법·정규화 기준·특성 선택 기준)이 변경될 때 해당 변경이 AI 모델 성능에 미친 영향을 추적하려면 변경 전·후의 데이터셋이 버전으로 보존되어 있어야 한다. 버전관리 없이 전처리 파이프라인을 덮어쓰면 변경 전 상태로 되돌릴 수 없어 영향 분석이 불가능하다.

▶ 고위험 AI 서비스의 데이터 거버넌스 감사 대응 불가 방지: 공공 AI·금융 AI·의료 AI 등 규제 대상 AI 서비스는 'AI 모델이 어떤 데이터로 학습되었는가'에 대한 규제 기관·감사 기관의 소명 요구에 응할 수 있어야 한다. 데이터 버전관리 이력이 없으면 현재 운영 중인 AI 모델의 학습 데이터를 특정할 수 없어 규제 대응이 불가능하다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 버전관리(DVC) 도입 요건 및 버전관리 대상 데이터 산출물 범위 정의	AI 학습에 관련된 데이터 산출물(원본 데이터셋·전처리 데이터셋·증강 데이터셋·학습/검증/테스트 분할셋·전처리 파이프라인 설정 파일)을 버전 관리하는 DVC 도입 요건과 버전관리 대상 범위가 {요구사항정의서}에 정의되었는지 확인
2	데이터셋 버전 명명 체계 및 버전 업데이트 트리거 기준 정의	데이터셋 버전 식별을 위한 명명 규칙(예: dataset-v1.0.0, 날짜 기반 태그)과 새로운 버전을 생성하는 트리거 조건(데이터 추가·삭제·전처리 방법 변경·라벨 수정 등)이 정의되었는지 확인
3	AI 모델 버전과 학습 데이터셋 버전 간 1:1 매핑 기록 요건 정의	AI 모델의 각 버전(v1, v2 등)이 어떤 데이터셋 버전·전처리 파이프라인 버전·하이퍼파라미터 조합으로 학습되었는지를 1:1로 기록·조회할 수 있는 매핑 요건이 {요구사항정의서}에 정의되었는지 확인
4	특정 모델 버전 재현을 위한 데이터셋 버전 복원 요건 및 복원 검증 계획 정의	특정 AI 모델 버전의 재현(동일 데이터·동일 파이프라인·동일 하이퍼파라미터로 재학습 시 동일 성능 재현)을 위해 해당 모델 버전에 매핑된 데이터셋 버전을 정확히 복원할 수 있는 요건과 복원 검증 계획이 정의되었는지 확인
5	데이터 버전 변경 이력 감사 추적(Audit Trail) 기록 및 접근 통제 요건 정의	누가·언제·어떤 이유로 학습 데이터셋을 변경·삭제·추가했는지를 감사 추적 로그로 기록하는 요건과 데이터 버전 저장소에 대한 접근 권한 통제(변경 권한·조회 권한 분리) 요건이 {요구사항정의서}에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	버전관리 대상 데이터 산출물 범위, 버전 명명 규칙·업데이트 트리거, 모델-데이터 버전 1:1 매핑 기록 요건, 특정 버전 재현을 위한 복원 요건·복원 검증 절차, 변경 이력 감사 추적 로그 요건, 접근 권한 통제 요건이 {요구사항정의서}에 모두 정의됨	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	DVC 도입 의향은 있으나 버전 명명 규칙·트리거 기준·모델-데이터 매핑 요건 미정의, 복원 검증 절차 누락, 감사 추적 로그 요건 미정의, 접근 통제 요건 누락 등 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	데이터 버전관리 요건이 {요구사항정의서}에 전혀 없거나 '데이터를 관리하겠다'는 선언만 있고 버전관리 체계·모델-데이터 매핑·복원 요건·감사 추적 전무	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'Git으로 코드 관리' ≠ '데이터 버전관리' - 반드시 구분하여 점검: 많은 AI 개발팀이 소스코드는 Git으로 관리하면서 수GB~수TB의 학습 데이터셋은 파일 시스템·NAS·클라우드 스토리지에 날짜 폴더 구조(예: /data/2024-11-01/) 방식으로 관리하는 경우가 많다. 이는 버전관리가 아니라 단순 파일 보관이므로, 감리원은 {요구사항정의서}에 데이터 전용 버전관리 요건이 별도로 정의되어 있는지 확인해야 한다.

2) 모델-데이터 버전 매핑 기록 없이는 AI 모델 성능 이슈 원인 추적 불가: AI 모델 v2.0이 v1.0보다 성능이 저하되었을 때 '코드가 바뀌었기 때문인가, 데이터가 바뀌었기 때문인가'를 파악하려면 모델 v2.0이 어떤 데이터셋 버전으로 학습되었는지 즉시 조회할 수 있어야 한다. 이 매핑 기록이 없으면 원인 분석에 수 주가 소요되거나 영구적으로 불가능해진다.

3) 데이터 복원 요건은 반드시 복원 검증 절차와 함께 정의되어야 유효: '이전 버전 데이터셋을 복원할 수 있다'는 요건만으로는 충분하지 않다. 복원 후 해당 데이터셋으로 모델을 재학습했을 때 동일한 성능이 재현되는지 검증하는 절차(재현 오차 허용 범위, 검증 방법)가 함께 정의되어야 실질적인 재현성이 보장되므로, 감리원은 복원 요건과 복원 검증 절차가 함께 정의되었는지 점검해야 한다.

4) 감사 추적 로그는 데이터 오염(Data Poisoning) 탐지의 핵심 수단: AI 학습 데이터에 악의적으로 오염 데이터가 삽입되는 데이터 오염 공격이 발생했을 때 감사 추적 로그(누가 언제 어떤 데이터를 추가했는지)가 있어야 오염 데이터 삽입 시점과 경로를 추적할 수 있다. 감리원은 데이터 버전 변경 이력 감사 추적 로그 요건이 정의되어 있는지 확인해야 한다.

5) 최초 1회 학습 사업도 DVC 요건이 필요한 이유 - 운영 중 재학습 가능성: 현재는 최초 1회 학습만 계획된 AI 사업이라도 운영 중 성능 저하, 데이터 드리프트, 새로운 카테고리 추가 등 이유로 모델 재학습이 필요해지는 경우가 빈번하다. 재학습 시 이전 학습 데이터와 현재 데이터를 연계·비교하려면 이전 데이터 버전이 보존되어 있어야 하므로, 감리원은 이러한 사업에서도 데이터 버전 보존 요건이 정의되어 있는지 확인해야 한다.

AG_SD41-4-03 | 예외대응 요건

품질보증	요구정의·분석	시스템 구현	AI QA체계

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · OWASP LLM Top 10

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 {요구사항정의서}에 예상 사용자 오류 대응 요건과 프롬프트 오류 등 AI 특유의 예외 상황에 대한 대응 방안을 구체적이고 측정 가능하게 명시하였는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 일반적인 시스템의 입력 오류 처리 수준을 넘어, 생성형 AI 서비스에서 발생 가능한 탈옥(Jailbreak) 시도·금칙어 입력·환각(Hallucination) 등 AI 고유 예외 상황에 대한 대응 방안이 요구사항 단계부터 구체적으로 정의되었는지를 확인하며, 해당 요건이 향후 개발·검수 단계에서 검증 가능한 수준으로 기술되었는지를 검토한다.

2. 필요성

▶ 생성형 AI 시스템은 사용자의 입력(프롬프트)에 따라 응답이 실시간으로 생성되는 특성 때문에, 일반 정보시스템과 달리 사전에 정의되지 않은 무수한 입력 형태에 노출된다. 이로 인해 예외 대응 요건을 요구사항 정의 단계에서 명확히 규정하지 않을 경우, 개발 단계에서 임의로 처리되어 서비스 안전성·신뢰성·법적 책임 문제로 이어질 수 있다.

▶ 개발 후 분쟁 예방: 예외 대응 요건이 모호하게 기술되면 향후 검수 또는 하자보수 단계에서 책임 소재가 불분명해져 분쟁이 발생할 수 있다. 감리 단계에서 선제적으로 수치화·구체화 여부를 확인하는 것이 중요하다.

▶ AI 서비스 안전성 확보: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 사용자 입력에 대한 안전 장치와 이상 입력 처리 방안을 개발 초기 단계부터 정의하도록 권고한다.

▶ 비기능 요건 누락 방지: 예외 대응 항목은 기능 요건이 아닌 비기능(품질·안전) 요건으로 간주되어 {요구사항정의서}에서 누락되는 경우가 빈번하다. 감리 시점에서 이를 명시적으로 점검해야 한다.

▶ NIST AI RMF 정합성: NIST AI RMF MAP 1.6에서는 시스템 요구사항에 사회기술적 위험을 반영하도록 요구하고 있으며, 사용자 오류 및 예외 대응 방안은 이에 해당하는 핵심 요건이다.

3. 검토방법

No	점검 항목	확인 방법
1	예상 사용자 오류 유형 식별 여부	AI 서비스 특성상 발생 가능한 예상 사용자 오류 유형을 체계적으로 분류하여 {요구사항정의서}에 명시하였는지 확인
2	프롬프트 오류 및 비정상 입력 대응 요건 명시 여부	생성형 AI 시스템에 특화된 프롬프트 관련 예외 대응 방안이 요구사항에 구체적으로 기술되었는지 확인
3	AI 응답 오류 대응 요건 명시 여부	AI 시스템이 오답·환각(Hallucination) 등 잘못된 응답을 생성할 경우 이를 사용자에게 고지하고 처리하는 방안이 명시되었는지 확인
4	오류 대응 요건의 측정 가능성	예외 대응 요건이 개발·검수 단계에서 검증 가능하도록 수치화·구체화되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	예상 오류 유형이 체계적으로 분류되어 {요구사항정의서}에 명시되고, 프롬프트 오류·환각 등 AI 특유 예외 처리 방안이 구체적으로 기술되었으며, 검수 가능한 수치화 기준이 포함된 경우	현행 유지 (요건 명시 사실과 근거 산출물의 페이지·항목을 기재하고 우수 사례로 기록)
보완권고 (△)	오류 유형이 일부 누락되었거나 일반 IT 수준의 예외 처리 요건만 기술된 경우, AI 특유 예외(탈옥, 환각 등)에 대한 대응 방안이 부재하거나 측정 불가능한 모호한 표현으로만 기술된 경우	누락된 항목을 구체적으로 열거하고, 보완 방향 및 조치 기한을 권고사항으로 기재 후 재확인
부적합 (X)	{요구사항정의서}에 예외 대응 요건이 전혀 기술되어 있지 않거나, RFP 요건 대비 중요 예외 처리 항목이 반영되지 않은 경우	문제점을 명확히 기술하고 시정 조치(즉시 보완)를 요구하는 감리 지적 사항으로 등록 후 이행 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 요건의 수치화 요구: "AI는 부적절한 응답을 하지 않아야 한다"와 같은 선언적 표현은 분쟁의 원인이 된다. "금칙어 목록 500개 이상 반영", "탈옥 시도 탐지율 95% 이상"과 같이 측정 가능한 기준으로 구체화하도록 유도해야 한다.

2) AI 특유 예외 항목의 선제적 체크리스트 활용: 일반 IT 시스템의 예외 처리 요건 점검 방식(입력 유효성 오류, Null 처리 등)으로는 AI 시스템의 탈옥·프롬프트 인젝션·환각 등 AI 고유 예외를

발견하기 어렵다. 생성형 AI 전용 상세 검토 항목을 별도로 준비하여 체크리스트 방식으로 활용할 것을 권장한다.

3) 비기능 요건 섹션을 중점 검토: 예외 대응 요건은 {요구사항정의서}의 기능 요건 항목에는 잘 포함되지 않는다. 비기능 요건(품질, 안전, 보안) 섹션을 중심으로 검토하되, 관련 내용이 전체 문서에 분산 기술되어 있을 수 있으므로 키워드(예외, 오류, 탈옥, 안전 등) 기반으로 전문 검색도 병행한다.

4) RFP 대비 요건 후퇴 여부 반드시 확인: 제안요청서(RFP)에 예외 대응 관련 요건이 명시되어 있었으나 {요구사항정의서}에서 약화·축소된 경우, 이는 단순 미흡을 넘어 계약 위반으로 이어질 수 있다. RFP와 {요구사항정의서}를 반드시 대조하여 요건 후퇴 여부를 확인해야 한다.

AG_SD41-4-04 | AI산출물 품질보증 계획

품질보증 | 요구정의·분석 | 생명주기 관리 | AI QA체계

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF GOVERN · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템 개발 과정에서 산출하는 AI 특유의 산출물(AI 모델 설계서, 학습 데이터 전처리 명세서, 모델 성능 평가 결과서, 프롬프트 설계서 등)을 품질보증(QA) 대상에 명시적으로 포함하였는지 여부를 제3자 관점에서 점검하는 것을 목적으로 한다.

특히, {사업수행계획서} 내 품질보증계획 작성 시 AI 산출물이 일반 IT 산출물과 구분하여 QA 대상 목록에 구체적으로 반영되었는지, 그리고 AI 특성에 맞는 QA 방법 및 합격 기준이 정량적으로 명시되었는지를 확인한다.

AI 산출물이 QA 대상에서 누락될 경우 품질 미검증 상태에서 개발이 진행되어 사업 완료 후 납품 지연 및 품질 분쟁이 발생할 위험이 높다.

2. 필요성

▶ AI 서비스 구축 사업은 일반 정보시스템 개발과 달리, AI 모델 학습·검증 결과, 데이터 전처리 명세, 프롬프트 설계 등 AI 특유의 산출물이 추가로 발생한다. 이 산출물들은 기존 품질보증 프레임워크(일반 소프트웨어 산출물 중심)에 포함되지 않는 경우가 많아 감리 시점에 명시적 점검이 필요하다.

▶ AI 산출물의 비가시성: 모델 가중치, 학습 로그, 프롬프트 설계서 등은 전통적 QA 체계에서 공식 산출물로 분류되지 않는 경우가 많아, 명시적으로 목록에 포함하지 않으면 품질 관리의 사각지대가 발생한다.

▶ QA 방법 부재에 따른 품질 위험: AI 산출물에 대한 QA 방법이 정의되지 않으면, "AI는 적절한 성능을 제공해야 한다"와 같은 모호한 요건이 검사 단계에서 분쟁의 원인이 된다. 정량적 합격 기준 없이는 납품 판정 자체가 불가능해진다.

▶ 요구정의 단계 조기 개입 효과: 품질보증 계획 수립 시점인 요구정의 단계에서 AI 산출물 목록과 QA 방법을 점검함으로써, 설계·개발 단계의 품질 문제를 사전에 예방하고 시정 비용을 최소화할 수 있다.

▶ 국내외 기준 준수 의무: TTA 신뢰할 수 있는 인공지능 개발안내서 및 NIST AI RMF는 AI 시스템 생명주기 전반에 걸친 품질 관리 체계 수립을 권고하고 있으며, 공공 AI 사업에서는 초기 단계부터 이를 계획에 반영해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 산출물의 QA 대상 목록 열거 여부	AI 산출물이 {사업수행계획서}에 QA 대상 목록으로 명시적으로 열거되었는지 확인
2	AI 특유 산출물의 항목별 포함 여부	모델 설계서·학습 데이터 전처리 명세서·모델 성능 평가 결과서·프롬프트 설계서 등 AI 특유 산출물이 {품질보증계획서}에 항목별로 포함되었는지 {사업수행계획서}와 대조
3	AI 특성 반영 QA 방법 정의 여부	단순 문서 검토 외 기능 검사·성능 평가 방법 등 AI 특성에 맞는 QA 방법이 {품질보증계획서}에 정의되었는지 확인하고 사업자 면담으로 보완
4	산출물별 검토 시점·주기·담당자 명시 여부	AI 산출물별 검토 시점·주기·담당자(역할)가 {사업수행계획서}·WBS·{품질보증계획서}에 구체적으로 명시되었는지 확인
5	QA 계획의 AI 생명주기 반영 여부	QA 계획이 데이터 수집·전처리→모델 개발→성능 검증→시스템 통합의 AI 시스템 생명주기를 반영하는지 {사업수행계획서}·WBS에서 확인
6	QA 자동화 도구 활용 계획 명시 여부	AI 산출물 QA 체계의 자동화 도구 활용 계획이 {사업수행계획서} 및 'QA 도구 활용 계획'에 명시되었는지 확인(계획의 적정성만 점검)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 특유 산출물이 {사업수행계획서}·{품질보증계획서}의 QA 대상 목록에 명시적으로 열거되고, AI 특성에 맞는 QA 방법과 정량적 합격 기준, 산출물별 검토 시점·주기·담당자가 구체적으로 명시되었으며, QA 계획이 AI 시스템 생명주기(데이터 수집·전처리 → 모델 개발 → 성능 검증 → 시스템 통합)를 반영하고, QA 자동화 도구 활용 계획이 명시된 경우	현행 유지
보완권고 (△)	AI 산출물 중 일부만 QA 대상에 포함되었거나, QA 방법·합격 기준·검토 시점·담당자 등 핵심 요소 중 일부가 누락되거나 불명확하게 기재된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 산출물이 QA 대상 목록에서 전혀 식별되지 않았거나, "AI 품질을 보증한다"는 선언적 기술만 있고 어떤 산출물을 누가, 언제, 어떤 기준으로 검토하는지에 대한 실질적 계획이 전무한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) QA 기준의 수치화 유도: "AI는 정확한 답변을 제공해야 한다"와 같은 모호한 요건은 "모델 정확도 90% 이상", "금칙어 탐지율 95% 이상" 등 측정 가능한 기준으로 구체화하도록 사업자에게 유도한다. 감리 단계에서 이를 확인하지 않으면 검사 단계에서 분쟁의 씨앗이 된다.

2) AI 산출물과 일반 IT 산출물의 혼합 주의: 사업자가 {사업수행계획서}에 일반 IT 산출물 목록만 제시하고 AI 산출물을 별도 관리하지 않는 경우가 많다. 산출물 목록에서 AI 생명주기 단계(데이터·모델·검증)에 해당하는 항목이 있는지를 별도로 확인한다.

3) 선언적 기술과 실질적 계획의 구분: "AI 품질을 보증한다"는 선언적 기술은 적합으로 판정하지 않는다. 어떤 산출물을, 누가, 언제, 어떤 기준으로 검토하는지가 명시된 경우에만 적합으로 판정한다.

4) 생성형 AI 서비스의 추가 점검 포인트: 외부 LLM API를 활용하는 서비스의 경우, 외부 API 응답에 대한 품질 관리 계획(모니터링, 오류 처리, 폴백 방안)이 QA 계획에 포함되어 있는지를 반드시 확인한다. 외부 API 품질 변화에 대한 대응 방안이 없는 경우 부적합으로 판정할 수 있다.

5) 면담의 중요성: 문서에는 QA 계획이 형식적으로 기재되어 있더라도, 실제 QA 담당자가 AI 산출물의 특성을 이해하고 검토할 역량을 갖추고 있는지를 면담을 통해 확인한다. 역량이 부족한 경우 교육·전문가 투입 계획도 QA 계획에 포함되어야 한다.

AG_SD41-4-05 | 산출물 작성표준

품질보증 | 요구정의·분석 | 생명주기 관리 | 형상관리

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

요구정의(기획/분석) 단계에서 AI 서비스 개발에 필요한 산출물 유형과 작성 기준이 적절히 정의되었는지, 그리고 선택한 AI 개발 방법론이 해당 사업의 특성에 맞게 테일러링되어 {사업수행계획서}에 반영되었는지를 제3자 관점에서 점검하는 데 목적이 있다.

이 점검을 통해, AI 프로젝트 고유의 산출물(편향 분석, 모델 학습 계획, 윤리 가이드 등)이 누락될 경우 이후 설계·구현 단계에서 품질 리스크가 증폭되는 것을 조기에 방지하고, 방법론 테일러링의 실질성을 담보하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템 개발은 일반 소프트웨어 개발 방법론(SDLC)과 구조적으로 다른 특성을 가지며, 이는 방법론 테일러링과 산출물 정의에 직접적인 영향을 미친다. 감리원은 이러한 차이를 명확히 인식하고 점검에 임하여야 한다.

▶ 비결정론적 개발 사이클: 모델 성능이 데이터 품질 및 하이퍼파라미터에 따라 달라지며, 반복 학습(iteration)이 사전에 일정으로 반영되어야 한다.

▶ AI 특화 산출물 요구: 데이터 수집·라벨링 계획, 모델 학습 계획서, 편향성 분석 보고서, AI 윤리가이드 등 일반 IT 방법론에는 포함되지 않는 산출물이 필수적으로 요구된다.

▶ 테일러링 미흡 시 리스크: 방법론 테일러링 없이 일반 산출물 기준을 그대로 적용하면, AI 특화 위험 관리에 필수적인 산출물이 아예 누락되거나 형식적으로만 작성될 가능성이 높다.

▶ 요구사항 모호성 리스크: AI 비기능 요건(공평성, 설명가능성, 안전성 등)이 정량적 목표 없이 선언적으로만 기술될 경우, 검사 단계에서 수용 기준 판단이 불가능해진다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 개발 방법론 정의 적정성	{사업수행계획서}에 AI 특화 방법론(CRISP-DM 등) 또는 기존 방법론 기반 테일러링 계획이 명시되어 있는지 확인
2	테일러링 근거 및 기준 명확성	테일러링 기준(규모, AI 유형, 위험도 등)이 구체적으로 기술되고, 검토·승인 절차가 포함되어 있는지 확인

No	점검 항목	확인 방법
3	AI 특화 산출물 목록 반영	산출물 목록에 AI 고유 산출물(모델 학습 계획, 편향 분석, 데이터 라벨링 가이드, 윤리가이드 등)이 명시적으로 포함되어 있는지 확인
4	산출물 작성 기준·표준 정의	각 산출물에 대한 작성 양식·책임자·검토자·작성 시점이 {사업수행계획서}에 정의되어 있는지 확인
5	AI 요구사항 완전성 및 추적 가능성	기능/비기능 요건(편향, 설명가능성, 안전성 등)이 정량적으로 정의되고, RTM에 반영되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 개발 방법론 정의, 테일러링 근거·기준, AI 특화 산출물 목록, 산출물 작성 기준·표준, 요구사항 추적 가능성 등 모든 점검 항목을 충족하고 관련 산출물이 완비된 경우	현행 유지
보완권고 (△)	AI 개발 방법론과 AI 특화 산출물 목록은 정의되었으나, 테일러링 기준·검토 승인 절차가 불명확하거나, AI 고유 산출물 일부 누락, 산출물별 작성 양식·책임자·검토자·작성 시점 미정의, 요구사항의 정량 정의 또는 RTM 반영이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 개발 방법론 또는 테일러링 계획이 {사업수행계획서}에 정의되지 않았거나, AI 특화 산출물 목록과 작성 기준·표준이 부재하고 요구사항이 RTM에 반영되지 않아 추적이 불가능한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 테일러링 항목이 AI 개발 특성을 반영하여 실질적으로 조정되었는지 확인한다: 일반 SW 방법론을 그대로 적용하고 AI 특화 산출물(학습데이터 정의서, 모델 명세서, 편향성 분석 보고서 등)이 누락된 경우가 많다. 감리원은 테일러링 기준(규모·AI 유형·위험도)과 검토·승인 절차가 문서로 남아 있는지 확인한다.

2) 형식적 표준 준수와 내용 충실성을 구분하여 점검한다: 양식은 갖추었더라도 항목이 '해당 없음'으로 채워져 있으면 실질 이행으로 인정하지 않는다. 감리원은 표본 산출물을 지정하여 항목별 기재 내용의 구체성을 확인한다.

3) AI 개발 방법론 정의가 팀 전체에 공유되는지 확인한다: 방법론을 PM만 인지하고 개발팀이 모르면 산출물의 일관성이 무너진다. 감리원은 착수 교육 자료·공유 회의록 등 전파 증빙을 요구한다.

4) 반복 학습(iteration) 일정이 사업 일정에 반영되었는지 확인한다: AI 개발은 모델 성능이 데이터 품질과 하이퍼파라미터에 따라 달라지는 비결정론적 사이클을 가진다. 감리원은 학습·평가·재학습의 반복 횟수와 기간이 WBS에 반영되었는지, 성능 미달 시의 일정 여유가 확보되었는지 확인한다.

5) AI 비기능 요건은 정량 목표와 함께 추적표에 연결되어야 한다: 공평성·설명가능성·안전성이 선언적으로만 기술되면 수용 기준을 판단할 수 없다. 감리원은 각 비기능 요건의 측정 지표와 목표값이 정의되고 {요구사항추적표}에 연결되었는지 확인한다.

AG_SD44-4-01 | 편향-차별훈련 이행 점검

품질보증 | 시험 | 인공지능 모델 개발 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · NIST AI RMF MEASURE

1. 목적

AI 서비스 구축 사업에서 편향 발생 시 즉각 조치 훈련이 계획에 따라 실제로 이행되었는지, 그리고 차별 발언 발생 시 차단 및 모델 업데이트 훈련이 수행되고 그 결과가 {결과보고서}에 적절히 기록되었는지를 점검하는 것이 목적이다.

감리원은 AI 성능의 기술적 평가자가 아닌, 관련 산출물의 완결성과 이행 증거 확보 여부를 제3자 관점에서 검토한다.

2. 필요성

▶ 생성형 AI 서비스는 학습 데이터·모델 구조·사용자 입력의 복합적 특성으로 인해 특정 성별·지역·계층·집단에 불리한 편향된 출력 또는 차별적 발언을 생성할 수 있다. 시험/전개 단계에서 이에 대한 훈련 이행이 적정하게 이루어지지 않으면 서비스 오픈 이후 사회적·법적 분쟁이 발생할 수 있으므로, 다음의 이유로 감리원이 이를 점검해야 한다.

▶ 훈련 이행이 문서화되지 않으면, 검사 단계에서 수행 여부를 입증할 근거가 없어 사업 완료 시 분쟁 소지가 된다.

▶ 차별 발언 차단 기능은 단순히 구현 여부가 아니라 시험 케이스를 통한 검증 결과가 산출물로 남아야 한다.

▶ 모델 업데이트 이력이 관리되지 않으면, 배포 이후 동일한 편향·차별 문제가 재발하더라도 원인 추적이 불가능해진다.

▶ TTA 생성AI 개발안내서(T5-5, T5-6)는 시험 단계에서 편향·차별 대응 훈련 이행과 결과 기록을 명시적으로 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	'편향·차별 대응 훈련 계획서'와 {결과보고서}의 이행 결과 연계 여부	'훈련 계획서'와 {결과보고서}를 교차 대조하여 계획 대비 이행 현황 확인
2	편향 탐지 기준 수치와 시험 결과 수치 기록 완결성	{테스트결과서} 내 편향 관련 측정 항목의 기준값·결과값 기록 여부 확인

No	점검 항목	확인 방법
3	편향 조치 훈련 이후 before/after 비교 기록 여부	{결과보고서} 내 편향 수치 개선 결과 비교 항목 존재 여부 확인
4	차별 발언 테스트 케이스 포함 및 합격 기준 명시 여부	'시스템 시험 시나리오'에서 차별 발언 케이스 및 합격 기준 존재 여부 검토
5	{테스트결과서} 내 차별 발언 테스트 케이스별 수행 결과 기록 여부	{테스트결과서}에서 차별 발언 관련 테스트 결과 항목 직접 검토
6	시험 실패 케이스에 대한 모델 업데이트 이행 여부	'AI 모델 업데이트 이력서'에서 실패 케이스 대응 업데이트 기록 확인, 필요 시 사업자 면담
7	모델 업데이트 후 재시험 결과 기록 여부	{결과보고서}에서 업데이트 후 재시험 결과 항목 존재 여부 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	훈련 계획서와 {결과보고서}가 1:1로 연계되고, 편향 기준값·결과값, before/after 비교, 차별 발언 테스트 결과, 모델 업데이트 및 재시험 이력이 모두 산출물로 기록되어 있는 경우	현행 유지
보완권고 (△)	훈련은 수행되었으나 일부 항목의 수치 기록, 비교 결과 또는 업데이트·재시험 이력이 누락되어 이행 증거가 불완전한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향·차별 대응 훈련이 이행되지 않았거나, 이행을 입증할 산출물이 존재하지 않는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 편향 및 차별의 기준이 글로벌 일반 기준이 아닌, 해당 공공/민간 서비스가 제공되는 대한민국의 특정 사회적·문화적·법적 맥락(예: 장애인 차별 금지, 특정 지역 비하 금지 등)을 제대로 반영한 시나리오로 훈련되었는지 문서화된 시나리오 명세서를 통해 검토해야 한다. (TTA 공공사회 분야 및 일반 분야 요구사항 반영)

2) 수치 없는 서술은 적합으로 판정 불가: "편향 조치 훈련을 수행하였다"는 선언적 기술만으로는 이행을 인정하기 어렵다. {결과보고서}에 "편향 테스트 케이스 50건 수행, 편향 발생률 3.2% → 1.1%로 개선" 등 측정 가능한 결과가 기록되어야 적합으로 판정할 수 있다.

3) 훈련 계획서와 결과보고서의 교차 점검 필수: '편향·차별 대응 훈련 계획서'에 기술된 훈련 항목과 {결과보고서}의 이행 결과를 반드시 1:1로 대조한다. 계획 대비 누락 항목이 발견되면 사업자에게 소명을 요구한다.

4) 모델 업데이트 이력의 산출물화: 구두 설명("수정했다")만으로는 인정 불가하다. 'AI 모델 업데이트 이력서'에 업데이트 일시, 버전, 변경 내용, 담당자가 기록되어야 한다.

AG_SD44-4-02 | AI 신뢰성 테스트·자동화 평가 파이프라인 점검

품질보증 | 시험 | 인공지능 모델 개발 | AI QA체계

근거 | EU AI Act · NIST AI RMF MEASURE·MANAGE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 시험 단계에서, AI 시스템의 편향·강건성·적대적 입력 등에 대한 신뢰성 시험이 수행되고 그 결과가 자동화 파이프라인을 통해 지속적으로 재검증되는 체계가 구축되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 'AI 신뢰성 평가 계획서', {테스트결과서}, 'AI 신뢰성 평가 결과서' 및 CI/CD 파이프라인 설정을 근거로 시험 대상 영역의 완전성, Golden Set의 형상관리, 회귀테스트 자동 실행 이력, 배포 차단 게이트의 실효성을 확인한다.

감리원의 점검은 평가 알고리즘 자체의 학술적 타당성을 검증하는 것이 아니라, 신뢰성 검증 체계의 구축·운영 여부와 그 증빙의 완결성을 확인하는 데 한정한다.

2. 필요성

▶ 공공 AI 서비스의 편향·불공정한 결과와 적대적 입력 취약성은 국민 신뢰도와 기관의 법적 책임에 직결된다. 채용·복지·심사 등 고영향 업무에서 편향된 출력이 반복되면 차별 시비와 행정심판으로 이어질 수 있으므로, 배포 전 신뢰성 시험의 수행 사실이 문서로 확인되어야 한다.

▶ 1회성 평가는 모델·데이터가 변경되는 순간 효력을 잃는다: AI 시스템은 모델 버전 교체, 프롬프트 수정, 학습 데이터 추가만으로도 출력 특성이 달라진다. 배포 시점의 평가 보고서 한 건으로는 운영 중 품질을 보증할 수 없으므로 자동화된 회귀 검증 파이프라인이 필요하다.

▶ EU AI Act 제15조는 고위험 AI 시스템에 대해 정확성·강건성·사이버보안 수준을 생애주기 전반에 걸쳐 유지하고 지속 검증할 것을 요구한다. 지속 검증 체계의 부재는 규제 준수 실패로 직결되며, 국내 AI 기본법의 고영향 AI 조항 역시 이에 준하는 관리 의무를 규정한다.

▶ NIST AI RMF의 MEASURE 함수는 AI 시스템의 강건성과 보안·복원력을 배포 환경과 유사한 조건에서 반복 측정하고 문서화할 것을, MANAGE 함수는 배포 이후의 지속 모니터링과 피드백 반영을 요구한다. 자동화 파이프라인은 이 요구를 실무적으로 구현하는 수단이다.

▶ TTA 신뢰할 수 있는 인공지능 개발안내서는 신뢰성 확보를 위해 편향·강건성·적대적 입력·설명가능성·프라이버시의 5대 영역 평가를 권고한다. 일부 영역만 시험하고 나머지를 누락하면 신뢰성을 확보하였다고 주장할 근거가 성립하지 않으므로, 영역별 시험 케이스의 매핑 여부를 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 신뢰성 시험 대상 영역의 완전성	'AI 신뢰성 평가 계획서'와 {테스트결과서}에서 편향·강건성·적대적 입력·설명가능성·프라이버시의 5대 영역별 시험 케이스가 매핑표로 제시되고 미시험 영역이 없는지 확인
2	Golden Set 구성 및 형상관리	업무별 표준 테스트셋(Golden Set)의 저장소 위치·버전 태그·갱신 주기·승인 절차가 'AI 신뢰성 평가 계획서'에 정의되고 형상관리 도구에 등록되어 있는지 확인
3	회귀테스트 자동화 및 실행 이력	CI/CD 파이프라인 설정에서 모델·데이터 변경 시 Golden Set 기반 회귀테스트가 자동 실행되도록 구성되었는지 확인하고, 최근 3회 이상의 실행 이력(실행 일시·결과)을 CI 로그로 확인
4	LLM-as-Judge 정량 평가 체계	평가 프롬프트와 채점 기준이 문서화되었는지 확인하고, 평가자 모델이 피평가 모델과 동일 계열인 경우 사람 채점과의 일치도를 확인한 교차검증 근거가 {테스트결과서}에 제시되었는지 확인
5	변경 영향분석 및 배포 차단 게이트	모델·프롬프트·데이터 변경 시 다운스트림 영향분석 절차가 정의되었는지 확인하고, 신뢰성 시험 미달 시 배포가 실제로 중단되도록 CI/CD 설정에 게이트가 코드 수준으로 강제되는지 확인
6	평가 결과의 임계값 충족 및 미달 조치	'AI 신뢰성 평가 결과서'에 영역별 목표 임계값과 실측값이 기재되고, 미달 항목의 원인 분석·조치·재시험 결과가 문서화되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	신뢰성 5대 영역(편향·강건성·적대적 입력·설명가능성·프라이버시)의 시험 케이스가 매핑표로 제시되어 미시험 영역이 없고, Golden Set이 형상관리되며, CI/CD에서 회귀테스트가 자동 실행되어 최근 실행 이력이 확인되고, LLM-as-Judge 채점 기준과 교차검증 근거가 문서화되었으며, 시험 미달 시 배포 차단 게이트가 코드 수준으로 강제되고, 평가 결과의 임계값 충족과 미달 항목 조치가 문서화된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	신뢰성 시험은 수행되었으나 일부 영역이 누락되었거나, Golden Set 형상관리가 미비하거나, 회귀테스트 자동화가 일부 미구축되어 수동 평가에 의존하거나, 배포 차단 게이트가 문서상 절차로만 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향·강건성·적대적 입력에 대한 신뢰성 시험이 수행되지 않았거나, 1회성 평가에 그쳐 모델·데이터 변경 시 재검증 체계가 전혀 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 신뢰성 평가는 '배포 시점 통과'가 아닌 '지속 통과'가 핵심이다: 1회성 평가 보고서만 제출되고 자동화 파이프라인이 없으면, 모델·데이터가 변경되는 순간 평가 결과는 효력을 잃는다. 감리원은 최근 3회 이상의 회귀테스트 자동 실행 이력(CI 로그·실행 일시·결과)을 증빙으로 요구한다.

2) Golden Set은 '존재'가 아니라 '형상관리'가 요건이다: 표준 테스트셋 파일이 있어도 버전관리와 변경 이력이 없으면 회귀 비교 자체가 성립하지 않는다. 감리원은 Golden Set의 저장소 위치, 버전 태그, 갱신 주기와 승인 절차가 정의되어 있는지 확인한다.

3) LLM-as-Judge는 평가자 모델의 편향을 함께 점검한다: 평가 모델이 피평가 모델과 동일 계열이면 자기 편향으로 점수가 부풀려질 수 있다. 감리원은 평가 프롬프트와 채점 기준의 문서화 여부, 그리고 사람 채점과의 일치도를 확인한 교차검증 근거를 요구한다.

4) 배포 차단 게이트의 실효성을 확인한다: 변경 영향분석 문서만 있고 임계 미달 시 실제로 배포를 중단시키는 게이트가 없으면 통제로 인정하지 않는다. 감리원은 CI/CD 설정에서 시험 실패 시 배포가 코드 수준에서 강제 중단되는지 확인한다.

5) 신뢰성 5대 영역의 누락 여부를 확인한다: 편향·강건성·적대적 입력만 시험하고 설명가능성·프라이버시 영역이 빠지는 경우가 많다. 감리원은 TTA 개발안내서 기준의 영역별 시험 케이스 매핑표를 요구하여 미시험 영역을 식별하고 보완을 권고한다.

4.3. 응용시스템

AI 기능의 요구정의·설계·구현·시험 전반의 점검 가이드를 수록한다. (총 49건)

요구정의·분석 (15건)

AG_SD41-2-01 편향식별 기능 요건 점검

AG_SD41-2-02 AI모델명세 투명성 점검

AG_SD41-2-03 XAI 설명가능성 점검

AG_SD41-2-04 신뢰점수 노출 요건

AG_SD41-2-05 환각 점검 요건

AG_SD41-2-06 부적절 프롬프트 차단 요건

AG_SD41-2-07 편견 차별 단어 필터링 요건

AG_SD41-2-08 금융 컴플라이언스 5대 필수 설명 의무

AG_SD41-2-09 윤리적 우려사항 보고채널

AG_SD41-2-10 다국어 및 접근성 지원 요건

AG_SD41-2-11 예외처리 정책 요건

AG_SD41-2-12 접속 로깅 요건

AG_SD41-2-13 시스템 오작동 시 기본값 요건

AG_SD41-2-14 Fail-safe 전환 요건

AG_SD41-2-15 UX로그 수집 요건

설계 (16건)

AG_SD42-2-01 UI 색상배치 편향배제 다크패턴 설계

AG_SD42-2-02 편향모니터링 지표 산식설계

AG_SD42-2-03 편향유발 입력 사전차단 로직설계

AG_SD42-2-04 환각발언 예방 파이프라인설계

AG_SD42-2-05 AI결정 오버라이드 기능설계

AG_SD42-2-06 AI한계점 안내 프론트엔드 설계

AG_SD42-2-07 XAI 설명가능성 적용 설계

AG_SD42-2-08 신뢰점수 노출 설계

AG_SD42-2-09 AI결정 오버라이드 수동변경 기능점검

AG_SD42-2-10 개인화 데이터처리 알고리즘 설계

AG_SD42-2-11 안전모드 Fail-safe 로직 설계

AG_SD42-2-12 악성 프롬프트 필터링 로직 설계

AG_SD42-2-13 실시간 성능지표 수집로직 설계

AG_SD42-2-14 타시스템 API연계 프로토콜설계

AG_SD42-2-15 AI모델 아키텍처 점검

AG_SD42-2-16 Agentic안전정지Loop통제점검

구현 (6건)

AG_SD43-2-01 튜토리얼 및 사용자가이드 작동 점검

AG_SD43-2-02 마스킹 비식별화 뷰 정상 조회 점검

AG_SD43-2-03 프롬프트 엔지니어링 관리체계 점검

AG_SD43-2-04 캐싱 히트율 및 속도개선

AG_SD43-2-05 비정형 문서 처리 AI 요건 점검

AG_SD43-2-06 AI 민원·상담 서비스 특화 요건 점검

시험 (12건)

AG_SD44-2-01 편향 모니터링 지표 정상 표출점검

AG_SD44-2-02 XAI 설명가능성 기능 동작확인 점검

AG_SD44-2-03 환각(Hallucination) 억제테스트 통과점검

AG_SD44-2-04 비동기 추론 프로그레스 Bar 상태 동작 점검

AG_SD44-2-05 시스템 오작동 기본값 응답점검

AG_SD44-2-06 안전모드 Fail-safe 전환테스트 점검

AG_SD44-2-07 실시간 지연시간 테스트 점검

AG_SD44-2-08 프롬프트 인젝션 방어필터 동작점검

AG_SD44-2-09 데이터 파이프라인 재학습용 동작 점검

AG_SD44-2-10 실시간 성능지표 로깅 점검

AG_SD44-2-11 외부 연계시스템 통합시험 점검

AG_SD44-2-12 API 연계데이터 정합성 점검

AG_SD41-2-01 | 편향식별 기능 요건 점검

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「개인정보 보호법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템에 내재될 수 있는 편향(Bias)의 발생 가능성을 사전에 식별하고 정의하는 활동을 구체적으로 수행하였는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 일반적인 데이터 품질 관리 수준을 넘어, AI 서비스가 적용되는 도메인(행정, 의료, 복지, 교육 등)에 특화된 편향 요인을 사전 정의하고 이를 {요구사항정의서}에 반영하였는지를 중점적으로 확인한다.

또한 편향 식별 결과가 데이터 수집·가공 계획 및 모델 개발 요건으로 연결되는지 그 체계성을 검토한다.

2. 필요성

▶ AI 시스템은 학습 데이터와 알고리즘의 특성에 따라 특정 집단에 대한 차별적 결과를 산출할 위험이 있으며, 이러한 편향은 일반 정보시스템과 달리 시스템 내부에 내재되어 외부에서 직접 확인하기 어렵다는 특성이 있다. 이에 따라 개발 후 수정보다 사전 식별·예방이 필수적이다.

▶ 공정성 확보 의무: TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI 분야) 요구사항 07은 데이터 수집·가공 시 편향 완화 방안 수립을 명시하고 있으며, 요구사항 09는 모델 편향 제거 기법 적용을 요구한다. 이를 위한 출발점은 편향 요인 사전 정의이다.

▶ 도메인별 위험의 상이성: 의료 AI와 행정 민원 AI는 편향이 미치는 사회적 파급 효과가 전혀 다르다. 도메인에 무관하게 일반적인 편향 관리 계획만 수립하면 실질적인 위험을 놓치게 된다.

▶ 법적·윤리적 리스크 최소화: 「인공지능기본법」, 「개인정보 보호법」 등은 AI 시스템이 특정 집단을 차별하거나 불이익을 주지 않도록 요구하며, 기획 단계의 편향 식별 활동은 이를 충족하는 핵심 근거가 된다.

▶ 감리 점검의 실효성: 편향 발생 가능성이 요구사항으로 명시되지 않으면, 이후 설계·개발·테스트 단계에서 편향 관련 검증 기준 자체가 없어지게 되어 전체 감리 실효성이 저하된다.

3. 검토방법

No	점검 항목	확인 방법
1	편향 발생 가능성 식별의 요구사항 명시 여부	단순 데이터 품질 언급이 아닌 편향 위험 식별 활동 자체가 요구사항으로 기술되었는지 확인
2	편향 유형별 구분 식별 여부	데이터·알고리즘·라벨링 등 편향 유형이 구분되어 식별되었는지, '편향을 방지한다'는 통합 서술만 있는지 확인
3	편향 유입 데이터 수집 과정 명시 여부	어떤 데이터 수집 과정에서 편향이 유입될 수 있는지 구체적 시나리오가 기술되었는지 확인
4	편향 식별 결과 문서화 요구 여부	편향 식별 결과가 산출물로 계획되어 단순 회의 논의로 끝나지 않는지 요구사항에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	편향 위험 식별 활동이 별도 요구사항으로 명시되고, 편향 유형(데이터·알고리즘·라벨링 등) 구분, 수집 단계별 유입 시나리오, 식별 결과 문서화 계획이 {요구사항정의서}에 구체적으로 반영된 경우	현행 유지
보완권고 (△)	편향 식별 관련 요구사항이 일부 존재하나 유형 구분 없이 "편향을 방지한다" 수준의 선언적 서술에 그치거나, 식별 결과 산출물 계획·측정 기준이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향 발생 가능성 식별 활동이 요구사항으로 전혀 정의되지 않았거나, 도메인 특화 편향 요인 검토와 식별 결과 문서화 계획이 모두 부재한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 편향 측정 기준이 "납득할 수 있는 수준"처럼 모호하게 정의된 경우 구체적 수치 기준 수립을 요구하라.

2) 편향 발생 가능성 식별 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) RFP와 요구사항정의서 교차 검토: RFP AI 윤리 요건이 기술 요구사항으로 전환되지 않는 경우가 빈번하다. 선언적 조항과 측정 기준의 불일치를 반드시 확인한다.

4) RFP에 편향 발생 가능성 식별 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

5) 사업자 면담 시 편향 식별 도구·방법의 직접 시연을 요청하여 실제 수행 여부를 확인한다.

AG_SD41-2-02 | AI모델명세 투명성 점검

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조 · NIST AI RMF GOVERN·MAP · 「인공지능기본법 이행 가이드라인」(과기정통부, 2026.1.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 모델의 작동 방식·성능 특성·한계점을 요구사항 수준에서 명확히 정의하였는지, 그리고 사용자 화면 차원에서 AI 시스템임을 안내하고 모델의 불확실성 및 오류 가능성을 사용자에게 고지하는 화면 요건이 요구사항정의서에 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

응용시스템 감리 관점에서 기술적 구현 방식이 아닌 요구정의의 충실성과 명세의 구체성을 중심으로 점검하며, TTA 2024 신뢰할 수 있는 인공지능 개발안내서(생성 AI 기반 서비스 분야) 요구사항 11-1(모델 명세 투명성 제공) 및 요구사항 11-3(사용자 판단 지원 설명 제공)을 기준으로 요구사항정의서의 적정성을 확인한다.

2. 필요성

▶ AI 시스템은 일반 정보시스템과 달리 동일한 입력에도 매번 다른 결과를 생성할 수 있고, 환각(Hallucination) 현상으로 인해 사실과 다른 정보를 그럴듯하게 출력하는 특성이 있다. 이러한 특성을 사용자가 인지하지 못한 채 AI 결과를 무비판적으로 수용할 경우 의사결정 오류와 피해가 발생할 수 있으므로, 요구정의 단계에서 모델 명세와 사용자 고지 요건을 명확히 정의해야 한다.

▶ AI 결과의 신뢰성·책임성 확보: AI 모델이 어떤 데이터로 학습되었고, 어떤 목적에 최적화되어 있으며, 어느 수준의 오류 가능성이 있는지를 명세화하지 않으면 사업 완료 후 결과 오류 발생 시 책임 소재가 불명확해지며, 발주기관의 서비스 품질 관리에 공백이 발생한다.

▶ 사용자 보호 및 알 권리 보장: TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI 분야)는 사용자가 AI 시스템임을 인지하고 결과의 불확실성을 이해할 수 있도록 안내하는 화면 요건을 요구사항으로 명시할 것을 권고한다. 이를 요구정의 단계에서 정의하지 않으면 개발 완료 후 화면 추가가 어렵다.

▶ 법적 리스크 최소화: AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법) 및 개인정보 보호법 관련 고지 의무가 강화되는 추세에서, AI 서비스임을 명시하지 않거나 오류 가능성을 고지하지 않을 경우 관련 법적 리스크가 발생할 수 있다.

▶ 국내외 기준 준수: NIST AI RMF의 GOVERN·MAP 기능, TTA 요구사항 11 등 국내외 기준은 AI 모델 명세의 투명한 제공과 사용자 안내를 공통으로 요구하고 있으며, 공공 AI 사업은 이를 준수할 의무가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 유형·목적 정의 여부	생성형·분류형 등 사용될 AI 모델의 유형과 목적이 {요구사항정의서}에 명시적으로 정의되었는지 확인
2	학습 데이터 개요 정보 명세 요구 여부	출처·규모·수집 기간 등 학습 데이터 개요 정보를 명세화하는 요구사항이 {요구사항정의서} 및 '제안요청서'에 포함되었는지 대조
3	성능 기준·허용 오류 범위 정의 여부	정확도·응답 품질 등 성능 기준과 허용 오류 범위가 측정 가능한 형태로 {요구사항정의서}에 정의되었는지 확인하고 사업자 면담으로 보완
4	모델 적용 범위·한계점 명시 여부	특정 언어·도메인·입력 조건 등에서의 제약 등 모델 적용 범위와 한계점이 {요구사항정의서} 및 'AI 모델 명세서'에 명시되었는지 확인
5	AI 모델 명세서(모델카드) 작성 요구 여부	개발 완료 후 'AI 모델 명세서' 또는 '모델카드(Model Card)' 산출물 작성이 요구사항에 명시되었는지 사업수행계획서 WBS·{요구사항정의서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 모델의 유형·목적, 학습 데이터 개요, 정량적 성능 기준과 허용 오류 범위, 적용 한계점이 {요구사항정의서}에 명시되고, 'AI 모델 명세서(모델카드)' 산출물 작성 계획이 WBS에 반영된 경우	현행 유지
보완권고 (△)	모델 유형·목적 등 일부 항목은 정의되어 있으나 성능 기준이 선언적 수준에 그치거나, 한계점 기술·모델 명세 산출물 계획 등 일부 요건이 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 모델 명세 관련 요구사항이 {요구사항정의서}에 전혀 반영되지 않았거나, 모델 유형·목적, 정량적 성능 기준, 적용 한계 명세가 모두 부재한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항정의서 구체성 기준 적용: 모델 명세 투명성 요건은 대상 이용자 수준(전문가/일반인)별 명세 항목·깊이를 명시해야 구현 지침이 된다.

2) 모델 명세 투명성 제공 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) AI 모델 명세 불일치 발견 시 개발사에 서면 해명을 요청하고, 해명 결과를 감리 근거 자료로 기록한다.

4) RFP에 모델 명세 투명성 제공 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

AG_SD41-2-03 | XAI 설명가능성 점검

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

본 점검 가이드라인은 AI 서비스 구축사업에서, 감리원이 사업자가 제출한 {요구사항정의서}를 검토하여 AI 시스템의 설명가능성(XAI) 기술 적용 요건과 추론 근거(수치·시각화) 제공 방안이 구체적이고 상세하게 정의되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

단순히 '설명 가능하게 구현한다'는 선언적 문구에 그치지 않고, 어떤 기능·화면에서, 누구를 대상으로, 어떤 방식(수치적 기여도, 그래프·차트 시각화, 자연어 요약 등)으로 추론 근거를 제공할 것인지가 요구사항으로 명시되었는지를 확인한다.

이를 통해 설계·개발 단계에서 XAI 요건이 구체적으로 이행될 수 있는 기반이 마련되었는지를 감리한다.

2. 필요성

▶ AI 시스템은 복잡한 수학적 모델을 기반으로 결과를 도출하기 때문에, 사용자가 해당 결과가 왜 나왔는지를 이해하기 어렵다. 특히 공공서비스·의료·금융 등 고위험 도메인에서는 AI 결과에 대한 '설명 요구권'이 법적·윤리적 의무로 자리 잡고 있으며, 이를 충족하기 위해서는 요구정의 단계에서부터 XAI 요건을 구체적으로 정의해야 한다.

▶ 사용자 신뢰 확보: AI 결과만 제공하고 근거를 설명하지 못하면, 사용자는 AI 판단을 신뢰하지 못하거나 맹목적으로 수용하는 양극단의 오류를 범할 수 있다. 추론 근거를 제공해야 사용자가 AI 결과를 적절히 검토·활용할 수 있다.

▶ 법적·행정적 설명 의무 이행: 행정 처분, 의료 보조 등에 AI가 활용될 경우, 당사자는 결정 근거를 요구할 권리가 있으며, AI 시스템은 이를 제공할 수 있도록 설계되어야 한다. TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI 분야) 요구사항 03은 이를 필수 요건으로 명시하고 있다.

▶ 감리원·운영자의 AI 동작 검증 가능성 확보: XAI가 설계되지 않은 AI 시스템은 오류 발생 시 원인 파악이 불가능하며, 감리 및 사후 감사 시 AI 결과의 적정성을 검증할 수 없다. NIST AI RMF는 설명가능성과 해석가능성을 신뢰할 수 있는 AI의 핵심 특성으로 규정하고 있다.

▶ 요구정의 단계 명시 필요성: XAI 요건은 설계 이전 단계에서 정의되지 않으면 구현 단계에서 후속 적용이 어렵거나 비용이 과도하게 증가한다. 따라서 감리원은 이 요건의 포함 여부를 반드시 점검해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	XAI 적용 대상·설명 제공 대상 명시 여부	최종 사용자·운영자·감사자 등 XAI 적용 대상 기능·화면과 설명 제공 대상이 {요구사항정의서}에 명시되었는지 확인
2	추론 근거 제공 방식 구체화 여부	수치적 기여도·그래프·차트 시각화·자연어 요약 등 추론 근거 제공 방식이 {요구사항정의서}에 구체화되었는지 확인하고 사업자 면담
3	XAI 요건의 검증 가능 기준 기술 여부	설명 표출 항목·수준·범위 등 설계·시험 단계에서 검증 가능한 기준으로 XAI 요건이 기술되었는지 {요구사항정의서}를 검토하고 검사기준서(안)와 대조
4	RFP 설명가능성 요구사항 매핑 여부	RFP의 설명가능성 관련 요구사항이 {요구사항정의서}에 누락 없이 매핑되었는지 {과업대비표}와 대조

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	XAI 적용 대상 기능·화면, 설명 제공 대상, 추론 근거 제공 방식이 {요구사항정의서}에 구체적으로 명시되고, 설계·시험 단계에서 검증 가능한 수준으로 기술된 경우	현행 유지
보완권고 (△)	설명가능성 요건이 일부 반영되었으나 '설명 가능하게 구현한다' 수준의 선언적 기술에 그치거나, 대상·방식·검증 기준 중 일부가 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	설명가능성(XAI) 관련 요구사항이 {요구사항정의서}에 전혀 반영되지 않은 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항정의서 구체성 기준 적용: 설명가능성 요건은 대상 이용자 수준(전문가/일반인)별 설명 방식·깊이를 명시해야 구현 지침이 된다.

2) 설명가능성 기술 적용 요건 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) LLM 기반 서비스는 SHAP·LIME 등 전통적 XAI 기법의 직접 적용이 어려우므로, 출처 인용(RAG 근거 명시)·자연어 이유 제시 등 대체 방식이 요건으로 정의되었는지 확인한다.

4) RFP에 설명가능성 기술 적용 요건 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

AG_SD41-2-04 | 신뢰점수 노출 요건

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템이 도출하는 출력 결과(분석·분류·예측·생성 등)에 대한 신뢰점수(Confidence Score) 또는 확률·불확실성 정보를 사용자에게 제공하는 요건이 {요구사항정의서}에 구체적으로 정의되었는지를 감리원의 관점에서 점검하는 것이 목적이다.

AI 시스템의 출력은 항상 일정 수준의 불확실성을 내포하며, 사용자가 이를 인식하지 못한 채 AI 결과를 절대적으로 신뢰하면 오류 있는 의사결정이 발생할 수 있다.

본 가이드라인은 신뢰점수 노출 방식, 불확실성 안내, 임계값 기준 등의 요건이 개발 착수 전 요구사항 수준에서 명확히 계획되었는지를 제3자 관점에서 확인하는 것이 핵심이다.

2. 필요성

▶ AI 시스템은 동일한 입력에 대해서도 확률적으로 다른 결과를 생성할 수 있으며 결과의 정확성이 항상 보장되지 않으므로, 오류가 포함된 AI 결과가 그대로 의사결정에 반영되는 것을 방지하기 위해 출력 결과와 함께 해당 결과의 신뢰 수준을 사용자에게 제공해야 하며, 이 요건이 반드시 정의되어야 한다.

▶ 의사결정 품질 향상: 신뢰점수가 함께 제공되면 사용자는 낮은 신뢰도의 결과를 참고 수준으로만 활용하거나 추가 확인을 요청할 수 있다. 신뢰점수 없이 결과만 제공되면 사용자는 AI 출력의 불확실성을 인식하지 못한 채 의사결정을 내리게 된다.

▶ 고위험 도메인 안전성 확보: 의료 진단 보조, 금융 심사, 행정 처분 지원 등 결과 오류가 치명적인 분야에서는 신뢰점수와 불확실성 안내가 필수적이다. 요구정의 단계에서 명시하지 않으면 서비스 완성 후 수정이 매우 어렵고 비용이 크게 증가한다.

▶ AI 투명성 및 설명가능성 요건 이행: TTA 신뢰할 수 있는 인공지능 개발안내서 및 NIST AI RMF는 AI 시스템이 출력 결과의 불확실성과 한계를 사용자에게 전달하도록 명시적으로 요구한다. 요구사항에 이를 반영하지 않으면 감리상 지적 사항이 된다.

▶ 사후 수정 비용 절감: 신뢰점수 표시는 화면 설계·백엔드 연동·사용자 안내 정책과 연계되므로, 개발 완료 후 추가하면 비용이 크게 증가한다. 사전에 확정해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	신뢰점수 표시 요건 명시 여부	AI 출력 결과 화면에 신뢰점수(확률값·정확도 등급)를 표시하는 요건이 {요구사항정의서}에 독립 항목으로 명시되었는지 '화면설계서'와 함께 확인
2	신뢰점수 임계값·처리 방안 정의 여부	60% 미만 시 주의 안내 등 신뢰점수 임계값이 정의되고 임계값 이하 결과에 대한 처리 방안이 {요구사항정의서} 및 '사용자 시나리오'에 포함되었는지 확인
3	예측 불확실성 안내 요건 정의 여부	신뢰구간·오차 범위·모델 한계 고지 등 AI 모델의 예측 불확실성을 사용자에게 안내하는 요건이 {요구사항정의서}에 정의되었는지 확인하고 사업자 면담
4	신뢰점수 산출 근거 문서화 요구 여부	신뢰점수 산출 근거·불확실성 산정 방식의 문서화가 산출물 계획에 포함되었는지 {요구사항정의서}에서 확인하고 RFP의 신뢰점수 항목이 {요구사항추적표}·{과업대비표}로 추적 가능한지 대조

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	신뢰점수·불확실성 제공 요건이 {요구사항정의서}에 구체적으로 명시되고 관련 산출물 계획이 포함된 경우 (신뢰점수 표시 방식·임계값·조건이 독립 요구사항으로 명시, 불확실성 안내 요건 포함, {요구사항추적표}로 RFP 항목 추적 가능)	현행 유지
보완권고 (△)	관련 요구사항이 일부 반영되었으나 구체성 또는 완전성이 미흡한 경우 ("결과에 신뢰도를 표시한다"는 선언적 문구만 존재, 표시 방식·임계값 미정의, 불확실성 안내 계획 누락 등)	해당 항목 보완 요청 후 재확인
부적합 (X)	신뢰점수 및 불확실성 관련 요구사항이 {요구사항정의서}에 전혀 반영되지 않은 경우 (신뢰점수 노출 요건 전무, AI 결과의 불확실성 고지 계획 부재, AI 출력임을 사용자에게 알리는 요건 없음)	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항 구체성 기준 적용: 신뢰점수·불확실성 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

2) 고위험 도메인 사전 평가: AI 서비스 적용 도메인(의료·금융·행정 등 고위험 여부)을 예비조사 단계에서 파악하여 신뢰점수 노출 요건의 중요도를 사전 평가하고, 고위험 분야는 필수 요건으로 판단한다.

3) RFP와 요구사항정의서 교차 검토: RFP에 AI 출력 투명성·신뢰점수 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

4) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 문서상 기재가 있어도 담당자가 내용을 인지하지 못하는 경우가 빈번하므로 실질 적정성 검증이 중요하다.

5) 판정 경계 사례 판단 기준: 신뢰점수 노출 요건이 일부만 반영된 경우는 보완권고(△)로 처리하고 보완 방향을 구체적으로 제시하며, 관련 요건이 전혀 없는 경우에만 부적합(X)으로 판정하고 재점검 일정을 명시한다.

AG_SD41-2-05 | 환각 점검 요건

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | 할루시네이션

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템이 산출할 수 있는 환각(Hallucination) 현상—즉 사실에 부합하지 않거나 근거 없는 정보를 그럴듯하게 생성하는 오류—을 방지하기 위한 기능 요구사항이 {요구사항정의서}에 구체적이고 상세하게 정의되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 단순한 "정확도 향상" 또는 "오류 최소화"와 같은 선언적 문구만으로는 충분하지 않으며, 허위정보 필터링 방식, 출처 명시 요건, 불확실성 고지 기준, 사람 개입(Human-in-the-loop) 프로세스 등 환각 방지를 위한 구체적인 기능 및 절차 요구사항이 명시되었는지를 검토한다.

2. 필요성

▶ 생성형 AI 시스템은 일반 정보시스템과 달리 모델이 학습 데이터의 패턴을 기반으로 텍스트·이미지 등을 자율적으로 생성하기 때문에, 사실과 다른 내용(허위 정보)을 마치 올바른 정보인 것처럼 자신 있게 제시하는 환각 현상이 구조적으로 발생할 수 있다. 이 문제를 요구정의 단계부터 명확히 대응하지 않으면 개발·운영 단계에서 심각한 품질·신뢰 문제로 확대된다.

▶ 공공 AI 서비스의 신뢰성 확보: 공공기관이 제공하는 AI 서비스에서 환각으로 인한 오정보가 제공될 경우, 국민의 권리·의무에 직접 영향을 미칠 수 있어 기획 단계에서 방지 요구사항 정의가 필수적이다.

▶ TTA 생성 AI 개발안내서 준거 의무: TTA「신뢰할 수 있는 인공지능 개발안내서 - 생성 AI 기반 서비스 분야」는 AI 출력 신뢰성 확보를 위해 허위정보 필터링 기능 및 출처 명시 요건을 요구사항에 반영하도록 명시적으로 권고하고 있다.

▶ NIST AI RMF 준거: NIST AI RMF의 Valid & Reliable 특성 및 MAP 2.2는 AI 시스템의 지식 한계와 출력 불확실성을 문서화하고 요구사항에 반영할 것을 요구한다.

▶ 개발 완료 후 수정 비용 최소화: 환각 방지 기능을 요구정의 단계에서 명확히 정의하지 않으면, 개발 완료 후 기능 추가·수정에 따른 비용과 일정 지연이 불가피하게 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	허위정보 필터링 방식 명시 여부	검색증강생성(RAG) 연동·팩트체크 모듈·사전 구축 지식베이스 기반 검증 등 허위정보 필터링 방식이 기능 요구사항으로 {요구사항정의서}에 명시되었는지 확인
2	AI 생성 정보 출처·근거 명시 기능 요구 여부	AI가 생성한 정보의 출처 또는 근거 데이터를 사용자에게 명시하는 기능 요구사항이 {요구사항정의서}에 포함되었는지 확인
3	AI 응답 신뢰도·불확실성 고지 기능 요구 여부	'이 정보는 AI가 생성한 것으로 실제와 다를 수 있습니다' 등 AI 응답의 신뢰도·불확실성을 사용자에게 고지하는 기능 요구사항이 {요구사항정의서}에 포함되었는지 확인하고 사업자 면담
4	고위험 업무의 사람 검토(HITL) 프로세스 포함 여부	민원 처리·법률 해석·의료·복지 판단 보조 등 고위험 업무 영역에 대해 AI 생성 결과를 담당자가 검토하는 사람 검토(HITL) 프로세스가 {요구사항정의서} 및 '사업수행계획서'에 포함되었는지 확인하고 발주기관 면담
5	환각 방지 기능 성능 측정 기준 명시 여부	정확도 목표치·오류율 허용 범위·테스트 시나리오 등 환각 방지 기능의 성능 측정 기준이 품질 요구사항으로 {요구사항정의서}에 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	허위정보 필터링 방식, 출처 명시, 불확실성 고지, 고위험 업무의 사람 검토(HITL) 절차, 성능 측정 기준이 {요구사항정의서}에 구체적 기능·수치 요건으로 명시된 경우	현행 유지
보완권고 (△)	환각 방지 요건이 일부 반영되었으나 필터링 방식·측정 기준이 선언적 수준에 그치거나, 출처 명시·불확실성 고지·사람 검토 절차 중 일부가 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	환각 방지 관련 기능 요구사항이 {요구사항정의서}에 전혀 반영되지 않았거나, "정확도 향상" 수준의 추상적 문구만 존재하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 환각 점검 요건은 핵심 위험 요소(필터링 방식·출처 명시·고지 기준)를 중심으로 점검 깊이를 조정하고, 발견사항을 구체적 근거와 함께 기록하라.

2) 생성 AI의 환각 방지 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) 면담 시 고위험 업무 영역의 목록과 각 영역별 사람 검토(HITL) 적용 여부를 확인하여 요건 정의의 누락을 식별한다.

4) RFP에 생성 AI의 환각 방지 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

AG_SD41-2-06 | 부적절 프롬프트 차단 요건

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | 금칙어/가드레일

근거 | OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, 사용자의 입력(프롬프트)을 통해 AI 시스템의 동작을 의도적으로 조작하려는 공격(프롬프트 인젝션, 탈옥 등)을 차단하기 위한 방어 요건이 {요구사항정의서}에 구체적이고 측정 가능한 수준으로 반영되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

특히 '부적절 입력 차단'이라는 선언적 문구만 존재하고 차단 대상 유형, 처리 방식, 판정 기준이 명시되지 않은 경우 설계·개발·검사 단계에서 구현 기준 부재 및 책임 공백 리스크가 발생한다.

이를 기획·분석 단계에서 사전에 식별하고, 이후 단계 추적 가능한 점검 기준을 확보하는 것이 핵심 목적이다.

2. 필요성

▶ 생성형 AI 시스템은 사용자의 자연어 입력을 그대로 AI 모델에 전달하는 구조적 특성 때문에, 악의적으로 조작된 입력(프롬프트 인젝션·탈옥 등)에 취약하다. 이 리스크를 요구정의 단계에서 요건으로 명시하지 않으면 다음 문제가 발생한다.

▶ 공공 서비스 신뢰성 훼손: 공격자가 AI 시스템의 안전 장치를 우회하여 금지된 정보(개인정보, 유해 정보 등)를 추출하거나, 시스템을 의도치 않은 방식으로 동작시키면 기관 신뢰도와 법적 책임 문제가 발생한다.

▶ 요건 부재로 인한 검사 기준 무력화: '부적절 입력을 차단한다'는 선언적 문구는 이행 여부를 판정할 수 없다. 차단 대상 유형(인젝션 패턴, 금칙어, 탈옥 시도 등), 차단 방식, 탐지율 기준이 요구사항에 명시되어야 측정 가능한 검사 항목이 된다.

▶ 보안·응용시스템 요건의 경계 확보: 프롬프트 차단 로직은 순수 보안 영역이 아니라, AI 서비스 응용 레벨에서 요구사항으로 정의해야 하는 사항이다. 요구정의 단계에서 응용시스템 요건으로 명시되지 않으면 설계 단계에서 누락되거나, 보안 영역과 응용시스템 영역 간 책임 공백이 발생한다.

▶ 사업 완료 후 분쟁 예방: 인젝션·탈옥 공격 방지 로직이 요구사항에 없다면, 구축 후 취약점 발견 시 사업자는 '요구사항에 없었다'고 주장할 수 있다. 명문화해야 이를 예방할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	금칙어·부적절 입력 차단 요건 명시 여부	금칙어·유해 표현·개인정보 요청·불법 정보 요청 등 차단 대상 입력 유형과 금칙어 목록의 초기 구성 기준·관리 절차(추가·수정 권한, 갱신 주기 등)가 {요구사항정의서}에 명시되었는지 'RFP'·{과업대비표}와 대조
2	프롬프트 인젝션 방지 로직 요건 반영 여부	시스템 프롬프트 변조·노출 인젝션 시도를 탐지·차단하는 요건과 탐지 방식의 적용 범위(입력 전처리·출력 후처리·양방향)·탐지율·오탐률 기준이 {요구사항정의서} 비기능 요구사항에 반영되었는지 확인하고 발주기관 면담
3	탈옥(Jailbreak) 공격 방지 요건 명시 여부	안전 장치 우회 탈옥 시도 방어 요건과 탈옥 유형(역할 전환·가상 시나리오 우회·반복 시도 등)별 방어·탐지 시 대응 방식이 {요구사항정의서}에 명시되었는지 확인하고 발주기관·사업자 면담
4	차단 로직 관리·갱신 요건 반영 여부	신규 공격 패턴 대응 차단 로직 갱신 절차, 오탐 예외 처리·이의 제기 절차, 차단·탐지 이력 로깅·모니터링 요건이 {요구사항정의서} 관리·유지보수 항목에 반영되었는지 {과업대비표}와 대조

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	차단 대상 입력 유형(금칙어·인젝션·탈옥 등)과 처리 방식, 탐지율·오탐률 등 측정 가능한 기준, 차단 로직 갱신·오탐 처리 절차가 {요구사항정의서}에 구체적으로 명시된 경우	현행 유지
보완권고 (△)	차단 요건이 일부 반영되었으나 특정 공격 유형(인젝션 또는 탈옥)이 누락되었거나, 탐지 기준·갱신 절차·오탐 처리 요건이 선언적 수준에 그쳐 검사 단계 판정 기준으로 활용하기 어려운 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	부적절 프롬프트 차단 관련 요구사항이 {요구사항정의서}에 전혀 반영되지 않았거나, '부적절 입력을 차단한다'는 선언적 문구만 존재하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 필터링 로직이 키워드 블랙리스트 방식만으로 구성된 경우, 우회 가능성을 지적하고 의미 기반 필터 또는 시스템 프롬프트 강화를 권고한다.

2) AI 서비스에 대한 프롬프트 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) RFP와 요구사항정의서 교차 검토: RFP에 명시된 LLM 품질 기준이 요구사항에 구체적 지표로 전환됐는지 확인한다. 모델 버전 고정 여부도 함께 점검한다.

4) RFP에 AI 서비스에 대한 프롬프트 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

5) 활용할 프롬프트 인젝션 테스트 케이스 목록의 제출을 요구하여 요건의 검증 가능성을 확보한다.

AG_SD41-2-07 | 편견 차별 단어 필터링 요건

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | 금칙어/가드레일

근거 | 「장애인차별금지 및 권리구제 등에 관한 법률」·「신뢰할 수 있는 인공지능 개발안내서」(TTA)·「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, AI 서비스가 사용자 입력 또는 AI 자체 출력을 통해 성별·연령·장애·출신지·인종·종교 등 특정 집단에 대한 편견·차별적 표현이 유통되지 않도록 금칙어 사전 연계 및 편향 표현 차단 기능에 관한 요건이 {요구사항정의서}에 차단 범위·운영 절차·검증 방법을 포함하여 구체적으로 정의되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

AI 서비스의 편견·차별 표현 문제는 학습 데이터에 포함된 사회적 편견이 모델에 내재화되어 의도치 않게 편향된 답변이 생성되는 경우와, 악의적 사용자가 의도적으로 차별적 표현을 생성하도록 AI를 유도하는 경우의 두 가지 경로로 발생한다.

이 두 경로를 모두 제어하기 위한 금칙어 사전과 편향 탐지·차단 기능이 요구정의 단계에서 명확히 요건화되지 않으면 설계·개발 단계에서 구조적으로 반영되지 않아 차별 표현 유통으로 인한 사회적 책임 및 법적 리스크가 발생하므로, 이를 사전에 식별하고 요건 확보를 유도하는 것이 핵심 목적이다.

2. 필요성

▶ 공공 AI 서비스에서 편견·차별 표현의 유통은 단순한 품질 결함을 넘어 특정 집단의 존엄성과 평등권을 침해하는 인권 문제로 직결되며, 금칙어 사전 연계 및 차단 기능 요건이 요구정의 단계에서 정의되지 않으면 다음과 같은 리스크가 발생한다.

▶ 법적·제도적 의무 위반 리스크: 「장애인차별금지법」, 「양성평등기본법」, 「국가인권위원회법」 등 관련 법령은 공공기관이 운영하는 서비스에서 특정 집단에 대한 차별적 표현이 생성·유통되지 않도록 관리할 의무를 부과한다. AI 서비스에서 편견·차별 표현이 출력되면 발주기관은 법적 의무 위반과 국가인권위원회 진정 대상이 될 수 있다.

▶ 학습 데이터 편향의 구조적 증폭 리스크: 생성형 AI 모델은 학습 데이터에 내재된 사회적 편견(성별 고정관념, 지역 편견, 장애인 비하 표현 등)을 그대로 흡수하여 출력에 반영한다. 이는 의도적 생성이 아니더라도 차별 표현이 출력될 수 있음을 의미하며, 금칙어 차단 기능과 편향 탐지 요건이 없으면 모델 내재 편향이 사용자에게 그대로 전달된다.

▶ 사용자 악용에 의한 차별 표현 생성 가능성: 악의적 사용자가 AI에게 특정 집단을 비하하는 표현을 생성하도록 유도(프롬프트 조작)하는 시도가 발생할 수 있다. 금칙어 차단 기능과 입력 필터링 요건이 {요구사항정의서}에 명시되지 않으면 이에 대한 대응 로직이 설계에 반영되지 않는다.

▶ 차단 기준의 불투명성에 따른 분쟁 리스크: 금칙어 사전의 구성 기준, 차단 범위, 갱신 절차가 명시되지 않으면 무엇이 차단되고 무엇이 허용되는지 기준이 없어 검사 단계에서 합격·불합격 판정 근거가 부재하며, 서비스 개통 후 차별 표현 발생 시 책임 소재가 불분명해진다.

3. 검토방법

No	점검 항목	확인 방법
1	금칙어 사전 구성 기준 및 차단 범위 요건 명시 여부	{요구사항정의서}·'금칙어 사전(초안)'·'AI 윤리 준수 계획서'에서 금칙어 구성 기준·포함 범주(성·연령·장애·지역·인종·종교 등), 최소 규모·적용 방향(출력·입력·양방향), 허용 표현 예외 처리 요건이 기술되었는지 확인
2	편향 탐지·차단 기능 및 AI 출력 후처리 요건 명시 여부	{요구사항정의서}·'AI 윤리 준수 계획서'에서 문맥 기반 편향 탐지 기능·적용 범위, 출력물 차별 표현 처리 방식(차단·대체·재생성)과 대체 문구, 탐지 임계값(탐지율·오탐률) 요건이 포함되었는지 확인
3	금칙어 사전 갱신 절차 및 관리 요건 명시 여부	{요구사항정의서}·'금칙어 사전(초안)'에서 갱신 주기·담당 주체·긴급 갱신 절차, 추가·삭제 사유 기록·이력 조회, 오탐 시 이의 제기·즉시 수정·모니터링 요건이 반영되었는지 확인하고 갱신 담당 주체를 면담으로 확인
4	편향 차단 요건의 검사 기준 및 차단 현황 모니터링 요건 명시 여부	{요구사항정의서}·{과업대비표}에서 범주별 대표 금칙어·우회 시도·오탐 테스트 검사 기준과 운영 단계 차단 현황 모니터링(차단 건수 리포트·급증 알림·이력 보관) 요건이 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	점검 항목 모두가 {요구사항정의서}에 서비스 특성을 반영하여 차단 범위·방식·갱신 절차·검증 방법이 구체적으로 기술된 경우 (금칙어 범주·방향·최소 규모·예외 처리 기준, 문맥 기반 탐지 및 후처리 방식·탐지 임계값, 갱신 주기·담당 주체·이력 관리·오탐 처리 절차, 검사 시나리오·모니터링 기능 요건이 독립 항목으로 반영)	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	일부 항목이 반영되었으나 구체성·차단 방향 명시·갱신 절차 완전성이 미흡한 경우 (금칙어 차단 요건의 선언적 기술, 차단 방향(입력·출력) 미명시, 오탐률 기준 부재, 갱신 담당 주체 불명확, 긴급 갱신 절차 누락, 검사 시나리오에 우회 시도 미포함, 모니터링 기능 요건 누락 중 하나 이상)	해당 항목 보완 요청 후 재확인
부적합 (X)	편견·차별 표현 차단에 관한 요건이 {요구사항정의서}에 전혀 반영되지 않아, AI 서비스가 차별 표현을 출력하거나 차별 표현 입력을 수용해도 이를 요건 위반으로 판정할 기준이 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항 구체성 기준 적용: 편견·차별 표현 차단 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

2) 탐지율·오탐률 수치 기준 확인: {요구사항정의서}에 탐지율(미탐 최소화 기준)과 오탐률 허용 기준이 수치로 기술되어 있는지 확인한다. "탐지율 95% 이상, 오탐률 5% 이하"와 같은 수치 기준이 있어야 필터링 기능의 적정성을 판정할 수 있다.

3) RFP와 요구사항정의서 교차 검토: RFP에 편견·차별 표현 차단 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

4) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 특히 금칙어 사전 갱신 담당 주체와 긴급 갱신 절차를 발주기관 면담으로 확인한다.

5) 판정 경계 사례 판단 기준: 편견·차별 표현 차단 요건이 일부만 반영된 경우는 보완권고(△)로 처리하고 보완 방향을 구체적으로 제시하며, 관련 요건이 전혀 없는 경우에만 부적합(X)으로 판정하고 재점검 일정을 명시한다.

AG_SD41-2-08 | 금융 컴플라이언스 5대 필수 설명 의무

응용시스템 | 요구정의·분석 | 인공지능 모델 개발 | 컴플라이언스/가드레일

근거 | 「신용정보의 이용 및 보호에 관한 법률」·「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」제31조·제34조

1. 목적

본 점검 가이드라인은 금융권 Agentic AI 상담·권유 서비스의 요구정의 단계에서, 자본시장법·금융소비자보호법상 설명의무와 부당권유 금지 등 컴플라이언스 통제가 AI 산출물에 강제 적용되도록 요건이 정의되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 {요구사항정의서}와 'AI 상담 컴플라이언스 요건 정의서'를 근거로 ① 5대 필수 설명의무(원금손실·수수료·중도환매·예금자보호·위험등급)의 정량 검증 요건, ② 고객 투자성향과 상품 위험등급의 적합성 판정 요건, ③ 부당권유 금지표현의 차단 요건, ④ 이행·차단 결과의 증적 기록 요건이 명시되었는지를 확인한다.

감리원의 점검은 개별 상품의 법적 해석이 아니라 통제 요건의 정의 여부와 검증 가능성에 한정한다.

2. 필요성

▶ 법적 의무의 시스템적 이행: 설명의무 누락과 부당권유는 금융소비자보호법 위반으로 직접 제재 대상이며, 상담 주체가 사람인지 AI인지를 불문하고 동일하게 적용된다. AI가 생성한 산출물에도 동일한 통제가 강제되지 않으면 기관이 법적 책임을 부담한다.

▶ LLM 자율 판단 배제: 필수 설명의무의 포함 여부를 LLM의 자율 생성에 맡기면 확률적 특성상 누락이 발생한다. 산출물에 5대 항목이 실제로 포함되었는지를 규칙 기반으로 정량 검증하고, 미포함 시 발송을 차단하는 요건이 정의되어야 한다.

▶ 적합성 원칙의 구현: 고객 투자성향과 상품 위험등급의 적합성은 LLM 추론이 아니라 Rule 기반 판정으로 결정되어야 하며, 부적합 권유는 생성 단계에서 차단되어야 한다. 자본시장법상 적합성 원칙은 사후 점검이 아닌 사전 통제를 요구한다.

▶ 부당권유 금지표현의 사전 차단: '원금 보장', '확정 수익' 등 금지표현은 생성 후 사람이 검수하는 방식으로는 실시간 상담에서 통제할 수 없다. 금칙어 사전과 출력 필터링 요건이 정의되어야 한다.

▶ 증적과 책임 추적: 설명의무 이행 여부와 차단 결과가 Case ID·상담 일시·고객 식별자와 함께 기록되어야 분쟁·감사 시 증적으로 활용할 수 있다. 기록 요건이 요구정의 단계에서 정의되지 않으면 사후 보강이 사실상 불가능하다.

3. 검토방법

No	점검 항목	확인 방법
1	5대 필수 설명의무 항목의 정량 검증 요건	{요구사항정의서}에서 원금손실·수수료·중도환매·예금자보호·위험등급 5개 항목의 포함 여부를 규칙 기반으로 검증하는 요건이 정의되고, 검증 실패 시 처리(차단·재생성) 방식이 명시되었는지 확인
2	원금손실 가능성 안내 요건	상품 유형별 원금손실 가능성 문구의 필수 포함 여부와 표현 기준이 'AI 상담 컴플라이언스 요건 정의서'에 정의되었는지 확인
3	수수료 구조 안내 요건	판매보수·운용보수 등 수수료 항목의 안내 범위와 표기 방식이 {요구사항정의서}에 정의되었는지 확인
4	중도환매 조건·불이익 안내 요건	중도환매 가능 여부, 환매수수료, 환매 제한 기간 등의 안내 요건이 {요구사항정의서}에 정의되었는지 확인
5	예금자보호 대상 및 위험등급 고지 요건	예금자보호법 적용 대상 여부와 상품 위험등급의 고지 요건이 상품 유형별로 구분되어 {요구사항정의서}에 정의되었는지 확인
6	투자성향-위험등급 적합성 판정 및 부적합 권유 차단	고객 투자성향 등급과 상품 위험등급의 적합성 판정이 Rule 기반으로 수행되고 부적합 권유가 생성 단계에서 차단되도록 {요구사항정의서}에 요건이 정의되었는지 확인. LLM 자율 판단에 의존하는 경우 부적합으로 판정
7	부당권유 금지표현 차단 및 증적 기록 요건	'원금 보장'·'확정 수익' 등 금지표현 사전과 출력 필터링 요건이 정의되었는지 확인하고, 설명의무 이행·차단 결과가 Case ID와 함께 기록되도록 요건이 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	5대 설명의무의 정량 검증 요건과 검증 실패 시 차단 방식, 투자성향-위험등급 적합성의 Rule 기반 판정, 부당권유 금지표현 차단, Case ID 기반 증적 기록 요건이 {요구사항정의서}에 구체적으로 정의된 경우	현행 유지
보완권고 (△)	5대 설명의무 항목은 정의되었으나 정량 검증 방식, 검증 실패 시 차단 동작, 적합성 판정의 Rule 기반 구현 또는 증적 기록 요건 중 일부가 선언적 수준에 그친 경우	요건 구체화 요청 후 재점검
부적합 (X)	설명의무 포함 여부를 LLM 자율 생성에 의존하여 정량 검증·차단 요건이 부재하거나, 적합성 판정과 금지표현 차단 요건이 전혀 정의되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 기계적 정량 검증: 설명의무는 '대체로 안내한다'가 아니라 '5개 항목이 모두 들어갔는가'로 기계적으로 검증되어야 한다. 누락은 곧 위법이다.

2) 프롬프트 지시만으로는 불충분: 설명의무 포함을 프롬프트 지시로만 처리하지 말고, 출력 직전 Regex+SLM으로 5대 항목 포함 여부를 정량 검증하며 누락 시 BLOCK/Rewrite가 동작하는지 확인한다.

3) 금지표현 자동 교정: '원금 보장'·'확정 수익' 등 단정적·오인 표현은 표준 대체문구(allowedRewrite)로 자동 교정되는지 점검한다.

4) 적합성 Rule 판정 신뢰: 적합성은 checkSuitability 룰 판정 결과만 신뢰하고, 투자성향 만료 상태에서 상품 추천이 지속되지 않고 즉시 프로세스가 중단되는지 확인한다.

5) 증적·승인 이력 확인: 준법감시부의 설명의무·금지표현 사전 승인 이력과 차단·재작성 결과의 Case ID 기반 기록, 분쟁 대비 증적 보존 정책을 확인한다.

AG_SD41-2-09 | 윤리적 우려사항 보고채널

응용시스템	요구정의·분석	생명주기 관리	법령/AI윤리

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보화 기본법」 · NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 시스템 운영 중 발생할 수 있는 윤리적 우려사항(차별·편견·불공정 출력 등)을 이용자가 직접 신고·보고할 수 있는 채널에 관한 요구사항이 {요구사항정의서}에 구체적이고 측정 가능한 수준으로 정의되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

단순한 기능 열거 또는 "신고 기능을 제공한다"와 같은 추상적 문구에 그치지 않고, 신고 채널의 운영 주체·처리 절차·응답 기한·결과 피드백 방안 등 실질적 이행 가능성을 판단할 수 있는 수준의 요구사항이 정의되었는지 확인하는 데 초점을 둔다.

2. 필요성

▶ AI 시스템은 학습 데이터의 편향성, 알고리즘의 불투명성 등으로 인해 이용자에게 차별적·불공정한 결과를 제공할 가능성이 내재해 있다. 이를 개발 단계에서 요구사항으로 정의하지 않으면 시스템 가동 이후 이용자 구제 체계가 부재한 채 피해가 누적되고, 검수·감사 단계에서의 분쟁으로 이어질 수 있다.

▶ 설계 단계 요구반영의 중요성: 신고 채널 기능은 설계 초기에 요구사항으로 명시되어야 추가 개발 비용 없이 구현 가능하다. 요구사항 누락 시 설계 변경 공수 증가 및 계약 분쟁 발생 가능성이 높아진다.

▶ 이용자 권리구제 체계 조기 확보: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI, 2024)는 AI 이용자가 차별·편견 등으로 피해를 입은 경우 신고 및 이의를 제기할 수 있는 채널 구축을 권고하고 있으며, 이는 요구사항 단계부터 반영되어야 한다.

▶ 공공 AI 서비스의 법적·정책적 의무 준수:「지능정보화 기본법」 및 관련 공공기관 AI 도입 기준은 차별적 결과에 대한 구제 수단 마련을 요구하며, NIST AI RMF 1.0(MAP 5.1)은 AI 시스템의 부정적 영향에 대한 피드백 채널 식별을 요구한다.

▶ 감리 실효성 확보: 요구사항이 모호하거나 측정 불가능하게 정의된 경우 검수 단계에서 이행 적부 판단이 불가능하다. 구체성 점검은 향후 감리 지적의 실효성을 높이는 핵심 요소이다.

3. 검토방법

No	점검 항목	확인 방법
1	윤리적 우려사항 보고 채널 기능의 요구사항 명시 여부	{요구사항정의서}·'제안요청서'에서 보고 채널 유형·접근 방법·익명 신고 가능 여부가 요구사항으로 명시되었는지 대조 검토
2	차별·편견 신고 처리 절차의 요구사항 명시 여부	{요구사항정의서}·{과업대비표}에서 신고 처리 절차가 접수→검토→조치→결과통보 4단계로 명시되었는지 확인
3	신고 처리 담당자 및 응답 기한 정의 여부	{요구사항정의서} 검토와 발주기관 면담으로 신고 처리 담당자(운영 주체)와 응답 기한이 수치로 정의되었는지 확인
4	신고 내용 기록·보존 및 이력 관리 요구사항 포함 여부	{요구사항정의서}·{요구사항추적표}에서 신고 내용의 기록·보존과 감사 추적이 가능한 이력 관리 요구사항이 포함되었는지 검토
5	검수 기준 활용 가능 측정 지표(KPI/SLA) 정의 여부	{요구사항정의서}·{과업대비표}에서 신고 처리율·응답 소요시간 등 검수 기준으로 활용 가능한 측정 지표가 정의되었는지 비교
6	신고 기능의 UI 설계 요구사항 반영 여부	'UI/UX 요구사항 정의서'·{요구사항정의서}에서 신고 기능이 화면 접근성·가시성을 포함해 UI 요구사항에 반영되었는지 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	채널 유형·접근 방법·익명 신고 가능 여부, 접수→검토→조치→결과통보 4단계 처리 절차, 신고 처리 담당자와 응답 기한, 신고 내용의 기록·보존 및 감사 추적 요건, 검수 기준으로 활용 가능한 측정 지표(KPI/SLA), 신고 기능의 화면 접근성·가시성 요건이 모두 {요구사항정의서}에 구체적으로 정의된 경우	현행 유지 (해당 요구사항 이행 여부를 계속 추적)
보완권고 (△)	일부 항목이 기재되었으나 처리 절차 또는 측정 지표 등이 누락되었거나 수준이 불충분한 경우	결함으로 등록하고 특정 기한까지 요구사항 보완을 권고하며, 보완 결과를 차기 감리에서 재확인
부적합 (X)	{요구사항정의서}에 관련 요구사항이 전혀 기재되지 않았거나, "공정성을 확보한다" 수준의 추상적 문구만 존재하는 경우	결함으로 지적하고 요구사항 재정의(보완) 조치를 요구하며, 검수 단계 전 재점검 실시

5. 감리 시 유의사항 및 실무 Tip

1) 보고채널의 형식적 존재와 실질적 활용 가능성을 구분하여 점검하라: '신고함 설치', '이메일 주소 공개' 등 채널이 있더라도 접수 후 처리 절차, 익명성 보장, 피드백 주기가 없으면 사실상 작동하지 않는 채널이다.

2) 윤리적 우려사항 보고 후 '불이익 없음' 보장 조항이 명시되었는지 확인하라: 내부고발 보호 규정이 없으면 임직원이 실제로 채널을 사용하지 않아 형해화된다.

3) 접수된 우려사항의 처리·피드백 추적 절차를 확인하라: 처리 결과 공개(익명화) 및 재발방지 조치 여부가 채널 실효성의 핵심 지표다.

4) 추상적 요구사항의 수치화 유도: "AI는 공정한 결과를 제공해야 한다"는 문구는 검수 기준으로 사용할 수 없다. 이러한 모호한 문구를 발견하면 "신고 처리율 95% 이상", "편향 신고 응답 기한 3영업일 이내" 등 측정 가능한 기준으로 구체화하도록 즉시 지적하고 요구사항 보완을 요청해야 한다.

5) 제안요청서(RFP) 원문 반드시 확보: 점검 전 RFP 원문을 확보하여 {요구사항정의서}와 대조한다. 사업자 제출 산출물만으로는 요구사항 누락 여부를 판단할 수 없으며, RFP에 명시된 요건이 {요구사항정의서}에 반영되지 않은 경우 명백한 결함이므로 반드시 지적한다.

AG_SD41-2-10 | 다국어 및 접근성 지원 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | UI/UX, 접근성

근거 | 「장애인차별금지 및 권리구제 등에 관한 법률」·「지능정보화 기본법」· KWCAG

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, 장애인·노약자·외국어 사용자를 포함한 다양한 사용자 계층이 AI 서비스를 동등하게 이용할 수 있도록 하는 접근성 지원 요건(음성 입출력, 대체 UI, 화면 낭독기 호환성 등)과 다국어 처리 요건이 {요구사항정의서}에 법적 기준에 부합하고 측정 가능한 수준으로 반영되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

AI 서비스는 텍스트 입력과 화면 표시를 기본으로 설계되는 경우가 많아, 시각·청각·지체 장애인이나 디지털 기기 사용에 익숙하지 않은 노약자가 서비스 이용에서 체계적으로 배제될 위험이 있다.

요구정의 단계에서 접근성 요건이 명시되지 않으면 설계·개발 단계에서 반영이 구조적으로 어렵고 법적 의무 위반 및 서비스 개통 지연 리스크가 발생하므로, 이를 사전에 식별하고 점검 기준을 확보하는 것이 핵심 목적이다.

2. 필요성

▶ 공공 AI 서비스는 전 국민을 대상으로 하는 서비스 특성상 모든 사용자 계층이 동등하게 이용할 수 있어야 하는 법적·사회적 의무를 지며, AI 기반 인터페이스는 일반 웹 서비스보다 접근성 설계가 더 복잡하기 때문에 요구정의 단계에서 명시적인 요건 정의가 더욱 중요하다.

▶ 법적 의무 이행 위험:「장애인차별금지 및 권리구제 등에 관한 법률」및「지능정보화 기본법」에 따라 공공기관이 운영하는 전자정보·서비스는 장애인도 동등하게 이용할 수 있어야 하며, 한국형 웹 콘텐츠 접근성 지침(KWCAG 2.2)을 준수해야 할 의무가 있다. 접근성 요건이 {요구사항정의서}에 반영되지 않으면 이 의무를 이행하지 않는 시스템이 개통될 위험이 있다.

▶ AI 고유 접근성 과제의 요건 설계 필요: AI 챗봇 인터페이스, 음성 입출력, AI 생성 이미지 등은 일반 웹 접근성 지침만으로는 충분히 커버되지 않는 AI 특유의 접근성 고려 사항이 존재한다. AI 응답 텍스트의 화면 낭독기 호환성, 음성 인식 오류 발생 시 대체 입력 수단 요건 등이 별도로 명시되어야 한다.

▶ 다국어 사용자의 서비스 배제 리스크: 외국인 주민, 결혼이민자, 외국인 근로자 등 한국어가 모국어가 아닌 사용자가 공공 AI 서비스를 이용해야 하는 상황이 증가하고 있다. 다국어 지원 요건이 요구정의 단계에서 명시되지 않으면 이들 사용자는 서비스에서 체계적으로 배제된다.

▶ 설계 단계 반영 시점 의존성: 접근성 기능(화면 낭독기 호환 마크업 구조, 음성 입출력 인터페이스, 고대비 모드 등)은 UI 설계 단계에서 구조적으로 반영해야 하며, 개발 완료 후 추가 구현이 매우 어렵고 비용이 크다. 명시가 유일한 사전 예방 수단이다.

3. 검토방법

No	점검 항목	확인 방법
1	웹 접근성 기준 준수 요건 및 적용 범위 명시 여부	{요구사항정의서}·'UI/UX 설계 방향서'에서 웹 접근성 준수 기준(KWCAG 2.2 이상)과 예외 범위, 챗봇·AI 생성 콘텐츠·음성 입력 등 AI 고유 인터페이스 접근성 요건, 스크린리더 호환성·검사 방법이 명시되었는지 발주기관 면담과 함께 확인
2	장애인·노약자 대체 입출력 수단 요건 명시 여부	{요구사항정의서}·'사용자 인터페이스 요구사항서'에서 시각장애인용 음성 출력(TTS)·음성 입력(STT) 요건과 고대비 모드·글자 크기 조정·키보드 접근성 요건이 명시되었는지 확인하고 발주기관 면담으로 보완
3	다국어 입출력 지원 요건의 구체성	{요구사항정의서}·'다국어 지원 계획서'에서 지원 언어 목록(언어 코드 수준), 동일 언어 응답·언어 자동 감지 요건, 다국어 응답 품질 기준과 고유명사 번역 정확성 요건이 명시되었는지 교차 검토
4	접근성·다국어 요건의 검사 기준 및 검증 방법 명시 여부	{요구사항정의서}·{과업대비표}에서 자동화 접근성 검사 도구 활용 기준·실사용 테스트 포함 여부·항목별 적합성 비율 기준과 다국어 품질 검증 방법·유지관리 요건이 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	점검 항목 모두가 {요구사항정의서}에 법적 기준을 포함하여 측정 가능한 수준으로 구체적으로 기술된 경우 (KWCAG 2.2 이상 기준 및 AI 인터페이스 별도 접근성 요건, TTS·STT·고대비·키보드 접근성 요건, 대상 언어 목록·응답 품질 기준·언어 전환 요건, 접근성 검사 기준과 다국어 품질 검증 방법이 독립 항목으로 반영)	현행 유지
보완권고 (△)	일부 항목이 반영되었으나 구체성·법적 기준 부합 여부 또는 완전성이 미흡한 경우 (KWCAG 기준 명시 없는 선언적 기술, AI 인터페이스 접근성 요건 누락, 대상 언어 목록 미명시, 다국어 응답 품질 기준 부재, 접근성 검사 방법 미반영 중 하나 이상)	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	접근성·다국어 관련 요건이 {요구사항정의서}에 전혀 반영되지 않은 경우, 특히 법적 의무 대상 공공 AI 서비스임에도 장애인 접근성 요건이 전혀 포함되지 않아 법적 위반 리스크가 명백한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 및 법적 리스크 보고

5. 감리 시 유의사항 및 실무 Tip

1) 다국어 지원 범위 구체화 유도: 다국어 지원 범위가 "주요 언어"처럼 선언적으로만 표현된 경우 지원 언어 목록과 우선순위를 명시하도록 요구한다.

2) 요구사항 구체성 기준 적용: 접근성·다국어 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) RFP와 요구사항정의서 교차 검토: RFP에 접근성·다국어 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

4) 접근성 자동화 도구 활용: 접근성 자동화 도구(WAVE, axe 등)를 활용한 사전 점검을 권장하되, 자동화 도구 검사만으로는 충분하지 않으므로 장애인 사용자 실사용 테스트 계획 여부를 발주기관 면담으로 확인한다.

5) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 문서상 기재가 있어도 담당자가 내용을 인지하지 못하는 경우가 빈번하므로 실질 적정성 검증이 중요하다.

AG_SD41-2-11 | 예외처리 정책 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | UI/UX, 접근성

근거 | OWASP LLM Top 10 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, AI 서비스에서 발생 가능한 오류 및 예외 상황(AI 모델 추론 실패, 환각 응답 탐지, 외부 API 연계 장애, 콘텐츠 필터링 차단, 사용자 입력 오류 등)에 대해 사용자에게 제공해야 할 예외 안내 메시지의 내용 기준·표시 방식·대체 행동 안내·시스템 내부 정보 노출 금지 요건이 {요구사항정의서}에 오류 유형별로 구체적으로 정의되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

AI 서비스의 예외 처리는 환각(Hallucination), 컨텍스트 초과, 스트리밍 중단 등 AI 특유의 오류 유형이 존재하고, 오류 메시지를 통해 내부 프롬프트 구조·모델명·버전 정보 등이 노출되면 AI 공격 표면(Attack Surface)이 확장된다는 점에서 일반 정보시스템의 오류 처리와 근본적으로 다르다.

이 두 가지 특수성이 요건화되지 않으면 설계·개발 단계에서 오류 메시지가 사업자 임의로 구현되어 사용자 혼란과 보안 취약점이 동시에 발생한다.

2. 필요성

▶ AI 서비스의 예외 처리 정책은 사용자 신뢰 유지, 보안 취약점 방지, 서비스 지속성 확보의 세 가지 목적을 동시에 충족해야 하는 복합 요건으로, 요구정의 단계에서 명확히 정의되지 않으면 다음과 같은 리스크가 발생한다.

▶ AI 특화 오류에 대한 사용자 안내 부재로 인한 신뢰 훼손: AI 서비스에서 오류가 발생했을 때 '오류가 발생했습니다. 잠시 후 다시 시도해 주세요'라는 일반 메시지만 표시되면, 사용자는 AI가 잘못된 정보를 제공하고 있는지, 서비스가 일시 중단된 것인지, 본인의 입력에 문제가 있는 것인지를 구별할 수 없다. 오류 상황에서 사용자가 취해야 할 행동을 명확히 안내하지 않으면 서비스 이탈로 이어진다.

▶ 환각 응답 탐지 시 처리 정책 미정의 리스크: AI 모델이 사실과 다른 정보를 생성하는 환각 현상은 AI 서비스 특유의 예외 상황이다. 환각이 탐지되었을 때 전체 응답을 차단할 것인지, 불확실성 경고와 함께 제공할 것인지, 사용자에게 재확인을 요청할 것인지에 대한 정책이 {요구사항정의서}에 없으면 개발자가 임의로 결정하게 되어 서비스 간 일관성이 없어진다.

▶ 내부 시스템 정보 노출에 따른 보안 취약점: 오류 메시지에 스택 트레이스(Stack Trace), 내부 프롬프트 구조, 모델명·버전, 데이터베이스 쿼리 정보 등이 포함되면 공격자는 이를 활용하여 탈

옥 시도, 프롬프트 인젝션, 시스템 침투 경로를 설계할 수 있다. 오류 메시지에서 노출 금지 항목을 요건으로 명시하지 않으면 보안 취약점이 설계 단계부터 내재된다.

▶ 오류 상황에서의 대체 서비스 안내 미정의: AI 서비스가 오류 상태일 때 사용자가 동일한 업무를 다른 채널(전화, 방문, 다른 URL 등)을 통해 처리할 수 있도록 대체 경로를 안내하는 요건이 없으면, 공공 서비스의 경우 민원인이 서비스 이용 불가 상태에서 어디로도 안내받지 못하는 상황이 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 특화 오류 유형 분류 및 유형별 안내 정책 요건 명시 여부	{요구사항정의서}·'AI 서비스 아키텍처 구성도'에서 일반 시스템 오류와 AI 특화 오류(입력·출력·연계 관련)의 분류 체계와, 귀책 주체(사용자·시스템·AI 불확실성)별 안내 차별화 요건이 기술되었는지 확인
2	예외 안내 메시지 구성 기준 및 표시 방식 요건 명시 여부	{요구사항정의서}·'AI 서비스 화면 설계서'에서 안내 메시지 구성 요소(상황 인식·원인·행동 안내·AI 특화 고지)와 표시 방식(위치·유지 시간·스트리밍 중단 처리), 문구 작성 원칙이 명시되었는지 확인
3	오류 메시지 내 내부 시스템 정보 노출 금지 요건 명시 여부	{요구사항정의서}에서 노출 금지 항목(스택 트레이스·시스템 프롬프트·API·모델명·쿼리·내부 IP 등)과 검증 방법(오류 유발 시나리오·응답 헤더 검사), 내부 오류 코드와 표시 메시지 분리 구조 요건이 포함되었는지 확인
4	오류 상황 대체 서비스 안내 및 일관성 관리 요건 명시 여부	{요구사항정의서}·{과업대비표}에서 전체 불가 시 대체 채널 안내, 부분 오류 시 가용 기능 유도, 점검 사전 공지 요건과 문구 표준화·다국어·변경 절차 및 검사 기준이 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	점검 항목 모두가 {요구사항정의서}에 AI 특화 오류 유형을 포함하여 메시지 구성 기준·보안 제약·대체 안내까지 구체적으로 기술된 경우 (AI 특화 오류 유형 분류, 귀책 주체별 안내 방향 차별화, 메시지 구성 4요소·표시 방식·문구 원칙, 내부 정보 노출 금지 목록·검사 시나리오, 대체 채널 안내·문구 표준화 원칙이 독립 항목으로 반영)	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	일부 항목이 반영되었으나 AI 특화 오류 유형 누락·구성 기준 미명시·보안 제약 부재 중 하나 이상에 해당하는 경우 (일반 IT 오류 처리만 있고 AI 특화 오류 유형 누락, 환각·스트리밍 중단 안내 정책 부재, 내부 정보 노출 금지 항목 부재, 대체 채널 안내 요건 누락 등)	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 서비스 예외 처리 정책 요건이 {요구사항정의서}에 전혀 반영되지 않아, 오류 발생 시 사용자 안내 방식이 사업자 임의로 결정되며 내부 정보 노출·사용자 미안내·대체 경로 부재 상태로 서비스가 개통될 리스크가 명백한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 특화 예외 유형 포함 여부 확인: LLM 타임아웃·컨텍스트 초과·유해 콘텐츠 감지 등 AI 특화 예외 유형이 예외처리 정책에 포함되지 않으면 보완을 요구한다.

2) 요구사항 구체성 기준 적용: 예외 처리 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) 처리 절차의 시각화 요구: 예외 유형별 처리 절차를 플로우차트로 제출하도록 요구하여 오류 발생 지점별 처리 흐름의 완전성을 확인한다.

4) RFP와 요구사항정의서 교차 검토: RFP에 예외 처리·오류 안내 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

5) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 문서상 기재가 있어도 담당자가 내용을 인지하지 못하는 경우가 빈번하므로 실질 적정성 검증이 중요하다.

AG_SD41-2-12 | 접속 로깅 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | 개인정보보호

근거 | 「개인정보 보호법」·「전자정부법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, AI 서비스 이용자의 신원 식별 정보(사용자 ID·접속 IP·세션 정보 등), 입력 프롬프트 전문, AI 응답 출력, 접속 일시·채널 등 접속 맥락 정보를 기록하는 로깅 요건이 개인정보 처리 기준·로그 보관 기간·접근 통제 기준과 함께 {요구사항 정의서}에 구체적으로 정의되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

AI 서비스의 접속 로그는 프롬프트에 이름·증상·민원 내용·계좌 정보 등 고도로 민감한 개인정보가 포함될 가능성이 높아 로깅 자체가 개인정보 수집·처리에 해당하며, 단순 보안 감사 목적을 넘어 AI 출력의 책임 소재 규명(어떤 입력에 어떤 출력이 생성되었는지)·AI 오남용 탐지·AI 편향 사후 검토의 핵심 증거 자료로 기능한다는 점에서 일반 접속 기록과 근본적으로 다르다.

이 특수성이 요구정의 단계에서 요건화되지 않으면 설계·개발 단계에서 로깅 항목과 개인정보 처리 기준이 일관성 없이 구현되어 사후 책임 추적이 불가능한 상태로 서비스가 개통된다.

2. 필요성

▶ AI 서비스의 접속 로깅 요건은 서비스 투명성·책임성·법적 의무 이행의 세 가지 축을 동시에 뒷받침하는 핵심 요건으로, 요구정의 단계에서 명확히 정의되지 않으면 다음과 같은 리스크가 발생한다.

▶ AI 출력 책임 소재 규명 불가 리스크: AI 서비스가 잘못된 정보를 제공하거나 차별적 표현을 출력한 경우, 해당 사건을 사후에 분석하려면 '어떤 사용자가 어떤 프롬프트를 입력했을 때 어떤 응답이 생성되었는지'를 로그로 재현할 수 있어야 한다. 프롬프트 로그가 없으면 문제 발생 시 원인 규명과 책임 소재를 확인할 수 없다.

▶ AI 오남용 탐지 및 보안 감사 수행 불가 리스크: 사용자가 탈옥 시도, 민감 정보 추출 시도, 반복적 편향 유도 질의 등 AI 오남용 행위를 할 경우, 이를 탐지하고 사후에 감사하려면 사용자 식별 정보와 입력 내용이 연계된 로그가 필요하다. 이 요건이 없으면 AI 오남용 행위가 발생해도 사후 추적이 불가능하다.

▶ 개인정보 처리 근거 부재 리스크: 프롬프트에 포함된 개인정보를 로그로 기록하는 행위는 개인정보 보호법상 개인정보 처리에 해당한다. 로깅 목적·보관 기간·접근 권한·파기 절차가 {요구사항정의서}에 명시되지 않으면 개인정보 처리 방침에 근거 없는 개인정보 수집·보유가 발생하여 법령 위반 리스크가 발생한다.

▶ 감사(Audit) 요건 미충족으로 인한 행정 제재 리스크: 공공기관이 운영하는 AI 서비스는 정보시스템 감리기준, 전자정부법, 행정안전부 AI 서비스 운영 지침 등에 따라 접속 기록 보관 의무가 있다. 보관 기간·항목·형식이 법령 기준에 맞게 {요구사항정의서}에 반영되지 않으면 법적 보관 의무 미이행으로 행정 제재를 받을 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	사용자 식별 정보 로깅 항목 및 수준 요건 명시 여부	{요구사항정의서}·'개인정보처리방침(초안)'에서 로그인·비로그인 사용자 식별 항목, 접속 채널·타임스탬프 등 기록 항목과 서비스 목적 적합성, IP 등 개인정보 처리 근거(목적·보관 기간)가 명시되었는지 확인
2	프롬프트 및 AI 응답 출력 로깅 요건 명시 여부	{요구사항정의서}·'개인정보처리방침(초안)'에서 프롬프트·응답 기록 방식(전문·요약·해시), 멀티턴 세션 맥락 기록, 요청 식별자(Request ID), 프롬프트 내 개인정보 저장·마스킹 요건이 포함되었는지 검토
3	로그 보관 기간·접근 통제·파기 요건 명시 여부	{요구사항정의서}·'개인정보처리방침(초안)'에서 로그 유형별 보관 기간(접속 6개월·감사 1년 이상 등), 역할별 조회 권한 구분, 보관 만료 후 파기 절차·이력 기록 요건이 명시되었는지 검토
4	로그 무결성 보장 및 AI 오남용 탐지 연계 요건 명시 여부	{요구사항정의서}·{과업대비표}에서 로그 위변조 방지(쓰기 전용·해시 검증·저장소 접근 통제), 프롬프트 로그 기반 오남용 탐지(탈옥·유해 입력·대량 호출)와 조회·리포트 기능 요건이 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	점검 항목 모두가 {요구사항정의서}에 AI 서비스 특화 로깅 특성과 개인정보 처리 의무를 반영하여 구체적으로 기술된 경우 (사용자 식별 항목·프롬프트 로깅 방식(마스킹 포함), 요청 식별자 연계, 로그 유형별 보관 기간, 역할별 접근 통제, 파기 절차, 무결성 보장 요건, AI 오남용 탐지 연계 요건이 독립 항목으로 반영)	현행 유지
보완권고 (△)	일부 항목이 반영되었으나 구체성·개인정보 처리 연계·AI 특화 요건이 미흡한 경우 (로깅 요건의 선언적 기술, 프롬프트 로깅 방식 미결정, 개인정보 처리 근거 연계 부재, 보관 기간 법적 기준 미충족, 역할별 접근 통제 미명시, 파기 절차 누락, 로그 무결성 요건 부재 중 하나 이상)	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	AI 서비스 접속 로깅 요건이 {요구사항정의서}에 전혀 반영되지 않아, AI 서비스 오남용 발생 시 사후 추적이 불가능하고 법적 접속 기록 보관 의무를 이행할 수 없는 상태인 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (개인정보 보호법 준수 촉구)

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항 구체성 기준 적용: 접속 로깅 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

2) 프롬프트 로깅 방식 결정 여부 확인: 프롬프트를 전문(Full-text) 저장할 것인지, 요약 저장할 것인지, 개인정보를 마스킹 후 저장할 것인지가 {요구사항정의서}에 명시되어 있는지 확인한다. 이 결정이 없으면 개발자가 임의로 결정하게 되어 개인정보 과도 수집 문제가 발생할 수 있다.

3) RFP와 요구사항정의서 교차 검토: RFP에 접속 기록·로그 관리 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

4) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 특히 로그 보관 기간의 결정 근거와 역할별 접근 권한 설계를 발주기관 면담으로 확인한다.

5) 산출물 간 정합성 교차 확인: {요구사항정의서}에 기재된 내용이 {과업대비표}·{사업수행계획서} 및 '개인정보처리방침(초안)'과 일치하는지 교차 확인한다. 산출물 간 불일치가 있으면 어느 쪽이 기준인지 명확히 하고, 혼선이 없도록 조기에 확인한다.

AG_SD41-2-13 | 시스템 오작동 시 기본값 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | 가용성/이중화

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MANAGE

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 감리원이, AI 시스템이 오작동·장애·비정상 입력·연계 시스템 오류 등의 이상 상태에 직면했을 때 사용자와 서비스에 미치는 피해를 최소화하기 위한 기본값(Fail-safe Default) 동작 요건 및 장애 대응 프로세스 요건이 {요구사항정의서}에 구체적이고 측정 가능한 수준으로 반영되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 일반 소프트웨어와 달리 모델 응답 불가, 환각(Hallucination) 출력, 외부 API 연계 장애 등 AI 특유의 오작동 유형이 존재하며, 이러한 상황에서 시스템이 어떻게 동작해야 하는지가 요구정의 단계에서 명문화되지 않으면 설계·개발 단계에서 장애 대응 로직이 누락되어 서비스 중단이나 잘못된 정보 제공 등의 피해가 발생한다.

이를 사전에 차단하고 이후 단계의 추적 가능한 점검 기준을 확보하는 것이 핵심 목적이다.

2. 필요성

▶ AI 서비스는 AI 모델, 외부 API, 데이터베이스, 연계 시스템 등 복수의 구성 요소로 이루어진 복합 시스템으로, 어느 한 구성 요소에 장애가 발생해도 전체 서비스에 영향을 줄 수 있으며 AI 특유의 비결정론적 응답 특성은 오작동 탐지와 대응을 더욱 어렵게 하므로, 요구정의 단계에서 기본값 및 장애 대응 요건을 명시해야 한다.

▶ AI 고유 오작동 유형에 대한 대응 체계 부재 리스크: 일반 IT 시스템 장애와 달리 AI 시스템은 모델이 부적절하거나 사실과 다른 응답(환각)을 자신 있게 출력하거나, 외부 지식 API 연계 실패 시 불완전한 응답을 정상 응답처럼 제공하는 오작동이 발생한다. 이에 대한 탐지 및 기본값 동작 요건이 없으면 사용자는 오작동 여부를 인지하지 못하고 잘못된 정보를 신뢰하게 된다.

▶ 요건 부재 시 설계·개발 단계 반영 불가: 장애 상황별 디폴트 동작(서비스 중단 안내, 대체 응답 제공, 담당자 연결 등)은 응용시스템 화면 설계 및 예외 처리 로직과 직결되므로, 요구정의 단계에서 명시되지 않으면 설계 단계 이후 반영이 현실적으로 어렵다.

▶ 서비스 연속성 및 사용자 보호 의무: 공공 AI 서비스는 장애 발생 시에도 사용자에게 최소한의 안내와 대체 경로를 제공해야 하는 의무가 있다. 장애 시 디폴트 행동이 정의되지 않으면 서비스가 무응답 상태로 방치되거나 오류 메시지가 사용자에게 그대로 노출되는 상황이 발생한다.

▶ 검사 단계 판정 기준 확보: '장애 발생 시 적절히 대응한다'는 선언적 문구는 검사 단계에서 이행 여부를 판정할 수 없다. 장애 유형별 대응 방식, 전환 조건, 사용자 안내 내용이 요건으로 명시되어야 검사 기준이 된다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 오작동·장애 유형 분류 및 탐지 요건 명시 여부	{요구사항정의서}·'AI 서비스 아키텍처 구성도'에서 모델 무응답·비정상 응답(환각·반복 루프)·외부 연계 실패·입력 품질 불량·과부하 등 장애 유형 분류와 유형별 탐지 기준(타임아웃·오류 코드·임계값)이 명시되었는지 확인
2	장애 유형별 디폴트 행동(Fail-safe Default) 요건 명시 여부	{요구사항정의서}에서 장애 유형별 기본값 동작(대체 응답·출력 차단 후 재시도·가용 기능 유지·요청 대기열 관리 등)과 디폴트 전환의 자동화 여부(자동 또는 관리자 승인)가 명시되었는지 검토하고 발주기관 면담으로 보완
3	장애 발생 시 사용자 안내 요건 명시 여부	{요구사항정의서}에서 장애 고지 문구·예상 복구 시간·대체 채널 안내, 지속 시간별 안내 단계 확대, AI 오작동임을 구분 고지하는 요건이 명시되었는지 확인
4	장애 기록(로깅) 및 복구 절차 요건 명시 여부	{요구사항정의서}·{과업대비표}에서 장애 이력 기록(유형·시각·지속시간·영향·조치), 관리자 자동 알림, 정상 복구 전환, 로그 보관 기간·접근 권한 요건이 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	점검 항목 모두가 {요구사항정의서}에 측정 가능한 수준으로 구체적으로 기술된 경우 (AI 특유 장애 유형 분류·탐지 기준 수치 명시, 장애 유형별 디폴트 행동·전환 조건 기술, 사용자 안내 문구·대체 채널·단계별 확대 기준 포함, 장애 로그·알림·복구 전환 요건이 독립 항목으로 반영)	현행 유지
보완권고 (△)	일부 항목이 반영되었으나 구체성 또는 완전성이 미흡한 경우 (장애 대응 요건은 있으나 AI 특유 유형(환각, 비정상 응답) 미분류, 디폴트 행동 요건의 선언적 기술, 사용자 안내 문구 미정의, 장애 로그 항목·보관 기간 기술 부재 중 하나 이상)	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 오작동 시 기본값(디폴트 행동) 관련 요건이 {요구사항정의서}에 전혀 반영되지 않았거나, 디폴트 행동·사용자 안내·장애 로그 등 핵심 요건이 모두 부재하여 서비스 연속성 체계가 결여된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 기본값(Default State)의 명확한 정의 확인: 기본값이 서비스 중단인지 제한 응답인지 명확히 정의되지 않으면 장애 시 이용자 피해 범위가 불명확해지므로, 장애 유형별 기본 상태가 구체적으로 정의되어 있는지 확인한다.

2) 요구사항 구체성 기준 적용: AI 오작동·장애 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) 시나리오 기반 안전성 검증: 오작동 시 기본값이 실제 안전한 상태인지 시나리오 테스트를 통해 확인하도록 권고하고, 전환 조건(시간 기준, 오류 횟수 등)이 수치로 기술되어 있는지 점검한다.

4) RFP와 요구사항정의서 교차 검토: RFP에 오작동·장애 대응 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

5) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 문서상 기재가 있어도 담당자가 내용을 인지하지 못하는 경우가 빈번하므로 실질 적정성 검증이 중요하다.

AG_SD41-2-14 | Fail-safe 전환 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | 가용성/이중화

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템에 치명적 오류(Critical Failure)가 발생하였을 때 시스템이 안전한 상태(Fail-safe State)로 자동 전환되도록 하는 요건을 {요구사항정의서}에 구체적이고 상세하게 정의하였는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 단순한 예외 처리(Exception Handling) 요구사항을 넘어, ① AI 모델의 오작동·환각(Hallucination)·응답 불능 등 AI 고유의 오류 유형에 대한 안전 전환 시나리오, ② 안전 상태의 구체적 정의, ③ 사용자 안내 및 대체 처리 방안, ④ 전환 후 로그 기록·보고 요건이 요구사항으로 명시되었는지를 확인한다.

2. 필요성

▶ AI 시스템은 일반 정보시스템과 달리 모델 예측의 불확실성, 학습 데이터 분포 변화, 환각(Hallucination) 등 구조적으로 예측 불가능한 오류가 상시 존재한다. 이러한 오류가 공공 서비스나 업무 의사결정에 직접 영향을 미치는 환경에서, 오류 발생 시 시스템이 스스로 안전한 상태로 전환되는 기능이 없다면 사용자 피해와 서비스 신뢰도 저하로 직결된다.

▶ AI 고유 오류의 비예측성 대응: 기계학습 기반 AI 시스템은 입력값의 범위 이탈, 모델 드리프트, 외부 LLM API 장애 등 일반 소프트웨어에서는 존재하지 않는 오류 유형이 존재하며, 요구정의 단계에서 이에 대한 안전 전환 요건이 정의되지 않으면 설계·개발 단계에서 기능이 누락될 위험이 높다.

▶ 서비스 연속성 및 사용자 보호: AI 응답 불능, 모델 서버 타임아웃 등 상황에서 대체 처리 경로(폴백)가 요구사항으로 명시되지 않으면, 사용자는 오류 원인을 알지 못한 채 서비스 중단 상태에 노출된다.

▶ 공공 AI 서비스의 법적·제도적 책임 이행: TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI 기반 서비스 분야)는 AI 시스템 안전성 요건으로 오류 발생 시 안전 상태 전환 및 사용자 안내 기능 정의를 명시적으로 요구하고 있으며, NIST AI RMF MEASURE 2.6은 시스템이 '안전하게 실패(Fail Safely)'하도록 설계할 것을 요구한다.

▶ 개발 단계 재작업 예방: 요구정의 단계에서 Fail-safe 요건이 누락되면 설계·구현 단계에서 기능 추가에 따른 공수 증가, 아키텍처 변경 등이 발생하여 사업 일정 및 품질에 부정적 영향을 미친다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 오류 유형별 안전 전환 시나리오 정의 여부	환각·응답 불능·모델 오작동·외부 API 장애 등 AI 고유 오류 유형별 안전 전환 시나리오가 {요구사항정의서} 비기능(안전성) 요구사항에 정의되었는지 확인하고 사업자 면담
2	안전 상태 정의·전환 조건 명시 여부	안전 상태(Fail-safe State)의 구체적 정의와 오류 판정 임계값 등 전환 조건이 {요구사항정의서}에 명시되었는지 확인하고 발주기관·사업자 면담
3	전환 시 사용자 안내·대체 처리(폴백) 명시 여부	Fail-safe 전환 시 사용자 안내 방식과 대체 처리(폴백) 방안이 {요구사항정의서} 기능 요구사항에 명시되었는지 확인하고 발주기관 면담
4	전환 로그 기록·관리자 보고 요건 포함 여부	Fail-safe 전환 발생 시 로그 기록 및 관리자 보고 요건이 {요구사항정의서} 관리·운영 요구사항에 포함되었는지 확인
5	RFP 안전성 요구사항 반영 여부	RFP의 안전성 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표}와 대조

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	Fail-safe 전환 요건이 구체적 시나리오·조건·대체 처리 방식과 함께 {요구사항정의서}에 명시되고, 사용자 통보·로그 기록 요건도 포함된 경우	오류 유형별 안전 전환 시나리오, 임계값 정의, 사용자 안내 문구 요건, 대체 수행 절차가 모두 포함된 사례 → 현행 유지
보완권고 (△)	Fail-safe 전환 관련 요구사항이 일부 반영되었으나 구체성·완전성이 미흡하거나, 특정 오류 유형에 국한된 경우	'시스템 오류 시 안전 처리한다'는 선언적 문구만 존재하고 전환 조건·임계값·대체 흐름이 미명시된 사례 → 해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	Fail-safe 전환 요건이 {요구사항정의서}에 전혀 반영되지 않은 경우	AI 오류 상황에 대한 처리 요구사항 자체가 부재하고 단순 '예외 처리' 언급도 없는 사례 → 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) Fail-safe 전환 요건은 핵심 위험 요소(오류 유형·전환 조건·대체 흐름)를 중심으로 점검 깊이를 조정하고, 발견사항을 구체적 근거와 함께 기록하라.

2) AI 치명적 오류 발생 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) 면담 시 안전 전환의 발동 조건(임계값), 안전 상태의 정의, 정상 복귀 절차를 사업자에게 직접 질문하여 요건의 실질적 구체성을 확인한다.

4) RFP에 AI 치명적 오류 발생 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

AG_SD41-2-15 | UX로그 수집 요건

응용시스템 | 요구정의·분석 | 시스템 구현 | 모니터링/재학습

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 제안된 AI 시스템이 사용자 경험(UX) 로그를 체계적으로 수집·분석할 수 있도록 요건을 구체적으로 정의하였는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 단순히 로그 수집 기능의 존재 여부를 확인하는 것이 아니라, 수집 항목의 구체성, 상호작용(멀티턴) 로그 구조의 반영, 수집 데이터의 AI 모델 개선 피드백 루프 설계, 개인정보 보호 및 사용자 동의 요건의 포함 여부를 종합적으로 검토한다.

2. 필요성

▶ 생성형 AI 서비스는 사용자의 입력(프롬프트), AI의 응답, 오류, 재시도, 사용자 피드백(평가) 등 다양한 상호작용 로그를 생성하며, 이러한 UX 로그는 AI 모델의 품질 개선, 서비스 안전성 확보, 이용자 만족도 향상에 직결되므로 수집 요건이 명확히 정의되어야 한다.

▶ AI 품질 개선 기반 확보: 수집된 UX 로그는 AI 모델의 환각(Hallucination), 오류 패턴, 응답 품질 저하를 조기에 감지하고 개선하는 핵심 데이터로, 요건 미정의 시 운영 이후 품질 개선이 사실상 불가능하다.

▶ 멀티턴 대화 특성 반영: 생성형 AI는 단일 Q&A가 아닌 연속대화(Multi-turn) 구조로 운영되므로, 세션 단위 로그와 컨텍스트 연속성을 반영한 로그 수집 요건이 명세화되어야 한다.

▶ 개인정보 보호 선제적 설계: UX 로그에는 사용자 입력 내용이 포함될 수 있어 개인정보 침해 가능성이 있다. 비식별화·가명처리 요건과 사용자 동의 절차를 요구정의 단계에서 설계하지 않으면 개인정보 보호법 위반 및 사후 재설계 비용이 발생한다.

▶ 피드백 루프 설계: TTA 신뢰할 수 있는 인공지능 개발안내서는 AI 서비스의 지속적 모니터링·개선 체계를 요구하며, 이를 위한 로그 분석 및 피드백 루프는 요구정의 단계에서 설계 기반이 마련되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	UX 로그 수집 항목 요건 정의 적정성	{요구사항정의서}에 사용자 입력 로그(프롬프트 텍스트, 입력 시각, 세션 ID), AI 응답 로그(응답 내용, 응답 시간, 모델 버전), 오류 및 재시도 로그, 사용자 피드백 로그(만족도 평가, 오류 신고 등) 등 주요 로그 항목이 유형별로 명시되었는지 확인
2	상호작용 로그의 멀티턴 대화 구조 요건 반영 여부	생성형 AI 특유의 멀티턴 대화 흐름(세션 단위 로그 수집 기준, 컨텍스트 연속성 식별자(Conversation ID 등), 대화 흐름 단절 시 로그 처리 방안)이 {요구사항정의서}에 포함되었는지 확인
3	로그 수집의 목적·활용 범위 명확성	{요구사항정의서} 또는 'AI 로그 수집 및 분석 계획서'에 수집 로그가 AI 모델 성능 개선, 서비스 오류 원인 분석, 사용자 경험 향상 등 구체적 활용 목적과 연결되어 명시되었는지 확인
4	로그 분석 요건의 구체성 및 피드백 루프 설계	로그 수집·분석 주기, 분석 결과의 모델 재학습 또는 프롬프트 개선 반영 절차, 피드백 루프 담당 조직·역할이 {요구사항정의서}에 요건으로 설계되었는지 확인
5	개인정보 보호 및 비식별화 처리 요건 포함 여부	{요구사항정의서} 또는 'AI 로그 수집 및 분석 계획서'에 로그 내 개인정보 포함 가능성 식별 및 처리 방침, 저장 전 자동 비식별화(마스킹·가명처리) 요건, 로그 보존 기간·파기 기준, 접근통제·위변조 방지 요건이 명시되었는지 확인
6	사용자 동의 기반 로그 수집 절차 설계 적정성	'데이터 수집·활용 동의 설계서'에 로그 수집 목적·범위의 명시적 고지, 사용자 동의·거부 선택권 보장, 미동의 시 핵심 서비스 이용이 제한되지 않는다는 원칙이 요건으로 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	UX 로그 수집 항목과 멀티턴 대화 구조 요건, 로그 수집 목적·활용 범위, 분석 결과의 모델 재학습·프롬프트 개선 피드백 루프가 {요구사항정의서}에 구체적으로 명시되고, 개인정보 보호·비식별화 및 사용자 동의 절차가 포함된 경우	현행 유지 (설계 및 요건 반영 여부를 지속 점검할 것을 권고)

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	UX 로그 수집 항목이 일부 정의되었으나 멀티턴 구조, 피드백 루프, 또는 개인정보 처리 요건 등 일부 항목이 누락된 경우	누락 항목을 {요구사항정의서}에 보완하고, 해당 요건이 설계 단계에 반영될 수 있도록 조치 요청 후 재확인
부적합 (X)	UX 로그 수집 요건 자체가 {요구사항정의서}에 없거나, 상호작용 로그 분석 계획이 전혀 수립되지 않은 경우	요구사항 정의 재수행을 권고하고 'AI 로그 수집 및 분석 계획서' 작성 및 사용자 동의 설계를 즉시 수립하도록 시정조치 요구 후 이행 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항 구체성 기준 적용: UX 로그 수집 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

2) 로그 항목 체크리스트 활용: 사용자 입력·AI 응답·오류 및 재시도·사용자 피드백의 4가지 로그 유형 중 2~3가지만 정의된 경우 보완권고, 1가지 이하만 정의되었거나 전혀 정의되지 않은 경우 부적합으로 판정하고, {요구사항정의서} 내 로그 수집 요건을 항목별로 매핑하여 누락 여부를 확인한다.

3) RFP와 요구사항정의서 교차 검토: RFP에 UX 로그 수집 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

4) 면담을 통한 이해도 및 구현 계획 확인: 사업자가 해당 요구사항의 의미를 정확히 이해하고 구현 계획을 보유하고 있는지 면담으로 확인한다. 문서상 기재가 있어도 담당자가 내용을 인지하지 못하는 경우가 빈번하므로 실질 적정성 검증이 중요하다.

5) 산출물 간 정합성 교차 확인: {요구사항정의서}에 기재된 내용이 {과업대비표}·{사업수행계획서}와 일치하는지 교차 확인한다. 산출물 간 불일치가 있으면 어느 쪽이 기준인지 명확히 하고, 혼선이 없도록 조기에 확인한다.

AG_SD42-2-01 | UI 색상배치 편향배제 다크패턴 설계

응용시스템 | 설계 | 시스템 구현 | 편향/공정성

근거 | EU AI Act · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

AI 서비스 구축 사업에서, {화면설계서}에 설계된 AI 서비스 UI가 색상·배치·크기·문구 등의 시각적 요소를 통해 이용자의 자유로운 판단과 선택을 방해하거나 특정 선택을 의도적으로 유도하는 다크패턴(Dark Pattern) 설계가 포함되어 있지 않은지를 감리원의 관점에서 점검하는 것이 목적이다.

AI 서비스에서 다크패턴은 AI 추천·AI 판단 결과에 대한 이용자의 자율적 수용 또는 거부 의사결정 과정에 직접 영향을 미치므로, AI 추천 결과의 편향된 기본 선택(Default) 설정, 동의·거부 버튼 간 크기·색상·위치의 현저한 불균형, 특정 결과를 유도하는 앵커링(Anchoring) 효과 활용 설계, AI 한계·오류 가능성 고지를 인지하기 어려운 시각적 처리(매우 작은 글씨·낮은 대비 색상·화면 하단 배치 등), AI 추천 거부·이의 제기 경로의 의도적 불편 설계가 없는지를 확인한다.

이 점검은 UI의 심미적 완성도나 색상 디자인의 창의성을 평가하는 것이 아니라, 이용자의 자율적 판단권을 보호하기 위한 공정한 UI 설계의 완결성을 확인하는 것을 목적으로 한다.

2. 필요성

▶ 심리적 권위 수용에 따른 이중 침해: AI 서비스에서 이용자는 AI가 제시하는 추천·판단·분류 결과에 대해 "AI가 분석했으니 맞겠지"라는 심리적 권위 수용(Authority Bias) 경향을 가지며, 이 상태에서 UI가 추가적으로 특정 선택을 유도하는 다크패턴을 적용하면 이용자는 이중으로 판단 자율성을 침해받는다.

▶ 중요 의사결정 검토 기회 차단: AI 서비스의 출력이 이용자의 중요한 의사결정(복지 수급 여부·업무 처리 방향·진단 결과 수용 등)에 활용되는 경우, 다크패턴이 포함된 UI 설계는 이용자가 AI 결과를 검토하고 이의를 제기할 기회를 실질적으로 차단한다.

▶ 비의도적 다크패턴 발생 가능성: 다크패턴은 의도적으로 설계되는 경우도 있지만 설계자가 인식하지 못한 채 관행적 UI 관례를 적용하다가 발생하는 경우도 많다. AI 추천 결과를 가장 눈에 잘 띄는 위치에 배치하고 나머지 선택지를 작게 표시하는 설계는 일반 UI 원칙처럼 보이지만, AI 서비스 맥락에서는 이용자가 AI 추천 외의 선택지를 고려할 기회를 줄이는 편향 유발 설계가 된다.

▶ 규제 대응 및 설계 단계 사전 차단의 효율성: 다크패턴은 이용자가 AI 시스템의 권고를 과신하거나 불리한 조건에 동의하도록 UI를 설계하는 방식으로, 관련 규제(EU AI Act 제5조 금지 행위 등)에서 명시적으로 문제시하고 있다. {화면설계서}에 이러한 설계가 포함되면 개발 이후 수정은 화면 전반의 재설계를 수반하므로, 설계 검토 단계에서의 사전 차단이 가장 효율적이다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추천·판단 결과의 편향된 기본값(Default) 설정 설계 여부	{화면설계서}의 AI 추천 결과 표시 화면 목업에서 기본 선택(체크박스 사전 선택·라디오 버튼 기본값·자동 적용 여부)과 선택지 간 시각적 균형 항목 육안 확인
2	AI 서비스 동의·거부 버튼 간 시각적 균형 설계 여부	{화면설계서}의 AI 서비스 동의 화면 목업에서 동의 버튼과 거부/비동의 버튼의 크기·색상·배치 위치 비교 육안 확인, 문구의 감정 유발 요소(Confirmshaming) 확인
3	AI 서비스 한계·오류 가능성 고지의 가시성 설계 여부	{화면설계서}에서 AI 서비스 한계·오류 가능성 고지 문구의 글씨 크기·색상 대비·화면 내 위치(본문 영역 내 vs 하단 각주 영역) 항목 육안 확인
4	AI 추천 결과 앵커링 효과 배제 설계 여부	{화면설계서}의 AI 분석 결과 표시 화면 목업에서 AI 수치·점수·추천 결과의 배치 크기·위치와 관련 맥락 정보(근거·불확실성 표시) 균형 항목 육안 확인
5	AI 추천 거부 및 AI 결정 이의 제기 경로의 접근성 설계 여부	{화면설계서}의 AI 추천 결과 화면과 AI 결정 통보 화면 목업에서 거부/이의제기 버튼 또는 경로의 존재 여부, 위치, 접근 단계 수 항목 육안 확인
6	긴급 고지·경고 팝업의 의도적 방해 설계 부재 여부	{화면설계서}의 AI 서비스 주요 안내 팝업 목업에서 팝업 크기·고지 내용의 가시성·닫기 버튼 크기와 위치(고지 내용 확인 전 닫기 가능 여부) 항목 육안 확인
7	{요구사항정의서} ↔ {화면설계서} 공정한 UI 및 이용자 자율 선택 요건 추적성 확보 여부	{요구사항정의서}의 AI 공정 인터페이스·이용자 선택권·이의 제기 경로 요건과 {화면설계서}의 관련 화면 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 추천 기본값 편향, 동의/거부 버튼 시각적 불균형, 한계 고지 가시성 미흡, 앵커링 효과의 과도한 강화, 거부·이의 제기 경로 접근성 불균형, 팝업 방해 설계가 {화면설계서}에서 확인되지 않고, {요구사항정의서}와의 요구사항 추적성이 확보된 경우	현행 유지
보완권고 (△)	명백한 다크패턴 설계 구조는 아니나 AI 한계 고지의 가시성 미흡, 거부·이의 제기 경로의 접근 단계 과다, 선택지 간 시각적 균형 부족, 요구사항 추적성 일부 누락 등 이용자 판단 자율성을 침해할 수 있는 설계 요소가 일부 확인된 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	AI 추천 결과가 이용자의 명시적 선택 없이 기본값으로 설정된 설계, 동의/거부 버튼 간 현저한 시각적 불균형, 실질적으로 읽히지 않는 AI 한계 고지, AI 추천 거부 경로가 수용 경로보다 현저히 불편한 설계, 중요 팝업 닫기를 유도하는 방해 설계 중 하나 이상이 {화면설계서}에서 확인된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 심미성과 공정성의 평가 기준 구분: UI 설계의 심미적 완성도와 공정성은 다른 평가 기준이며, 구체적인 색상 값이나 디자인 스타일을 요구하는 것은 감리 범위를 벗어난다. 감리원이 집중해야 하는 것은 "AI 추천 선택지와 다른 선택지 사이에 이용자의 판단을 구조적으로 편향시킬 만큼 현저한 시각적 불균형이 있는가"이며, "현저한"의 기준은 이용자가 화면을 보는 순간 다른 선택지가 사실상 보이지 않거나 클릭하기 불편한 구조인 경우로 한정하여 적용한다.

2) 중립적 표현으로 결과 기술: 다크패턴은 설계자가 의도하지 않은 채 관행적 UI 관례를 적용하다가 발생하는 경우가 많으므로, 감리 결과서에 "의도적인 이용자 기만"이라는 표현 대신 "이용자의 자율적 판단권을 침해할 수 있는 설계 요소가 발견되었으며 개선이 필요합니다"와 같은 중립적이고 구체적인 표현을 사용한다.

3) 한계 고지 가시성의 실용적 판정 기준: AI 서비스 한계 고지 가시성은 "{화면설계서} 목업에서 육안으로 읽히는가"를 기준으로 판단하되, 웹 콘텐츠 접근성 지침(WCAG)의 4.5:1 색상 대비 권고 기준을 참조 근거로 활용한다. 모바일 환경에서는 AI 추천을 거부·수정할 수 있는 버튼의 터치 타겟(Touch Target) 크기가 접근성 지침을 충족하는지 모바일 뷰 목업으로 추가 점검한다.

4) 강조 설계와 억제 설계의 구분: AI 추천 결과를 시각적으로 강조하는 것 자체는 정당한 설계 목적이 있으며, 문제는 이 강조가 다른 선택지를 보이지 않게 만드는 방식으로 이루어지는 경우이다. 점검 관점을 "다른 선택지가 이용자에게 동등하게 인식될 수 있는가"로 설정하면 더 명확한 판정이 가능하다.

5) 공공 AI 서비스의 이의 제기 경로 필수 확인: AI 결정이 행정 처분의 근거로 활용되는 공공 AI 서비스에서 이의 제기 경로 부재는 행정법상 이의신청·행정심판 권리 행사를 실질적으로 방해하는 문제로 발전할 수 있으므로, {화면설계서}에 AI 결정 결과 통보 화면과 이의 제기 경로 화면이 별도로 설계되어 있는지를 필수 확인 항목으로 점검한다.

AG_SD42-2-02 | 편향모니터링 지표 산식설계

응용시스템 | 설계 | 인공지능 모델 개발 | 편향/공정성

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

AI 서비스 구축 사업에서, 편향(Bias) 모니터링에 필요한 정량적 지표와 그 산식이 {화면설계서}에 구체적으로 반영되었는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 편향 감지에 활용되는 지표(예: 통계적 동등성, 균등화 교차 확률 등)의 산식 정의가 설계서에 명시되었는지, ② 대시보드 화면 내에서 해당 지표가 어떤 형태(수치, 차트, 알림 등)로 출력되도록 설계되었는지, ③ 지표 임계값(Threshold) 및 경보 조건이 사전에 정의되어 화면 로직에 반영되었는지 여부를 확인한다.

이를 통해 운영 단계에서 편향을 탐지·조치할 수 있는 기반이 설계 단계에서부터 갖추어졌음을 검증하는 것이 이 가이드라인의 핵심 목적이다.

2. 필요성

▶ AI 시스템의 편향 문제는 서비스 운영 중에 발현되지만, 그 탐지와 시각화를 위한 로직은 반드시 화면설계서에 반영되어야 한다. 편향 모니터링 지표와 산식이 설계 단계에서 정의되지 않으면, 개발자가 임의로 지표를 구현하거나 대시보드가 단순 통계 조회 화면에 그쳐, 실질적인 편향 감시 기능이 누락된 채 시스템이 구축될 위험이 있다.

▶ 또한 {화면설계서}에 편향 지표의 산식과 임계값이 문서화되어 있어야 검수 기준으로 활용할 수 있으며, 사후 감사(Audit) 시에도 설계 근거 자료로 제출할 수 있다. 감리원이 이 단계에서 산식의 정의 여부와 화면 로직 반영 여부를 제3자 관점으로 점검함으로써, 설계 미비로 인한 품질 리스크를 사전에 차단할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	편향 모니터링 지표 항목 명시 여부	{화면설계서} 내 대시보드 화면 목록 및 항목 검토 (세부: 성별·연령·지역 등 집단별 비교 지표 포함 여부)
2	편향 지표 산식(계산 공식) 정의 여부	{화면설계서} 또는 'AI 공정성 요건 정의서' 내 산식 정의 기술 여부 확인 (세부: 예: 통계적 동등성, 균등화 교차 확률, 예측 균등성 등)
3	임계값(Threshold) 및 경보(Alert) 조건 설계 반영 여부	{화면설계서} 내 알림·경보 조건 설계 항목 검토
4	대시보드 출력 형태(시각화 방식) 설계 적정성	{화면설계서} 내 대시보드 UI 와이어프레임 또는 목업 검토 (세부: 수치 표시, 그래프 유형, 색상 경보 등)
5	{화면설계서} ↔ {AI모델설계서} 간 편향 지표 추적성 확보 여부	{AI모델설계서}의 공정성 설계 항목과 {화면설계서}의 지표 항목 교차 검토

No	점검 항목	확인 방법
6	집단 정의(Cohort) 및 비교 기준 설계 명시 여부	{화면설계서} 또는 'AI 공정성 요건 정의서' 내 집단 정의 기술 여부 확인 (세부: 비교 집단의 분류 기준, 기준 집단 지정 여부)
7	편향 지표 산출 주기 및 데이터 범위 설계 여부	{화면설계서} 내 데이터 갱신 주기 및 산출 범위 기술 여부 확인 (세부: 실시간·일별·주별 산출, 적용 데이터 기간 등)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	편향 모니터링 지표와 산식, 임계값·경보 조건, 대시보드 출력 형태, 집단 정의 및 산출 주기가 {화면설계서}에 구체적으로 설계되고 {AI모델설계서}와의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	편향 모니터링 화면은 설계되어 있으나 산식 미정의, 임계값·경보 조건 누락, 집단 정의 불명확 등 일부 항목이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향 모니터링 지표·산식 설계가 {화면설계서}에 반영되어 있지 않거나 단순 통계 조회 화면에 그쳐 편향 감시 기능을 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 산식의 수학적 타당성은 감리 범위 밖임을 인식할 것: 통계적 동등성 등 편향 지표 산식 자체의 학술적 타당성 평가는 데이터 과학 전문가의 영역이다. 감리원은 산식이 설계서에 정의되어 있는지, 임계값·경보 조건·산출 주기가 명시되어 있는지 등 설계 문서의 완결성만 점검하고, 필요 시 전문가 자문을 감리 의견에 병기한다.

2) 편향 대시보드가 실제 모니터링 기능인지 확인할 것: 일부 사업에서는 {화면설계서}에 "편향 모니터링 화면"이 포함되어 있으나, 실제로는 단순 데이터 조회 화면에 불과하고 임계값 기반 경보 로직이 없는 경우가 있다. 화면 제목에 "모니터링"이 포함된다고 해서 자동으로 적합 판정하지 않고, 반드시 알림 조건과 대응 로직 설계 여부를 확인해야 한다.

3) 개인정보 포함 여부를 집단 선제 확인할 것: 성별·연령·지역 등 비교 집단(Cohort) 정의에는 개인정보 또는 민감정보의 처리가 수반될 수 있다. 집단 분류에 사용되는 데이터 항목의 수집 근거와 비식별화 처리 방식이 설계에 반영되어 있는지 확인하고, 미반영 시 개인정보 보호법 준수 관점에서 보완을 권고한다.

AG_SD42-2-03 | 편향유발 입력 사전차단 로직설계

응용시스템 | 설계 | 인공지능 모델 개발 | 편향/공정성

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

AI 서비스 구축 사업에서, AI 모델에 입력되는 질의 중 특정 집단에 대한 편향된 응답을 유발할 가능성이 높은 입력(이하 "편향 유발 입력")을 AI 모델에 전달하기 이전 단계에서 탐지·차단하는 필터링 로직이 {응용시스템설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 편향 유발 입력의 유형 및 차단 대상 금칙어·금지 표현의 범주 정의, ② 입력 수신부터 AI 모델 전달 또는 차단 처리까지의 필터링 처리 흐름도, ③ 금칙어 목록의 관리 체계(버전 관리·갱신 절차·승인 방법), ④ 일반 악성 프롬프트 필터링과 별개의 독립적인 설계 항목으로 구분되어 있는지를 확인한다.

TTA 개발안내서와 NIST AI RMF 3.2절(Bias Management)은 AI 시스템이 생애주기 전반에 걸쳐 편향을 식별·관리할 것을 요구한다.

2. 필요성

▶ AI 모델은 학습 데이터에 내재된 사회적 편견을 반영하여 특정 인종·성별·나이·지역·종교 등의 집단에 대해 차별적이거나 편향된 응답을 생성하는 경향이 있다. 특히 이용자가 의도적으로 또는 무의식적으로 특정 집단에 대한 고정관념을 내포한 질의를 입력할 경우, AI 모델은 그 편향을 증폭시켜 응답하는 경향이 있어 서비스 차원의 신뢰도 저하와 사회적 논란으로 이어질 수 있다.

▶ 설계 단계에서 편향 유발 입력 차단 로직이 {응용시스템설계서}에 구체화되지 않으면, 개발자는 악성 프롬프트 차단 로직(욕설·음란물·해킹 시도 등)만 구현하고 편향 관련 입력 차단은 설계 근거가 없어 생략하거나, 차단 대상 범주가 체계적으로 정의되지 않아 특정 문화권·맥락의 편향 유발 표현이 차단 범위에서 누락될 위험이 있다.

▶ 또한 편향 유발 금칙어 목록은 사회적 논의와 법적 기준의 변화에 따라 지속적으로 갱신되어야 하므로, 갱신 절차가 설계 단계에서 정의되지 않으면 운영 중 목록의 유효성이 급격히 저하된다. 이 항목은 악성 프롬프트 필터링과 기술적 구현 방식이 유사하지만 차단 목적과 대상 분류 체계가 다르다는 점에서 독립적인 설계 항목으로 점검할 필요가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	편향 유발 입력 유형 및 금칙어 범주 정의 여부	{응용시스템설계서} 또는 '편향 유발 금칙어 목록 및 필터링 정책서'에서 편향 유형 분류 및 범주별 금칙어 설계 항목 확인, 담당자 인터뷰
2	악성 프롬프트 필터링과의 독립적 설계 구분 여부	{응용시스템설계서} 내 입력 필터링 관련 섹션에서 악성 프롬프트 필터링과 편향 유발 입력 차단의 설계 항목 분리 여부 확인
3	금칙어 기반 필터링 처리 흐름도 설계 여부	{응용시스템설계서} 내 편향 입력 차단 처리 흐름도 항목 확인 및 흐름 단계별 분기 조건 기술 여부 확인
4	편향 유발 입력 탐지 방식의 범주별 적용 설계 여부	{응용시스템설계서} 내 탐지 방식 설계 항목에서 범주별 탐지 수준(단순 키워드 매칭·패턴 매칭·문맥 분석 등) 기술 여부 확인
5	차단 후 이용자 안내 메시지 및 대안 안내 설계 여부	{응용시스템설계서} 내 차단 발동 시 이용자 안내 메시지 설계 항목 및 {화면설계서} 연계 여부 확인
6	금칙어 목록 버전 관리 및 갱신 절차 설계 여부	{응용시스템설계서} 또는 '편향 유발 금칙어 목록 및 필터링 정책서'에서 목록 갱신 절차·버전 관리 방법 항목 확인
7	{요구사항정의서} ↔ {응용시스템설계서} 편향 방지 요건 추적성 확보 여부	{요구사항정의서} 공정성·편향 방지 관련 요건과 {응용시스템설계서} 편향 입력 차단 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	편향 유형·금칙어 범주 정의, 처리 흐름도, 탐지 방식, 이용자 안내, 목록 버전 관리·갱신 절차가 {응용시스템설계서}에 악성 프롬프트 필터링과 구분되어 설계되고 요구사항과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	편향 입력 차단 설계는 존재하나 금칙어 범주 체계 불완전, 갱신 절차 미정의, 이용자 안내 설계 누락 등 일부 항목이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향 유발 입력 차단 로직이 설계에 반영되어 있지 않거나 악성 프롬프트 필터링에 혼재되어 독립적인 편향 방지 설계를 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 악성 프롬프트 필터링 가이드라인과의 점검 중복을 방지하되, 두 항목의 목적 차이를 명확히 구분할 것: 악성 프롬프트 필터링(보안 목적)과 편향 유발 입력 차단(공정성 목적)은 감리 시 별개의 점검 항목으로 다루어야 한다. 사업자가 두 기능을 통합 구현하는 경우에도 설계서에서 각 목적별 금칙어 범주가 명확히 구분되어 있는지를 확인한다. "이 항목은 보안 필터링 가이드라인에서 이미 확인했으므로 생략"하는 것은 편향 방지 설계의 독립성을 무시하는 오류이다.

2) 금칙어 목록의 내용 자체를 감리원이 직접 평가하려는 시도를 자제할 것: "이 금칙어 목록이 충분한가", "이 표현이 편향 유발에 해당하는가"와 같은 판단은 사회학·언어학·법학 전문 지식을 필요로 하며 감리 범위를 벗어난다. 감리원은 "금칙어 범주 분류 체계가 설계서에 존재하는가", "서비스 도메인에 맞는 범주가 포함되어 있는가", "갱신 절차가 설계되어 있는가"만을 확인하고, 내용의 적절성에 의문이 있는 경우 AI 윤리 전문가 자문 필요성을 감리 의견에 병기한다.

3) "차단 기준이 모호하다"는 사업자의 주장을 경계하고, 모호성을 설계로 해결하도록 유도할 것: 기준이 모호할수록 범주 정의·판정 절차·갱신 체계를 설계 문서로 명문화해야 한다. 모호성을 이유로 설계를 생략하는 것은 인정하지 않으며, 범주별 예시와 판단 기준을 문서화하고 운영 중 갱신할 수 있는 절차를 갖추도록 권고한다.

AG_SD42-2-04 | 환각발언 예방 파이프라인설계

응용시스템 | 설계 | 인공지능 모델 개발 | 편향/공정성

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

AI 서비스 구축 사업에서, 대규모 언어 모델(LLM) 기반 AI 서비스가 사실과 다르거나 존재하지 않는 정보를 자신 있게 생성하는 환각(Hallucination) 현상을 예방하기 위해, 검색 증강 생성(RAG, Retrieval-Augmented Generation) 기법에 기반한 문서 검색 및 응답 검증 로직이 'AI모델설계서'에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 이용자 질의에 대한 관련 문서를 검색하여 AI 모델의 생성 컨텍스트로 제공하는 문서 검색 로직(Retrieval)이 'AI모델설계서'에 설계되어 있는지, ② AI 모델이 생성한 응답이 검색된 문서에 근거하고 있는지를 검증하는 출력 검증 로직(Output Verification)이 설계에 포함되어 있는지, ③ 관련 문서가 없거나 검증에 실패한 경우 이용자에게 "근거 없음"을 안내하거나 응답 생성을 차단하는 폴백(Fallback) 처리가 설계되어 있는지, ④ RAG 파이프라인에 사용될 지식 베이스의 문서 범위·갱신 주기·품질 기준이 설계 단계에서 정의되어 있는지를 확인한다.

이 점검은 기술적 구현 방식의 타당성 평가가 아니라, 환각 예방에 필요한 설계 요소들이 'AI모델설계서'에 빠짐없이 포함되어 있는지의 완결성을 확인하는 것을 목적으로 한다.

2. 필요성

▶ LLM은 방대한 텍스트 데이터로 학습된 언어 패턴을 기반으로 응답을 생성하므로 특정 사실에 대한 정확한 정보를 항상 보유하고 있지 않으며, 학습 데이터에 포함되지 않은 최신 정보, 기관 특유의 내부 정책·규정, 매우 구체적인 수치가 필요한 질의에서 그럴듯하지만 사실이 아닌 내용을 자신 있는 어조로 생성하는 환각 현상을 보인다.

▶ 공공 AI 서비스에서 환각이 발생하면 이용자가 잘못된 정책 정보나 법령 해석을 신뢰하여 행정적 불이익을 받거나 법적 분쟁의 소지가 생기는 심각한 문제로 이어진다.

▶ RAG 기법은 신뢰할 수 있는 문서 저장소를 검색하여 관련 문서를 AI 모델의 응답 생성 컨텍스트로 제공함으로써 환각 가능성을 구조적으로 줄이는 가장 검증된 방법이나, RAG를 도입한다고 해서 환각이 완전히 사라지지는 않으므로 AI 모델이 생성한 응답이 실제로 검색된 문서에 근거하고 있는지를 검증하는 출력 검증 로직이 함께 설계되어야 한다.

▶ 설계 단계에서 'AI모델설계서'에 RAG 파이프라인과 출력 검증 로직이 포함되어 있지 않으면 개발 단계에서 환각 예방 기능이 누락되거나, 검색 기능만 있고 검증 로직은 없는 불완전한 구조로 구현되는 위험이 있다.

▶ 환각 예방 파이프라인이 없는 LLM 서비스는 사실과 다른 정보를 신뢰도 있는 형식으로 제공하여 이용자 피해·기관 신뢰도 손상·법적 분쟁을 유발하므로, RAG·출력 검증·신뢰도 임계값 등 다층 방어 체계는 결정해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	RAG 파이프라인 전체 흐름도 설계 여부	'AI모델설계서'에서 RAG 파이프라인 처리 흐름도 항목 확인. 6단계 각각이 흐름도에 포함되어 있는지 단계별 확인, 담당자 인터뷰
2	문서 검색 방식 및 검색 대상 지식 베이스 설계 여부	'AI모델설계서' 또는 '지식 베이스 구축 계획서'에서 검색 방식·대상 문서 범위·갱신 주기 항목 확인
3	검색 결과의 관련성 임계값(Threshold) 및 컨텍스트 구성 방식 설계 여부	'AI모델설계서'에서 검색 결과 관련성 임계값 설정값, 컨텍스트에 포함할 문서 수(Top-K) 기준 항목 확인
4	AI 모델 출력의 근거 문서 검증 로직(Output Verification) 설계 여부	'AI모델설계서'에서 AI 응답 생성 이후 단계의 출력 검증 로직 항목 확인. 검증 방법·임계값·검증 실패 판정 기준 포함 여부 확인
5	환각 의심 응답 및 검증 실패 시 폴백(Fallback) 처리 흐름 설계 여부	'AI모델설계서' 및 {응용시스템설계서}에서 검증 실패 시 처리 흐름(이용자 안내 문구·응답 차단·관리자 알림) 항목 확인
6	지식 베이스 문서 품질 기준 및 버전 관리 설계 여부	'지식 베이스 구축 계획서' 또는 'AI모델설계서'에서 문서 등재 기준(출처·최신성·정확성)과 문서 버전 관리 방식 항목 확인
7	{요구사항정의서} ↔ 'AI모델설계서' 환각 예방 요건 추적성 확보 여부	{요구사항정의서} 환각 예방·응답 정확도·근거 제시 관련 요건과 'AI모델설계서' RAG 파이프라인 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	RAG 파이프라인 6단계 전체 흐름도가 설계되어 있고, 문서 검색 방식·관련성 임계값·Top-K 기준, 출력 검증 로직, 두 가지 시나리오(관련 문서 없음·검증 실패)에 대한 폴백 처리, 지식 베이스 품질 기준·버전 관리 프로세스, 요구사항 추적성이 모두 확인된 경우	현행 유지
보완권고 (△)	RAG 파이프라인 전체 흐름도와 주요 설계 요소는 존재하나, 출력 검증 통과 기준이 측정 가능한 방식으로 정의되지 않았거나 지식 베이스 갱신 주기·폴백 안내 문구 등 일부 설계 항목이 구체화되지 않은 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	RAG 파이프라인 흐름도가 없는 경우, 문서 검색 기능만 설계되고 출력 검증 로직이 없는 경우, 관련성 임계값이 수치로 정의되지 않은 경우, 폴백 처리 흐름이 없는 경우, 지식 베이스 등재 기준이나 갱신 프로세스가 없는 경우, 요구사항의 환각 예방 요건이 RAG 설계에 반영되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "RAG를 도입했다"는 진술만으로 환각 예방 설계 완료로 판단하지 말 것: RAG는 환각 가능성을 구조적으로 줄이는 방법이지 완전히 차단하는 방법이 아니므로, 사업자가 "RAG를 적용했으므로 환각 문제는 해결됩니다"라고 답변하더라도 출력 검증 로직·폴백 처리·지식 베이스 품질 기준이 'AI모델설계서'에 포함되어 있는지를 별도로 확인한다. 특히 출력 부재가 가장 빈번한 설계 미비이다.

2) 'AI모델설계서'가 없는 사업은 {응용시스템설계서} 내 AI 설계 섹션을 대체 확인하되 설계 누락과 문서 분류 문제를 명확히 구분할 것: {응용시스템설계서}에도 RAG 파이프라인 설계가 전혀 없는 경우는 "'AI모델설계서'가 없음"이 아니라 "RAG 파이프라인 설계 자체가 어느 문서에도 존재하지 않음"으로 판정하여, 단순한 문서 분류 문제가 아닌 설계 누락으로 감리 의견을 기록한다.

3) 지식 베이스에 어떤 문서가 포함되는지와 문서 범위가 서비스 목적과 일치하는지 점검할 것: 설계서에 등재 문서 목록이 명시되어 있더라도 그 문서들이 서비스가 대답해야 하는 질의 범위를 실제로 커버하는지 담당자 인터뷰로 확인한다. 예를 들어 복지 서비스 안내 AI에 홍보용 브로셔만 등재되어 있으면 수급 자격 기준 등 구체적 질의에서 환각 발생 위험이 높다.

4) 출력 검증 로직은 설계 존재 여부와 검증 기준의 측정 가능성에 집중할 것: 검증 방법(의미 유사도·인용 확인·검증 모델 등)의 기술적 우열 평가는 감리 범위를 벗어나며, 출력 파이프라인 설계에 존재하는지와 통과·실패 기준이 측정 가능한 방식(예: 유사도 0.65 이상, 출처 인용 1건 이상)으로 정의되어 있는지를 확인한다. 아울러 채택된 문서 검색 알고리즘(Keyword·Vector·Hybrid Search 등)이 요구사항에 명시된 응답 정확도 목표 달성에 논리적으로 타당한 설계인지 검토 회의 등을 통해 적정성을 점검한다.

5) 환각 예방이 요구되는 AI 기능 범위를 먼저 특정할 것: 이미지 분류처럼 환각이 발생하지 않는 기능과 사실 정보를 생성하는 챗봇 기능이 한 서비스에 혼재할 수 있으므로, AI가 이용자에게 텍스트 형식으로 사실 정보를 직접 생성·제공하는 기능 범위를 먼저 특정한 후 그 범위 안에서 RAG 파이프라인 설계 여부를 확인한다.

AG_SD42-2-05 | AI결정 오버라이드 기능설계

응용시스템 | 설계 | 인공지능 모델 개발 | AI윤리/xAI

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

AI 서비스 구축 사업에서, AI 시스템이 생성한 출력 결과를 관리자 또는 권한 있는 담당자가 수동으로 수정하거나 대체하는 오버라이드(Override) 기능이 {화면설계서} 및 {응용시스템설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 오버라이드 기능에 접근할 수 있는 권한 체계와 접근 화면이 {화면설계서}에 설계되어 있는지, ② AI 출력 화면에서 수정 가능한 항목의 범위와 수정 방식(직접 편집, 선택지 교체 등)이 목업으로 설계되어 있는지, ③ 오버라이드 실행 시 수정자·수정 일시·수정 전후 내용이 포함된 감사 추적(Audit Trail) 로그가 저장되도록 설계되어 있는지, ④ 오버라이드 이력이 AI 모델 개선 피드백으로 연계되는 절차가 설계서 수준에서 정의되어 있는지를 확인한다.

NIST AI RMF 3.1절(Accountable)과 TTA 개발안내서는 AI 시스템의 결정이 이해관계자에게 중요한 영향을 미칠 때 사람이 개입하여 수정·통제할 수 있는 체계를 갖출 것을 요구한다.

2. 필요성

▶ AI 시스템은 학습 데이터의 한계, 입력 맥락의 복잡성, 도메인 특수성 등 여러 요인으로 인해 실제 업무 환경에서 오류가 있거나 맥락에 부적합한 출력을 생성하는 경우가 발생한다. 이러한 상황에서 담당자가 AI 출력을 그대로 이용자에게 전달하거나 행정 처리에 적용할 수밖에 없는 구조라면, AI의 오류가 실제 업무상 피해로 직결된다. 따라서 권한 있는 담당자가 AI 출력을 수정·대체할 수 있는 오버라이드 기능과 그 감사 추적 체계가 구체화되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	오버라이드 권한 체계 및 접근 화면 설계 여부	{화면설계서} 내 관리자 오버라이드 접근 화면 목업 확인, {응용시스템설계서} 권한 체계 항목 확인, 담당자 인터뷰
2	AI 출력 수정 가능 항목 범위 및 편집 UI 설계 여부	{화면설계서} 내 AI 출력 수정 화면 목업에서 편집 가능 영역, 수정 방식, AI 원본 출력 표시 항목 확인
3	오버라이드 이유(사유) 입력 설계 여부	{화면설계서} 내 오버라이드 수정 화면에서 사유 입력 필드 또는 사유 선택 UI 항목 확인 (세부: 수정 사유 필수 입력 또는 선택 입력 설계, 사유 분류 체계 포함)

No	점검 항목	확인 방법
4	오버라이드 실행 시 감사 추적(Audit Trail) 로그 설계 여부	{응용시스템설계서} 내 감사 추적 로그 항목 구성 설계 및 로그 저장 방식 확인 (세부: 수정자·수정 일시·AI 원본 출력·수정 후 출력·수정 사유가 모두 포함된 로그 구조 설계)
5	오버라이드 확정 전 검토·승인 단계 설계 여부	{화면설계서} 내 수정 확정 흐름(즉시 반영 또는 승인 대기) 및 승인 화면 목업 확인 (세부: 수정 완료 즉시 반영 vs. 2차 승인자 검토 후 반영 분기 설계)
6	오버라이드 이력 조회 화면 설계 여부	{화면설계서} 내 오버라이드 이력 조회 화면 목업, 조회 조건(기간·수정자·사유 유형 등) 필터 항목 확인
7	{요구사항정의서} ↔ {화면설계서} ↔ {응용시스템설계서} 오버라이드 기능 요건 3방향 추적성 확보 여부	{요구사항정의서} 관련 요건, {화면설계서} 오버라이드 UI, {응용시스템설계서} 처리 흐름을 스프레드시트로 3방향 매핑하여 누락·불일치 여부 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	오버라이드 권한 체계·수정 범위·사유 입력·감사 추적 로그·승인 단계·이력 조회 화면이 {화면설계서}와 {응용시스템설계서}에 모두 설계되고, 요구사항과의 3방향 추적성이 확보된 경우	현행 유지
보완권고 (△)	오버라이드 기능은 설계되어 있으나 감사 추적 로그 항목·사유 입력·이력 조회 화면 등 일부 요소가 누락되었거나, 화면 설계와 시스템 설계 간 불일치가 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 출력에 대한 오버라이드 기능 설계 자체가 부재하거나, 수정 기능은 있으나 감사 추적 로그 설계가 전혀 없어 수정 이력 추적이 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 오버라이드 기능과 감사 추적 로그는 반드시 "쌍(Pair)"으로 설계되어 있는지 확인할 것: 현장에서 자주 발생하는 패턴은 {화면설계서}에 수정 편집 UI는 설계되어 있으나 {응용시스템설계서}에 감사 추적 로그 설계가 없거나, 반대로 시스템 설계에 로그 저장은 있으나 화면 설계에 이력 조회 화면이 없는 경우이다. 오버라이드 기능의 존재 의의는 "수정의 자유"와 "수정의 투명한 기록"이 동시에 확보되는 데 있으므로, 두 가지가 반드시 쌍으로 설계되어 있는지 확인해야 한다.

2) 서비스 위험도에 따라 2차 승인 단계 설계의 필요성을 차등 판단하여 의견을 제시할 것: 행정 처분·복지 급여 판정·의료 보조 등 이용자에게 중요한 영향을 미치는 고위험 AI 서비스에서 담당자 단독 즉시 수정 반영 구조는 오버라이드 권한 남용이나 실수로 인한 피해를 초래할 수 있으므로

2차 승인자 검토 단계 설계를 강하게 권고한다. 반면 AI가 초안 문서 작성을 보조하는 수준의 저위험 서비스에서는 즉시 반영 구조가 업무 효율 측면에서 적절할 수 있다.

3) AI 원본 출력이 반드시 감사 로그에 포함되어야 함을 사업자에게 명확히 안내할 것

AG_SD42-2-06 | AI한계점 안내 프론트엔드 설계

응용시스템 | 설계 | 시스템 구현 | AI윤리/xAI

근거 | NIST AI RMF · 「지능정보화 기본법」 · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

AI 서비스 구축 사업에서, AI 시스템 결과의 불확실성·오류 가능성·적용 한계 등 이용자가 반드시 인지해야 할 정보가 {화면설계서}의 화면 목업에 경고 문구(Warning Message) 또는 고지 UI 형태로 구체적으로 반영되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① AI 결과가 참고용임을 명시하는 면책 고지 문구가 결과 화면 어느 위치에 어떤 형태로 표시되는지, ② 모델의 예측 정확도 한계·학습 데이터 기간·적용 제외 사례 등 AI 고유의 한계 정보가 이용자에게 제공되는 화면이 설계되어 있는지, ③ 불확실성 수준이 높은 결과(저신뢰도 출력)에 대해 해당 결과에 시각적으로 연동된 경고 UI가 별도 설계되어 있는지를 확인한다.

NIST AI RMF는 투명성(Transparency)의 핵심 요소로 "AI 시스템이 무엇을 할 수 있고 무엇을 할 수 없는지를 이용자가 이해할 수 있도록 정보를 제공해야 한다"고 명시하고 있으며, 이 가이드라인은 그 요구사항이 설계 단계에서 화면에 반영되었는지를 검증하는 근거가 된다.

2. 필요성

▶ AI 시스템은 학습 데이터의 시간적·지역적 범위를 벗어난 입력에 대해 높은 신뢰도 점수를 출력하면서도 실제로는 오류를 낼 수 있다. 이 특성을 이용자가 알지 못한 채 AI 결과를 확정적 사실로 받아들이면, 의료·법률·금융 등 고위험 도메인에서 심각한 피해로 직결된다. NIST AI RMF는 투명성(Transparency)을 신뢰할 수 있는 AI의 핵심 특성으로 요구한다.

▶ AI 한계점 고지 의무는「지능정보화 기본법」및 관계 부처 가이드라인에서 요구하는 사항이다. 설계 단계에서 고지 문구·위치·표시 조건이 명확히 정의되지 않으면 구현 단계에서 누락되거나 형식적 텍스트로 처리될 위험이 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 시스템 결과 화면 내 면책 고지 문구 설계 여부	{화면설계서} 내 AI 결과 화면 목업 검토, 담당자 인터뷰 (세부: AI 결과가 참고용이며 최종 결정은 이용자 책임임을 명시)
2	AI 모델 한계 정보 제공 화면 설계 여부	{화면설계서} 내 한계 정보 제공 화면 또는 팝업 설계 항목 검토 (세부: 학습 데이터 기간, 정확도 범위, 적용 제외 케이스 등)

No	점검 항목	확인 방법
3	불확실성 수준별 경고 UI의 시각적 차등 설계 여부	{화면설계서} 내 신뢰도 수준별 UI 표현 방식 설계 항목 검토 (세부: 저신뢰도 결과에 아이콘·색상·문구를 결과에 직접 연동)
4	이용자 맥락별 고지 문구 차등 설계 여부	{화면설계서} 내 화면 흐름도 및 단계별 고지 방식 기술 여부 확인 (세부: 최초 진입 시 전문 고지 / 결과 조회 시 요약 고지 구분)
5	고지 문구의 가독성 설계 적정성	{화면설계서} 내 고지 문구 디자인 스펙(폰트 크기, 색상코드, 배치 위치) 항목 검토 (세부: 글꼴 크기, 색상 대비, 화면 내 노출 위치의 적정성)
6	이용자 확인 인터랙션 설계 여부	{화면설계서} 내 고위험 기능 결과 화면의 이용자 확인 UI 설계 항목 검토 (세부: 고지 내용 확인 체크박스·동의 버튼 등)
7	{요구사항정의서} ↔ {화면설계서} AI 한계 고지 항목 추적성 확보 여부	{요구사항정의서} 고지 요건, {AI모델설계서} 한계 정의, {화면설계서} 고지 문구의 3방향 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	면책 고지 문구, 모델 한계 정보 화면, 신뢰도 수준별 경고 UI, 이용자 맥락별 고지 문구 차등, 가독성 스펙, 이용자 확인 인터랙션이 {화면설계서}에 구체적으로 설계되고 {AI모델설계서}의 한계 내용과 정합성이 확보되어 있는 경우	현행 유지
보완권고 (△)	고지 문구는 설계되어 있으나 표시 위치·가독성 스펙 미기재, 저신뢰도 경고 UI 미설계, 고지 내용과 모델 한계 정보의 불일치 등 일부 항목이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 한계점 고지 관련 설계가 {화면설계서}에 반영되어 있지 않아 이용자 보호 및 투명성 요건을 충족할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 고지 문구의 "존재"와 "실효성"을 함께 점검할 것: 고지 문구가 화면에 존재하더라도 글꼴 크기 8pt, 연한 회색 텍스트로 화면 최하단에 배치된 경우는 이용자가 실제로 읽지 않는다. {화면설계서}에 고지 문구의 글꼴 크기와 색상 코드가 명시되어 있는지 확인하고, 주요 결과 텍스트 대비 현저히 작거나 눈에 띄지 않게 설계된 경우 보완을 요구한다. "존재"와 "가시성"은 별도 기준으로 판정한다.

2) 생성형 AI(LLM) 서비스의 환각(Hallucination) 고지를 별도로 확인할 것: LLM 기반 서비스에서 가장 중요한 한계 고지는 환각—AI가 사실이 아닌 내용을 자신 있게 출력하는 현상—에 대한 경고이다. {화면설계서}에 "AI가 생성한 내용은 사실과 다를 수 있으며 반드시 원본 자료를 확인하세요"와 같이 환각 위험을 명시하는 고지 문구가 생성 결과 화면에 설계되어 있는지를 별도 항목으로 확인한다.

3) 고지 문구 내용의 정확성보다 {AI모델설계서}와의 정합성을 우선 점검할 것: 화면의 고지 문구가 {AI모델설계서}에 정의된 실제 모델 한계(학습 데이터 기간, 적용 제외 케이스 등)와 일치하는지 교차 확인한다. 설계서의 한계 정의와 무관하게 작성된 일반적·형식적 문구는 보완 대상으로 판정한다.

AG_SD42-2-07 | XAI 설명가능성 적용 설계

응용시스템 | 설계 | 인공지능 모델 개발 | AI윤리/xAI

근거 | NIST AI RMF · 「개인정보 보호법」 · GDPR · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

AI 서비스 구축 사업에서, {요구사항정의서}에 정의된 설명가능성(XAI) 요건이 {AI모델설계서}에 구체적이고 빠짐없이 반영되었는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 서비스 목적과 이용자 유형에 적합한 XAI 기법(SHAP, LIME, Attention 시각화 등)이 선정 근거와 함께 명시되었는지, ② 선정된 XAI 알고리즘의 적용 대상 모델과 출력 형태(특성 중요도 차트, 이유 텍스트, 하이라이팅 등)가 설계서에 상세화되었는지, ③ 설명 결과를 어떤 수준의 이용자(최종 사용자, 운영자, 감사자)에게 어떤 방식으로 제공할지에 대한 설계 방향이 기술되었는지를 확인한다.

이를 통해 개발 착수 전에 XAI 설계 미비로 인한 "블랙박스 AI" 출시 리스크를 사전에 식별하고, 시험·검수 단계에서 활용할 수 있는 구체적인 검증 기준이 마련되었음을 검증한다.

2. 필요성

▶ AI 시스템은 복잡한 수학적 연산을 거쳐 결과를 도출하기 때문에, 별도의 설계 없이는 "왜 이 결과가 나왔는가"를 이용자나 관리자가 이해하기 어렵다. 이러한 불투명성은 공공·금융·의료 분야에서 특히 심각한 신뢰 리스크와 법적 분쟁의 원인이 되며, NIST AI RMF는 설명가능성(Explainability)과 해석가능성(Interpretability)을 신뢰할 수 있는 AI의 핵심 특성으로 규정한다.

▶ 설명가능성(XAI) 요건은 설계 단계에서 구체적 메커니즘(SHAP, LIME, Attention 시각화 등)이 결정되지 않으면 구현 단계에서 형식적 출력에 그치게 된다. AI 결정에 이의를 제기할 이용자 권리(GDPR, 개인정보 보호법)와도 연결되는 핵심 요건이다.

3. 검토방법

No	점검 항목	확인 방법
1	XAI 적용 필요성 검토 및 기법 선정 근거 명시 여부	{AI모델설계서} 및 'AI 위험성 평가서' 검토, 담당자 인터뷰 (세부: 서비스 도메인·위험 수준에 따른 XAI 적용 판단 근거)
2	XAI 알고리즘 유형 선정 및 적용 대상 명시 여부	{AI모델설계서} 또는 'XAI 설계 명세서' 내 기법 선정 항목 검토 (세부: SHAP, LIME, Attention 시각화 등)
3	XAI 출력 형태(설명 결과 표현 방식) 설계 구체성	{AI모델설계서} 내 XAI 출력 설계 항목 검토 (세부: 특성 중요도, 자연어 이유 설명, 시각적 하이라이팅 등)

No	점검 항목	확인 방법
4	이용자 유형별 설명 수준 차등 설계 여부	{AI모델설계서} 내 이용자 역할별 XAI 제공 방식 기술 여부 확인 (세부: 최종 사용자용 / 운영자·감사자용)
5	XAI 설명 범위 및 적용 시점 설계 여부	{AI모델설계서} 내 XAI 적용 범위 및 처리 시점 기술 여부 확인 (세부: 개별 예측 설명 vs. 모델 전체 설명, 실시간 vs. 사후 설명)
6	{요구사항정의서} ↔ {AI모델설계서} XAI 항목 추적성 확보 여부	{요구사항정의서}의 XAI 요건과 {AI모델설계서} 설계 항목 간 매트릭스 교차 검토
7	XAI 성능 저하 및 트레이드오프 대응 설계 명시 여부	{AI모델설계서} 내 XAI 적용에 따른 성능 영향 기술 여부 확인 (세부: 설명가능성과 예측 정확도 간 균형 방안)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	XAI 기법 선정 근거, 적용 대상·출력 형태, 이용자 유형별 설명 수준, 적용 범위·시점, 성능 트레이드오프 대응이 {AI모델설계서}에 구체적으로 설계되고 요구사항과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	XAI 적용 방침은 기술되어 있으나 기법 선정 근거 불명확, 출력 형태·이용자별 제공 방식 미상세 등 설계 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{요구사항정의서}의 XAI 요건이 설계서에 반영되지 않았거나 '추후 개발' 수준의 선언적 기술에 그쳐 설계 반영으로 인정할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) XAI 기법의 기술적 적정성 평가는 감리 범위 밖임을 인식할 것: SHAP 값의 수학적 타당성 등 기법 자체의 검증은 전문가 영역이다. 감리원은 기법 선정 근거·출력 형태·추적성 등 설계 문서의 완결성과 구체성만을 점검하고, 기술적 의문이 있는 경우 전문가 자문 필요성을 감리 의견에 병기한다.

2) 생성형 AI(LLM)에서의 XAI는 전통적 ML과 접근 방식이 다름을 인지할 것: GPT 계열 LLM에는 SHAP·LIME을 직접 적용하기 어렵고, Attention 가중치 시각화, 응답 근거 인용(RAG 기반 출처 명시), 자연어 이유 생성 등이 XAI 대안으로 활용된다. 사업자가 LLM 기반 서비스에 SHAP 적용을 명시한 경우, 실제 구현 가능성에 대해 인터뷰로 확인이 필요하다.

3) "설명가능성 기능 개발 예정"은 설계 미반영으로 처리할 것: 구체적 기법·출력 형태·적용 대상이 설계서에 기술되지 않은 채 "추후 개발" 수준으로만 기재된 경우 설계 반영으로 인정하지 않는다. 검증 가능한 수준의 명세를 요구한다.

AG_SD42-2-08 | 신뢰점수 노출 설계

응용시스템 | 설계 | 인공지능 모델 개발 | AI윤리/xAI

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

AI 서비스 구축 사업에서, AI 모델이 생성하는 추론 결과의 신뢰도 수치(Confidence Score)가 이용자에게 의미 있는 형태로 화면에 표시되도록 {화면설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 AI 추론 결과와 함께 신뢰점수를 표시하는 UI 요소(수치, 게이지 바, 색상 등급 등)가 화면 목업에 명시되어 있는지, 신뢰점수의 표현 형식(백분율·0~1 소수·5단계 등급 등)과 이용자 유형별 표시 수준이 설계서에 정의되어 있는지, 신뢰점수의 출처인 {AI모델설계서}의 모델 출력 정의와 화면 표시 방식 간 정합성이 확보되어 있는지를 확인한다.

NIST AI RMF는 AI 시스템의 의미 있는 투명성(Meaningful Transparency)을 위해 이용자의 역할과 지식 수준에 맞게 정보가 제공되어야 한다고 명시하며, 본 가이드라인은 그 핵심 구현 수단인 신뢰점수 표시에 관한 설계 요건이 설계 단계 산출물에 빠짐없이 반영되었는지를 검증하는 데 목적이 있다.

2. 필요성

▶ AI 과신(Over-reliance) 방지: AI 결과 화면에 추론 결과값만 표시되고 신뢰점수가 없으면, 이용자는 AI가 99% 확신하는 결과와 51% 확신하는 결과를 동일하게 받아들이게 된다. 이는 NIST AI RMF가 경고하는 AI 과신 현상을 직접적으로 유발한다.

▶ 설계 구체화 미비 시 형식적 구현 위험: 신뢰점수가 제공되더라도 설계 단계에서 표시 위치·형식·이용자 맥락이 구체화되지 않으면, 개발자가 임의로 소수점 6자리 수치를 화면 구석에 작은 글씨로 배치하는 방식으로 구현하여 실질적 정보 전달 효과가 없게 된다.

▶ 시험·검수 기준 확보: 신뢰점수 노출 방식이 설계서에 명시되어야 시험·검수 단계에서 "신뢰점수가 결과값 옆에 퍼센트 형식으로 표시되는가"와 같은 구체적인 테스트 케이스를 작성할 수 있으며, 미비 시 인수 기준 분쟁의 빌미가 된다.

▶ 신뢰 가능한 AI 서비스 설계 요건 이행: TTA 개발안내서는 AI 결과의 신뢰도 수준을 이용자에게 시각적으로 전달하는 것을 신뢰 가능한 AI 서비스의 설계 요건으로 규정하고 있어, 감리원이 이 항목을 설계 단계에서 체계적으로 점검하는 것이 필수적이다.

3. 검토방법

No	점검 항목	확인 방법
1	신뢰점수 표시 UI 요소 설계 여부	{화면설계서} 내 AI 결과 화면 목업에서 신뢰점수 UI 요소 존재 여부 확인 (세부: 수치·게이지 바·색상 등급 등 신뢰도를 나타내는 시각 요소 존재 여부)
2	신뢰점수 표현 형식 정의 여부	{화면설계서} 내 신뢰점수 표현 형식 스펙 기술 항목 검토 (세부: 백분율·0~1 소수·등급 레이블 등 출력 형식과 소수점 자리 수 명시)
3	신뢰점수 표시 위치 및 결과값과의 배치 관계 설계 여부	{화면설계서} 내 AI 결과 화면 레이아웃 목업에서 신뢰점수 위치 확인 (세부: 결과값 옆·아래·툴팁 등 위치와 시각적 연결 방식 정의)
4	이용자 유형별 신뢰점수 노출 수준 차등 설계 여부	{화면설계서} 내 역할별 화면 분기 항목 및 목업 검토 (세부: 최종 사용자용 간략형 / 운영자·감사자용 상세형 분기 설계)
5	신뢰점수 의미 설명 UI 설계 여부	{화면설계서} 내 신뢰점수 관련 툴팁·도움말 UI 설계 항목 확인 (세부: 툴팁·팝업 등을 통해 이용자가 수치 해석 방법을 이해할 수 있는 보조 설명)
6	{AI모델설계서} 모델 출력 형식 ↔ {화면설계서} 표시 형식 정합성 여부	{AI모델설계서}의 신뢰도 출력 정의와 {화면설계서}의 표시 형식 스프레드시트 교차 검토 (세부: 모델 출력 값 범위·형태와 화면 표시 방식의 일치 여부)
7	복수 후보 결과의 신뢰점수 비교 표시 설계 여부	{화면설계서} 내 다중 후보 결과 화면 목업 검토, 담당자 인터뷰 (세부: 상위 N개 예측 결과와 각 신뢰점수를 나란히 표시하는 설계)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	신뢰점수 UI 요소가 화면 목업에 존재하고, 표현 형식과 배치 위치가 설계서에 구체적으로 명시되어 있으며, {AI모델설계서}의 출력 형식과 화면 표시 형식 간 변환 규칙이 정의되고 이용자 유형별 차등 설계와 의미 설명 UI가 갖추어진 경우	현행 유지
보완권고 (△)	신뢰점수 표시 UI 요소는 설계되어 있으나 표현 형식·소수점 자리 수 미명시, 이용자 유형별 차등 설계 누락, 의미 설명 UI(툴팁·범례) 부재, 모델 출력과 화면 표시 형식 간 변환 규칙 미정의 등 일부 항목의 구체성이 미흡한 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	신뢰점수 표시 UI가 화면 목업에 없거나 "신뢰도를 표시한다"는 추상적 기술만 있는 경우, 표현 형식이 미정인 경우, {AI모델설계서}의 출력과 화면 표시 형식 간 불일치가 있는 경우, 이용자 대상 의미 설명 UI가 전혀 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 신뢰점수 수치의 정확성은 감리 범위 밖: 감리원은 "87%라는 신뢰점수가 실제로 맞는 수치인가"를 검증하는 것이 아니라, "신뢰점수가 설계서에 정의된 형식으로 결과 화면에 표시되도록 설계되어 있는가"를 확인하는 역할이다. 신뢰점수 산출 방법의 타당성에 의문이 있는 경우 AI 전문가 자문을 별도로 요청하도록 감리 의견에 기재하면 충분하다.

2) "신뢰도를 표시함" 텍스트 기술은 설계로 불인정: {화면설계서}에 "AI 결과 화면에 신뢰도를 함께 표시한다"고만 기재되어 있고 실제 화면 목업이 없는 경우는 부적합으로 판정하고 구체적인 화면 목업 작성을 요구한다. 목업 없이 텍스트로만 기술된 설계는 개발자마다 다른 구현으로 이어지므로 설계의 목적인 개발 지침 제공 기능을 이행하지 못한 것으로 본다.

3) LLM 서비스는 신뢰점수 산출 방법 자체를 먼저 확인: 생성형 AI(LLM) 기반 서비스는 전통적 분류 모델처럼 softmax 신뢰점수를 자동으로 제공하지 않는 경우가 많다. {AI모델설계서}에 LLM 신뢰도 산출 방법(별도 분류기 모델의 응답 품질 평가, 자기 평가 프롬프트 활용 등)이 기술되어 있는지를 먼저 확인하고, 그 방법이 없으면 화면 설계 이전에 모델 설계 자체가 불완전함을 지적한다.

4) Fail-safe·한계 고지와의 연계 설계 확인: 신뢰점수 노출 설계는 Fail-safe 로직 및 AI 한계 고지 UI와 긴밀히 연계된다. 신뢰점수가 임계값 이하일 때 Fail-safe가 발동하고 동시에 저신뢰도 경고 UI가 화면에 표시되는 흐름이 {화면설계서}와 {응용시스템설계서}에 일관되게 설계되어 있는지 종합적으로 확인하고, 설계서 간 기준이 다르면 설계 불일치로 지적한다.

5) 분기 처리(Routing) 설계 검증: 단순히 신뢰점수를 화면에 노출하는 것에 그치지 않고, 특정 점수(예: 60% 이하)일 경우 답변을 숨기거나 상담원 연결 버튼을 활성화하는 등 신뢰점수에 따른 UI 분기 처리 흐름이 {화면설계서}의 상태도(State Diagram)에 반영되어 있는지 확인한다.

AG_SD42-2-09 | AI결정 오버라이드 수동변경 기능점검

응용시스템 | 설계 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 종료(시험/전개) 단계에서, 관리자가 AI의 결정(판정·분류·추천·등급 등)을 수동으로 수정·저장할 수 있는 오버라이드 기능이 설계 의도대로 {시스템화면}에 정상 구현되고 {테스트결과서}를 통해 적정하게 시험되었는지를 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 수정 로직의 기술적 구현 방식이 아닌 다음 세 가지 관점에 집중한다.

① 수동변경 권한이 지정된 관리자 역할에만 부여되고 변경 가능 항목·범위가 'AI 오버라이드 기능 설계서'에 명확히 정의되어 있는가, ② 수동변경 이력(변경 전·후 값, 변경자, 변경 사유, 변경 일시)이 별도 로그·이력 테이블에 기록되어 추적 가능한가, ③ AI 결정이 수동변경되었음이 이후 화면·처리 결과에 명확히 표시되어 AI 결정과 인간 결정이 혼동되지 않는가. 이 세 가지 관점을 중심으로 설계 산출물과 {시스템화면}·{테스트결과서}의 일치 여부 및 시험의 충분성을 확인한다.

2. 필요성

▶ AI 결정 오버라이드(수동변경) 기능 동작 점검이 필요한 이유는 다음과 같다.

▶ 인간 감독(Human Oversight) 체계의 핵심 구현 요건: AI가 잘못된 결정을 내렸을 때 관리자가 이를 수정할 수 있는 오버라이드 기능은 AI 서비스의 안전망이다. 이 기능이 구현되지 않거나 특정 관리자에게만 제한 없이 개방되면, AI 오류에 대한 신속한 수정이 불가능하거나 반대로 무분별한 수동 수정으로 AI 결정 체계의 일관성과 신뢰성이 훼손될 수 있다.

▶ 수동변경 이력 미기록은 책임 추적 불가 및 감사 대응 실패 위험: 관리자가 AI 결정을 수동으로 변경하였을 때 변경 전·후 값, 변경자, 변경 사유가 기록되지 않으면, 이후 이의 제기·감사·사고 발생 시 어떤 결정이 AI에 의한 것인지, 어떤 결정이 인간에 의해 수정된 것인지 구분할 수 없어 책임 귀속이 불명확해진다.

▶ AI 결정과 수동변경 결정의 혼동은 이용자·후속 처리 시스템에 오인 유발: 수동변경 이후에도 화면·처리 결과에 "AI 결정"으로 표시되면, 이용자는 AI가 해당 결과를 생성했다고 오인한다. 이는 AI 결정 투명성 요건 위반이자, 수동변경 결과가 AI 성능 지표에 포함되어 AI 모델 평가가 왜곡되는 이중 문제를 야기한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {시스템화면} 등 산출물이 설계 의도대로 구현되었는지 여부와 시험의 적정성을 주요 점검 항목으로 규정하며, AI 서비스의 오버라이드 기능은 AI 특화 핵심 기능으로서 이 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	오버라이드 기능 접근 권한의 관리자 역할 제한 여부	오버라이드 기능 접근 권한이 'AI 오버라이드 기능 설계서'에 정의된 관리자 역할에만 부여되고 일반 이용자 계정으로는 수동변경 기능에 접근할 수 없는지 {시스템화면}에서 일반·관리자 계정을 비교하여 확인
2	수동변경 가능 항목·범위 정의 및 구현 여부	수동변경 가능한 AI 결정 항목·입력 허용 값 범위가 'AI 오버라이드 기능 설계서'에 정의되고 설계서에 없는 항목은 변경할 수 없도록 {시스템화면}에 구현되었는지 대조
3	변경 사유 필수 입력·미입력 저장 차단 여부	관리자 수동변경 시 변경 사유를 필수 입력하도록 구현되고 사유 미입력 시 저장이 차단되는지 {시스템화면}에서 확인
4	수동변경 이력의 감사 로그 기록 여부	변경 전 AI 결정값·변경 후 수동값·변경자 식별자·변경 사유·변경 일시가 이력 테이블 또는 감사 로그에 기록되는지 {테스트결과서}·{프로그램 명세서}에서 확인
5	수동변경 결과의 구분 표시 여부	수동변경된 결과가 이후 화면·후속 처리에서 'AI 결정'이 아닌 '수동변경됨' 등으로 구분 출력되는지 {시스템화면}·{테스트결과서}에서 확인
6	수동변경 이력 조회 화면 제공 여부	변경자·변경 사유·변경 전후 값 등 수동변경 이력이 관리자 화면에서 조회 가능한지 {시스템화면}에서 확인
7	오버라이드 운영 절차·승인 체계 정의 및 구현 여부	수동변경 허용 조건·변경 사유 기재 의무·변경 승인 체계(1인 단독 변경 불가 등)·이력 보관 기한이 'AI 오버라이드 운영 절차서'에 정의되고 이를 지원하는 시스템 기능이 구현되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	오버라이드 접근 권한이 관리자 역할로 제한되고, 변경 항목·범위 정의, 변경 사유 필수 입력, 변경 이력 기록, 수동변경 표시와 이력 조회 화면이 설계서대로 구현되며, 운영 절차서에 수동변경 허용 조건·변경 승인 체계·이력 보관 기한이 정의되어 {시스템화면}·{테스트결과서}에서 모두 확인되는 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	오버라이드 기능은 구현되었으나 변경 사유 필수 입력 미적용, 이력 항목 일부 누락, 수동변경 표시 미흡, 운영 절차서 미비 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	오버라이드 기능이 미구현되었거나, 권한 통제·이력 기록이 전혀 없어 수동변경의 추적과 책임 귀속이 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 관리자·일반 이용자 계정을 사전 확보하여 현장에서 권한 제어를 직접 실증한다: 착수 회의에서 두 계정을 모두 요청하여 각 계정으로 접속해 접근 차단 여부를 확인하며, URL 직접 입력을 통한 우회 접근 시도도 포함하여 점검한다.

2) 변경 사유 필수 입력 미구현은 책임 추적 불가로 이어지는 핵심 위험이다: 사유 입력이 선택 사항으로 구현된 경우 필수 입력 강제와 저장 차단 구현을 보완권고에 포함하고, 이를 배포 전 완료 조건으로 명시한다.

3) 이력 기록 시험은 "기록 정상"이라는 총평이 아니라 항목별 기록 성공 여부로 확인한다: 변경 전 AI 결정값, 변경 후 수동값, 변경자 식별자, 변경 사유, 변경 일시가 각각 저장되는지를 {테스트결과서}에서 항목 단위로 확인한다.

4) 수동변경 결과가 "AI 결정"으로 표시되지 않는지 확인한다: 변경 후에도 화면·후속 처리에서 AI 결정으로 표기되면 이용자가 오인하고, 수동변경 결과가 AI 성능 지표에 섞여 모델 평가가 왜곡되는 이중 문제가 발생한다. 감리원은 "수동변경됨" 등 구분 표시가 화면과 후속 처리에 반영되는지 확인한다.

5) 변경 승인 체계와 이력 조회 화면을 함께 점검한다: 1인 단독 변경이 무제한 허용되면 오버라이드 기능이 오히려 통제 실패 지점이 된다. 감리원은 'AI 오버라이드 운영 절차서'의 승인 체계가 화면 기능으로 구현되었는지, 이력 조회 화면에서 변경자·사유·변경 전후 값이 표시되는지 확인한다.

AG_SD42-2-10 | 개인화 데이터처리 알고리즘 설계

응용시스템 | 설계 | 인공지능 모델 개발 | 비식별화

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)

1. 목적

AI 서비스 구축 사업에서, AI 개인화 추천·맞춤형 서비스 제공을 위해 사용자 프로파일(이용자 행동 이력·선호도·인구통계 정보 등)을 활용할 때 개인정보에 해당하는 항목이 AI 모델 추론 과정에서 불필요하게 노출되거나 로그에 기록되지 않도록 마스킹(Masking) 또는 비식별화 처리하는 로직이 {응용시스템설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 개인화 알고리즘이 활용하는 사용자 프로파일 항목 중 개인정보 해당 항목의 목록과 항목별 마스킹 방식의 명세 여부, AI 모델 추론 입력값에 원본 개인정보가 포함되지 않고 마스킹·해시화·범주화된 값이 전달되도록 하는 처리 시점과 로직의 설계 여부, 마스킹 수준과 개인화 정밀도의 균형 검토 여부, 마스킹 처리 이전의 원본 프로파일 데이터가 AI 추론 과정의 임시 버퍼·로그·캐시에 잔류하지 않도록 하는 데이터 생명주기 설계 포함 여부를 확인한다.

이 점검은 마스킹 알고리즘의 기술적 강도를 평가하는 것이 아니라, 개인화 서비스에서 개인정보 최소 처리 원칙이 {응용시스템설계서}의 설계에 반영되어 있는지의 완결성을 확인하는 것을 목적으로 한다.

2. 필요성

▶ AI 개인화 서비스는 이용자별로 다른 콘텐츠·추천·응답을 제공하기 위해 이용자의 행동 이력(클릭 패턴·검색어·구매 이력 등), 인구통계 정보(연령대·지역·성별 등), 관심사·선호도를 수집하고 프로파일로 구성하여 AI 모델에 입력하며, 이 과정에서 개인정보 처리에 관한 구조적 위험이 발생한다.

▶ 원본 개인정보의 외부 전달·로그 기록 위험: AI 모델의 입력값에 마스킹되지 않은 원본 개인정보(이름·주민등록번호·연락처 등)가 포함되는 경우, AI 모델 서비스가 외부 API 형태로 운영되거나 로그가 수집되는 환경에서는 그 원본 개인정보가 제3자 시스템에 전달되거나 로그에 기록되어 개인정보 보호법상 제공·저장 근거 없는 처리가 이루어질 수 있다.

▶ 조합 재식별(Re-identification) 위험: 사용자 프로파일이 여러 개인정보 항목을 조합한 형태로 AI 모델에 전달될 때, 각 항목 자체는 마스킹되어 있더라도 조합으로 인해 특정 개인이 식별될 수 있는 재식별 위험이 있다.

▶ 설계 단계 미반영 시 구현 누락 위험: 개인화 서비스의 개인정보 마스킹 설계는 형식적 절차가 아니라 AI 서비스의 신뢰성과 이용자 권리 보호를 위한 핵심 설계 요소이다. 설계 단계에서 마스킹 로직이 {응용시스템설계서}에 포함되지 않으면 기능 구현이 우선되어 출시 시점까지 마스킹이 구현되지 않는 경우가 발생하며, 서비스 운영 중 발견 시에는 이미 개인정보가 처리된 이후가 된다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 입력 프로파일 항목 목록 및 개인정보 해당 항목 식별 설계 여부	{응용시스템설계서}의 AI 개인화 처리 흐름 또는 입력 특성(Feature) 명세에서 프로파일 항목 목록과 개인정보 구분 항목 확인, 담당자 인터뷰
2	개인정보 항목별 마스킹 방식 및 처리 시점 설계 여부	{응용시스템설계서}에서 AI 추론 입력 전 마스킹 처리 단계 및 각 항목별 마스킹 기법(해시화·범주화·삭제 등) 항목 확인
3	마스킹 처리 후 AI 모델 추론 입력값 구조 설계 여부	{응용시스템설계서}의 AI 추론 요청 입력 전문 구조에서 마스킹된 프로파일 항목의 포함 형태 확인
4	AI 추론 과정의 임시 데이터 잔류 방지(원본 프로파일 정리) 로직 설계 여부	{응용시스템설계서}의 AI 추론 처리 흐름에서 마스킹 처리 후 원본 데이터 삭제·정리 단계 항목 확인
5	조합 재식별(Combination Re-identification) 위험 항목 검토 및 처리 설계 여부	{응용시스템설계서}의 사용자 프로파일 항목 조합 구성에서 재식별 위험 항목 식별 및 처리 방안 항목 확인, 담당자 인터뷰
6	개인화 서비스 미동의 또는 개인정보 처리 철회 시 프로파일 처리 흐름 설계 여부	{응용시스템설계서}의 개인화 동의 철회 또는 개인정보 삭제 요청 처리 흐름에서 AI 추론 입력 프로파일 변경 로직 항목 확인
7	{요구사항정의서} ↔ {응용시스템설계서} 개인정보 마스킹 요건 추적성 확보 여부	{요구사항정의서}의 개인정보 처리·마스킹·개인화 동의 관련 요건과 {응용시스템설계서}의 마스킹 로직 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 개인화 입력 프로파일 항목 전체 목록과 개인정보 구분이 설계되어 있고, 항목별 마스킹 기법이 명시되어 있으며, 마스킹 후 AI 추론 입력 전문에 원본 개인정보가 포함되지 않는 구조가 설계되어 있고, 임시 데이터 잔류 방지·재식별 위험 검토·동의 철회 처리 흐름·요구사항 추적성이 모두 확인된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	마스킹 처리 단계는 설계되어 있으나 항목별 마스킹 기법 명세가 불충분하거나, 조합 재식별 위험 검토 미흡, 로그 마스킹 설계 불명확, 동의 철회 시 캐시·임시 프로파일 처리 흐름 일부 누락, 요구사항 추적성 일부 미확보 등 구체성·완전성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 모델 입력 프로파일 항목 목록이 없는 경우, 항목별 마스킹 기법이 "개인정보를 보호한다"는 추상적 기술에 그치는 경우, AI 추론 요청 입력 전문에 원본 개인정보 필드가 포함된 경우, 재식별 위험 검토가 수행되지 않았거나 동의 철회 시 AI 캐시 프로파일 처리 흐름이 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (개인정보 보호법 준수 촉구)

5. 감리 시 유의사항 및 실무 Tip

1) "개인정보를 보호합니다"는 설계가 아닌 의도의 표현: 설계서에 "사용자 프로파일의 개인정보는 마스킹 처리하여 AI 모델에 제공합니다"라는 한 줄 기술만 있는 경우가 많다. 감리원은 어떤 프로파일 항목에 대해, 어떤 마스킹 기법을(삭제·해시화·범주화·일반화 중 무엇을), 처리 흐름의 어느 단계에서 적용하는지가 항목 단위로 명세되어 있어야 설계가 완료된 것으로 판정한다.

2) 마스킹 설계의 존재 여부와 항목 단위 명세 여부에 집중: 해시 알고리즘 종류나 범주화 구간의 적절성 등 마스킹 기법의 기술적 강도 평가는 감리 범위를 벗어난다. 각 프로파일 항목에 마스킹 처리가 설계되어 있는지와 구체적 기법으로 명세되어 있는지에 집중하고, 마스킹 대상 항목·기법이 '개인정보영향평가(PIA)' 결과 또는 발주기관의 '개인정보보호지침'과 일치하는지 요구사항 추적 매트릭스로 교차 검증한다.

3) 외부 AI API 사용 시 원본 개인정보 전달 여부 중점 점검: 외부 상용 AI API를 활용하는 경우 API 요청 본문에 원본 개인정보가 포함되면 제3자인 AI API 제공사에 개인정보가 제공되는 것으로 별도의 제3자 제공 동의가 필요하다. {응용시스템설계서}의 AI API 호출 입력 전문 구조에서 원본 개인정보 포함 여부를 반드시 확인하고, 마스킹 없이 전달하도록 설계된 경우 개인정보 보호법 위반 위험을 명시하여 시정을 강력히 권고한다.

4) 개인화 동의 철회와 회원 탈퇴 처리의 구분 확인: "개인정보 처리 철회 = 서비스 탈퇴"로 설계된 경우 이용자가 개인화 기능만 철회하고 서비스를 계속 이용할 수 없게 되어 사실상 동의 철회를 어렵게 만든다. 철회 후 비개인화(기본값) 모드로 서비스를 계속 이용할 수 있는 구조인지, 개인화 적용 여부를 관리하는 상태값이 별도로 설계되어 있는지 확인한다.

5) 데이터베이스 영역 감리와의 점검 범위 구분: 프로파일 테이블 암호화·DB 접근 권한 통제는 데이터베이스 영역 감리원의 범위이다. 응용시스템 감리원은 AI 추론 전 마스킹 처리 단계의 설계 완결성, AI 추론 입력 전문의 개인정보 포함 여부, 임시 데이터 잔류 방지 로직, 동의 철회 시 프로파일 처리 흐름에 집중하고, 두 영역의 점검 범위를 감리 결과서에 명확히 구분하여 기재한다.

AG_SD42-2-11 | 안전모드 Fail-safe 로직 설계

응용시스템 | 설계 | 시스템 구현 | 가용성/이중화

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

AI 서비스 구축 사업에서, AI 모델의 추론 결과가 신뢰도 임계값(Confidence Threshold)에 미달하거나 오류·이상 상황이 발생했을 때 시스템이 안전하게 동작할 수 있도록 Fail-safe 로직이 {응용시스템설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① AI 추론 신뢰도가 사전 정의된 임계값에 미달할 경우 적용되는 기본값(Default Value) 반환 또는 서비스 제공 보류 로직이 설계되어 있는지, ② 임계값 미달 또는 시스템 오류 발생 시 관리자에게 전환(Human-in-the-Loop)하는 로직과 그 절차가 명시되어 있는지, ③ 안전모드 진입 조건과 해제 조건, 이용자·관리자에게 전달되는 안내 메시지가 설계서에 정의되어 있는지를 확인한다.

NIST AI RMF는 AI 시스템이 지식 한계를 초과하여 동작할 경우 안전하게 실패(Fail Safely)할 수 있어야 하며, 이를 위한 인간 개입 절차를 사전에 마련할 것을 요구한다.

2. 필요성

▶ AI 모델은 학습 데이터의 범위를 벗어난 입력이 들어오거나, 모델 자체의 확신 수준이 낮은 상황에서도 형식적으로 결과를 출력하는 특성이 있다. 이 결과가 적절한 안전 로직 없이 서비스에 그대로 반영되면, 대출 심사 오결정·의료 권고 오류·교통 제어 오작동 등 사회적으로 중대한 피해로 이어질 수 있다.

▶ NIST AI RMF는 안전성(Safety)을 신뢰할 수 있는 AI의 핵심 특성으로 규정하고, AI 시스템이 기대 동작 범위를 벗어날 때 인간이 개입하거나 시스템을 종료·수정할 수 있는 절차를 사전에 마련할 것을 강조한다. 설계 단계에서 Fail-safe 로직이 {응용시스템설계서}에 반영되지 않으면, 개발자가 임의의 예외 처리 수준으로 구현하여 안전모드가 형식적으로만 존재하게 될 위험이 있다.

▶ 또한 임계값 수치와 관리자 전환 절차가 설계 단계에서 문서화되지 않으면, 시험·검수 단계에서 합격/불합격 기준을 정할 수 없어 인수 시 극심한 분쟁이 발생한다. 감리원이 이 단계에서 Fail-safe 설계의 완결성을 조기에 점검하는 것이 필수적인 이유가 여기에 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 신뢰도 임계값 정의 여부	{응용시스템설계서} 내 임계값 정의 항목 검토, 담당자 인터뷰 (세부: 신뢰도가 몇 % 미만일 때 안전모드를 발동할지 수치로 정의)
2	임계값 미달 시 기본값(Default Value) 반환 로직 설계 여부	{응용시스템설계서} 내 임계값 미달 분기 처리 로직 검토 (세부: 안전한 사전 정의 값 또는 서비스 보류 처리 방식)
3	관리자 전환(Human-in-the-Loop) 로직 및 절차 설계 여부	{응용시스템설계서} 내 관리자 전환 시나리오 및 흐름도 검토 (세부: 전환 발동 조건, 전환 대상 담당자, 처리 방식)
4	Fail-safe 진입 조건과 해제 조건의 명시 여부	{응용시스템설계서} 내 상태 전환 조건 항목 검토 (세부: 안전모드 진입 트리거 및 정상 복귀 조건)
5	이용자 및 관리자 안내 메시지 설계 여부	{응용시스템설계서} 및 {화면설계서} 연계 항목 검토 (세부: Fail-safe 발동 시 이용자에게 표시할 문구와 관리자 알림 방식)
6	Fail-safe 이력 기록(로깅) 설계 여부	{응용시스템설계서} 내 로그 설계 항목 검토 (세부: 안전모드 발동·해제 일시, 원인, 처리 결과 기록)
7	{요구사항정의서} ↔ {응용시스템설계서} Fail-safe 항목 추적성 확보 여부	{요구사항정의서}의 Fail-safe 요건과 {응용시스템설계서} 설계 항목 간 매트릭스 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	신뢰도 임계값이 수치로 정의되고, 기본값 반환·관리자 전환 로직, 안전모드 진입·해제 조건, 안내 메시지, 발동 이력 로깅이 {응용시스템설계서}에 모두 설계되고 요구사항과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	Fail-safe 설계는 존재하나 임계값 수치 미정의, 해제 조건·안내 메시지 누락, 로깅 설계 미흡 등 일부 항목이 불완전한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 특화 Fail-safe 로직이 설계되어 있지 않거나 일반 에러 처리로 대체되어 있어 저신뢰도 출력에 대한 안전장치를 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 임계값 수치의 적정성 판단은 감리 범위 밖임을 인식할 것: 신뢰도 임계값을 몇 %로 설정할지는 도메인 특성과 서비스 위험도에 따른 사업자·발주기관의 판단 영역이다. 감리원은 임계값이

수치로 정의되어 있는지, 그 설정 근거와 발주기관 협의·승인 이력이 문서화되어 있는지만 점검한다.

2) Fail-safe 로직과 단순 에러 처리(Error Handling)를 구분하여 점검할 것: AI 신뢰도 임계값 미달에 따른 Fail-safe 로직은 서버 오류(500 Error)나 네트워크 장애에 대한 일반적인 예외 처리와는 구별되어야 한다. 일부 설계서에서 AI 특화 안전모드 설계 없이 일반 에러 처리 로직을 Fail-safe로 표기하는 경우가 있으므로, 설계서에 신뢰도 점수 기반의 분기 로직이 별도로 존재하는지 반드시 확인한다.

3) 임계값 복수 단계 설계(경고 구간 → 안전모드 구간) 여부를 확인할 것: 단일 임계값보다 경고 구간과 안전모드 구간을 구분한 다단계 설계가 운영 안정성 확보에 유리하다. 다단계 설계 여부와 구간별 처리 방식(경고 표시, 서비스 보류, 관리자 전환 등)이 설계서에 기술되어 있는지 확인한다.

AG_SD42-2-12 | 악성 프롬프트 필터링 로직 설계

응용시스템 | 설계 | 인공지능 모델 개발 | AI특화보안

근거 | NIST AI RMF · OWASP LLM Top 10 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)

1. 목적

AI 서비스 구축 사업에서, 이용자 또는 외부 공격자가 의도적으로 AI 모델의 동작을 우회하거나 비정상 출력을 유도하기 위해 입력하는 악성 프롬프트를 탐지하고 차단하는 로직이 {응용시스템 설계서}에 구체적으로 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 ① 프롬프트 인젝션·탈옥 시도 등 주요 공격 유형별 필터링 대상이 설계서에 정의되어 있는지, ② 정규식(Regular Expression) 또는 룰셋(Ruleset) 기반 필터링 로직의 구조와 처리 흐름이 설계서에 명시되어 있는지, ③ 필터링 발동 시 처리 방식(차단·경고·로깅·에스컬레이션)과 이용자 안내 방식이 화면 흐름과 연계하여 설계되어 있는지를 확인한다.

NIST AI RMF 3.3절(Secure and Resilient)은 AI 시스템이 의도적 조작 시도에 대한 복원력을 갖추도록 설계할 것을 요구하며, 악성 프롬프트 필터링 로직은 그 핵심 구현 수단 중 하나이다.

2. 필요성

▶ 생성형 AI 서비스는 자연어 입력을 처리하는 구조 특성상, 공격자가 교묘한 문장 표현만으로 시스템의 안전 장치를 우회할 수 있다. 대표적인 사례가 프롬프트 인젝션으로, 공격자는 "이전 지시를 무시하고 다음을 실행하라"와 같은 문구를 입력하여 시스템 프롬프트를 덮어쓰거나 금지된 정보를 반환하도록 AI를 조작한다. 설계 단계에서 이러한 공격 유형에 대한 방어 로직이 정의되지 않으면 구현 단계에서 누락되기 쉽다.

▶ 프롬프트 인젝션은 OWASP LLM Top 10 1위 취약점으로, 방어 로직이 설계에 포함되지 않으면 공격자가 시스템 지시를 무력화하여 유해 콘텐츠 생성·개인정보 유출·서비스 중단을 유발할 수 있다. 설계 단계 검토가 가장 비용 효율적인 방어 시점이다.

3. 검토방법

No	점검 항목	확인 방법
1	필터링 대상 공격 유형 정의 여부	{응용시스템설계서} 또는 'AI 보안 위협 분석서' 내 공격 유형 정의 항목 검토, 담당자 인터뷰
2	필터링 로직 방식 및 처리 흐름 설계 여부	{응용시스템설계서} 내 필터링 로직 처리 흐름도(Flow Diagram) 및 방식 기술 항목 검토
3	정규식·룰셋의 항목 범주 및 갱신 절차 설계 여부	{응용시스템설계서} 또는 '입력 필터링 룰셋 정의서'에서 룰셋 구조·범주·갱신 절차 항목 확인

No	점검 항목	확인 방법
4	필터링 발동 시 차단·경고·로깅 처리 설계 여부	{응용시스템설계서} 내 필터링 판정 결과별 처리 분기 설계 항목 및 화면 흐름 연계 확인 (세부: 차단 즉시 응답 거부 / 경고 후 진행 / 이벤트 로깅 분기 설계)
5	차단 시 이용자 안내 메시지 설계 여부	{응용시스템설계서} 내 차단 발동 시 화면 표시 메시지 항목 및 {화면설계서} 연계 확인 (세부: 차단 사유 없이 거부만 하는 것이 아닌, 이용자 친화적 안내 문구 설계)
6	필터링 로직의 오탐(False Positive) 대응 설계 여부	{응용시스템설계서} 내 오탐 대응 처리 항목 및 {화면설계서} 이의 제기 UI 설계 확인 (세부: 정상 입력이 잘못 차단될 경우 이용자 이의 제기 또는 재시도 절차)
7	{요구사항정의서} ↔ {응용시스템설계서} 보안 필터링 요건 추적성 확보 여부	{요구사항정의서} 보안 요건 항목, 'AI 보안 위협 분석서' 위협 목록, {응용시스템설계서} 필터링 설계 항목 3방향 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	공격 유형 목록, 탐지 방식과 처리 흐름도, 룰셋 범주·갱신 절차, 차단·경고·로깅 분기, 이용자 안내 및 오탐 대응이 {응용시스템설계서}에 모두 설계되고 보안 요건과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	필터링 로직은 설계되어 있으나 공격 유형 정의 불충분, 룰셋 갱신 절차 미정의, 오탐 대응·이용자 안내 설계 누락 등 일부 항목이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	악성 프롬프트 필터링 설계가 {응용시스템설계서}에 반영되어 있지 않거나 일반 입력 유효성 검증으로 대체되어 있어 AI 특화 보안 요건을 충족할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 일반 입력 유효성 검증(Validation)과 AI 악성 프롬프트 필터링을 반드시 구분하여 점검할 것: 사업자가 "SQL 인젝션·XSS 방어 설계가 있으므로 악성 입력 차단은 완료되었다"고 주장하는 경우가 있다. 그러나 웹 보안의 입력 유효성 검증은 데이터베이스·HTML 렌더링 맥락의 공격을 막는 것이고, AI 프롬프트 필터링은 자연어 문장 구조를 통해 AI 모델의 동작을 조작하는 공격을 막는 것으로 완전히 다른 설계 항목이다. 두 가지가 {응용시스템설계서}에 각각 별도 항목으로 설계되어 있는지 확인한다.

2) 룰셋 문자열 자체의 기술적 타당성은 감리 범위 밖임을 명확히 인식할 것: "이 정규식 패턴이 실제로 공격을 탐지하는 데 충분한가"를 검증하는 것은 보안 전문가의 역할이다. 감리원은 "룰

셋이 범주별로 구조화되어 있는가", "갱신 절차가 설계서에 정의되어 있는가", "보안 요건과 필터링 설계 항목 간 추적성이 확보되어 있는가"만을 확인한다. 룰셋의 기술적 충분성에 의문이 있는 경우 보안 전문가 자문을 감리 의견에 병기하면 충분하다.

3) LLM 서비스에서는 정규식만으로는 탐지가 불충분함을 사업자에게 안내할 것: 자연어의 변형·우회 표현은 무궁무진하므로 정규식·룰셋 단독 방식은 근본적 한계가 있다. AI 기반 분류기 병행, 시스템 프롬프트 방어 강화 등 다층 방어 설계 여부를 확인하고, 단일 방식에만 의존하는 설계는 보완을 권고한다.

AG_SD42-2-13 | 실시간 성능지표 수집로직 설계

응용시스템 | 설계 | 시스템 구현 | 모니터링/재학습

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「개인정보 보호법」

1. 목적

AI 서비스 구축 사업에서, AI 서비스의 품질과 이용자 반응을 지속적으로 측정하기 위해 이용자 행동(클릭, 페이지 체류 시간 등)과 AI 추론 성능 지표(응답 시간, 추론 정확도 등)를 실시간으로 수집하는 로그 수집 객체(Log Collector Object)가 {응용시스템설계서}에 구체적으로 설계되어 있는지를 점검하는 것이 목적이다.

구체적으로는 ① 수집 대상 이벤트와 각 이벤트의 로그 객체 구조(필드명·타입·수집 시점)의 명세 여부, ② 로그 수집이 AI 추론 응답 성능에 영향을 주지 않는 비동기·비차단 방식 설계 여부, ③ 수집 로그 내 개인정보 항목과 비식별화 처리 방식의 정의 여부, ④ {요구사항정의서}의 지표 요건이 로그 수집 객체 설계로 빠짐없이 반영되었는지를 확인한다.

TTA 개발안내서와 NIST AI RMF는 운영 단계 품질 모니터링 체계의 사전 설계를 요구한다.

2. 필요성

▶ AI 서비스는 일반 정보 시스템과 달리 동일한 입력에도 상황에 따라 다른 응답을 생성하는 비결정론적 특성을 가진다. 이 때문에 시스템이 "정상 가동 중"이더라도 AI 모델의 응답 품질이 점진적으로 저하되거나(모델 드리프트), 특정 유형의 질의에서 이용자 이탈률이 급증하는 현상이 운영 중에 조용히 발생할 수 있다.

▶ 이러한 품질 변화를 조기에 탐지하려면 이용자 행동(클릭, 체류 시간, 재질문 여부 등)과 AI 추론 성능 지표(응답 시간, 오류 발생 횟수, 신뢰도 점수 분포 등)를 수집하는 로그 체계가 서비스 출시 첫날부터 작동해야 한다. 설계 단계에서 로그 수집 객체 설계가 {응용시스템설계서}에 포함되어 있지 않으면, 개발자가 "나중에 추가하면 된다"는 식으로 구현을 미루어 출시 시점에 수집 체계가 누락될 위험이 크다.

▶ 또한 AI 서비스 로그에는 이용자의 질의 내용, 클릭 이력, 체류 시간 등 개인정보에 해당하는 데이터가 포함될 수 있으므로, 설계 단계에서 수집 항목과 비식별화 방식을 정의하지 않으면 개인정보 보호법 위반 위험이 발생한다. 로그 수집 객체 설계는 단순한 개발 편의 기능이 아니라 AI 서비스의 품질 보증과 법적 준수를 위한 필수 설계 항목이다.

3. 검토방법

No	점검 항목	확인 방법
1	수집 대상 이벤트 목록 및 이벤트별 로그 객체 구조 설계 여부	{응용시스템설계서}에서 수집 이벤트 목록과 각 이벤트의 로그 객체 필드 명세 항목 확인, 담당자 인터뷰
2	사용자 행동 지표(클릭·체류 시간) 수집 로직 설계 여부	{응용시스템설계서}에서 클릭 이벤트 수집 조건 및 체류 시간 계산 기준(시작·종료 이벤트) 항목 확인
3	AI 추론 성능 지표(응답 시간·오류율·신뢰도) 수집 로직 설계 여부	{응용시스템설계서}에서 AI 추론 성능 관련 이벤트의 수집 위치(미들웨어·서비스 레이어 등) 및 측정 기준 항목 확인
4	로그 수집의 비동기·비차단(Non-blocking) 방식 설계 여부	{응용시스템설계서}의 로그 수집 처리 흐름도에서 비동기 처리 여부 및 메인 서비스 스레드와의 분리 구조 확인
5	로그 데이터 내 개인정보 항목 정의 및 비식별화 처리 설계 여부	{응용시스템설계서} 및 '개인정보 처리방침 설계서'에서 로그 수집 항목 중 개인정보 해당 여부 및 처리 방식 확인
6	로그 수집 객체의 버전 관리 및 스키마 변경 대응 설계 여부	{응용시스템설계서}에서 로그 객체 스키마 버전 관리 방식 및 하위 호환성 처리 항목 확인 (세부: AI 모델 버전 교체나 UI 변경 시 로그 객체 스키마를 버전으로 관리하는 설계)
7	{요구사항정의서} ↔ {응용시스템설계서} 성능 지표 수집 요건 추적성 확보 여부	{요구사항정의서} 성능 지표·품질 모니터링 관련 요건과 {응용시스템설계서} 로그 수집 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	수집 이벤트 목록과 로그 객체 구조, 행동·추론 성능 지표 수집 로직, 비동기 처리 구조, 개인정보 비식별화 방식, 스키마 버전 관리가 {응용시스템설계서}에 모두 설계되고 요구사항과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	로그 수집 설계는 존재하나 이벤트별 필드 명세 미흡, 비동기 처리 구조 불명확, 개인정보 처리 방식 일부 미정의 등 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	로그 수집 객체 설계가 {응용시스템설계서}에 반영되어 있지 않거나 수집 도구 명칭만 언급되어 있어 품질 모니터링 기반을 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "지표 수집 기능이 있다"와 "지표 수집 설계가 되어 있다"를 명확히 구분하여 점검할 것: {응용시스템설계서}에 "GA(Google Analytics) 또는 유사 도구를 사용하여 이용자 행동을 수집한다"는 한 줄 기술은 수집 도구 도입 의도일 뿐, 수집 이벤트 목록·로그 객체 구조·비동기 처리 방식이 설계된 것이 아니다. 감리원은 수집 도구 언급이 아닌, 수집 대상 이벤트 목록과 각 이벤트의 구조 명세가 설계서에 포함되어 있는지를 기준으로 판정한다.

2) 이용자 질의 텍스트의 로그 수집 여부를 반드시 확인하고 처리 방식이 없는 경우 보완을 강력히 권고할 것: 이용자가 AI 챗봇이나 AI 검색 서비스에 입력하는 텍스트에는 이름, 건강 정보, 위치 정보 등 민감한 개인정보가 포함될 수 있다. 이를 사전 동의 없이 수집하면「개인정보 보호법」 위반이 되므로, 로그 수집 객체 설계에서 질의 텍스트 필드의 처리 방식(수집 제외·해시화·마스킹 등)이 정의되어 있는지 확인한다.

3) "나중에 필요한 지표가 생기면 수집을 추가하면 된다"는 사업자 의견에 대해 사전 설계의 필요성을 설득력 있게 안내할 것: 로그 스키마가 사전에 설계되지 않으면 출시 이후 수집을 소급 적용할 수 없어 초기 운영 데이터가 영구 유실된다. 스키마 버전 관리 설계와 함께 출시 시점부터 수집이 작동해야 하는 이유를 안내하고 설계 반영을 권고한다.

AG_SD42-2-14 | 타시스템 API연계 프로토콜설계

응용시스템 | 설계 | 시스템 구현 | AI모델 API

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」 · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

AI 서비스 구축 사업에서, AI 서비스가 외부 시스템 또는 공공 데이터 API와 연계할 때 사용하는 인터페이스 프로토콜의 입출력 전문, 오류 코드 체계, 예외 처리 흐름이 {인터페이스설계서}에 명확하고 완전하게 설계되어 있는지를 감리원의 관점에서 점검하는 것이 목적이다.

구체적으로는 각 연계 인터페이스별 입력 전문 구조(필드명·타입·길이·필수 여부)와 출력 전문 구조(AI 추론 결과 포함 응답 스키마)의 명세 여부, HTTP 상태 코드 외 AI 서비스 전용 애플리케이션 오류 코드 체계의 정의 여부, 외부 API 응답 지연·타임아웃·부분 실패·AI 추론 결과 이상값 등 AI 서비스 특유의 예외 상황에 대한 처리 흐름 포함 여부, {요구사항정의서}의 연계 요건이 {인터페이스설계서}의 설계 항목에 빠짐없이 반영되어 있는지를 확인한다.

이 점검은 AI 서비스 특유의 연계 설계 품질(AI 모델 입출력 전문 구조의 완결성, AI 추론 실패 시 연계 시스템의 대응 방안)에 집중하며, 일반적인 IT 시스템의 API 설계 적정성 점검 항목은 범위에서 제외한다.

2. 필요성

▶ AI 서비스는 일반 웹 애플리케이션과 달리 외부 AI 모델 API(OpenAI, HyperCLOVA 등), 공공 데이터 API, 지식 베이스 연계 API 등 다양한 외부 시스템과 동적으로 연계되는 구조를 가지며, 이 연계 지점에서 발생하는 오류는 단순한 "API 호출 실패"에 그치지 않고 AI 모델이 잘못된 입력을 받아 편향되거나 불완전한 응답을 생성하는 데이터 품질 문제로 이어질 수 있다.

▶ 공공 API에서 수신한 데이터에 결측값(NULL)이 포함되어 있을 때 AI 모델이 이를 어떻게 처리하는지, 외부 AI API가 부분 응답을 반환했을 때 서비스가 이를 완성된 응답으로 처리하지 않도록 검증하는지가 설계 단계에서 명확히 정의되어야 한다.

▶ AI 서비스는 추론 과정이 블랙박스 특성을 가지므로, 연계 API로부터 비정상 응답이 수신되었을 때 이를 탐지하고 이용자 또는 관리자에게 적절히 알리는 예외 처리 흐름이 없으면 이상 응답이 이용자에게 그대로 전달되는 사고가 발생한다.

▶ 설계 단계에서 {인터페이스설계서}에 이러한 AI 서비스 특유의 예외 상황이 포함되어 있지 않으면 개발 단계에서 예외 처리 로직이 생략되거나 임시방편으로 구현되어 운영 장애의 원인이 된다.

3. 검토방법

No	점검 항목	확인 방법
1	연계 인터페이스 목록 및 입력 전문 구조 완결성	{인터페이스설계서}에서 연계 인터페이스 목록 및 각 인터페이스의 요청 전문 상세 명세 항목 확인, 담당자 인터뷰
2	출력 전문 구조 및 AI 추론 결과 응답 스키마 완결성	{인터페이스설계서}에서 각 인터페이스의 응답 전문 명세 항목 확인, AI 결과 필드(추론 결과, 신뢰도, 근거 텍스트 등) 구조 포함 여부 확인
3	HTTP 상태 코드 외 애플리케이션 오류 코드 체계 정의 여부	{인터페이스설계서}의 오류 코드 정의 섹션에서 애플리케이션 레벨 오류 코드 목록 및 각 코드별 설명 항목 확인
4	타임아웃 기준 및 재시도 정책 설계 여부	{인터페이스설계서} 내 각 연계 인터페이스별 타임아웃 설정값 및 재시도 정책 항목 확인 (세부: 외부 AI API·공공데이터 API 각각의 타임아웃 임계값, 재시도 횟수·간격·백오프 정책)
5	AI 추론 결과 이상값(Abnormal Output) 탐지 및 처리 흐름 설계 여부	{인터페이스설계서} 또는 {응용시스템설계서}에서 AI 추론 결과 이상값 탐지 기준 및 처리 흐름 항목 확인
6	외부 데이터 수신 시 입력 유효성 검증(Validation) 설계 여부	{인터페이스설계서} 또는 {응용시스템설계서}에서 외부 수신 데이터의 AI 모델 전달 전 유효성 검증 항목 및 검증 실패 시 처리 흐름 확인
7	{요구사항정의서} ↔ {인터페이스설계서} 연계 요건 추적성 확보 여부	{요구사항정의서}의 API 연계·오류 처리 관련 요건과 {인터페이스설계서} 설계 항목을 스프레드시트로 2방향 매핑하여 누락·불일치 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	연계 인터페이스 전체 목록이 있고, 각 인터페이스의 입출력 전문 구조(AI 모델 전용 항목 포함)가 필드 단위로 명세되어 있으며, 애플리케이션 오류 코드 체계·타임아웃 기준·AI 이상값 탐지·외부 데이터 유효성 검증·요구사항 추적성이 모두 확인된 경우	현행 유지
보완권고 (△)	연계 인터페이스와 입출력 전문의 기본 설계는 있으나 AI 모델 입력 전문의 프롬프트 구조 미명세, 타임아웃·재시도 정책의 수치 기준 미정의, AI 이상값 탐지 조건의 추상적 기술, 유효성 검증 단계 일부 누락, 요구사항 추적성 일부 미확보 등 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	연계 대상 시스템이 설계서에 누락된 경우, 입출력 전문이 필드 단위로 명세되지 않고 개요 수준에 그친 경우, 애플리케이션 오류 코드가 없는 경우, 타임아웃 기준값이 없는 경우, AI 이상값 탐지 조건이 없는 경우, 외부 데이터 유효성 검증 단계가 처리 흐름도에 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 일반 IT 항목과 AI 특화 항목의 구분: 인터페이스설계서 점검은 일반 IT 감리에서도 수행하는 항목이지만, 본 가이드라인은 AI 서비스 특유의 연계 설계 품질(AI 모델 입출력 전문 구조, AI 이상값 탐지, GIGO 방지 유효성 검증)에 집중한다. 필드 명세 누락이나 HTTP 오류 코드 정의와 같은 일반적인 설계 미비 사항은 감리 의견으로 기록하되, 핵심 AI 특화 항목과 명확히 구분하여 감리 결과서를 작성한다.

2) 외부 AI API 오류 코드의 공식 문서 근거 확인: 외부 AI API(OpenAI, HyperCLOVA 등)는 rate_limit_exceeded, context_length_exceeded 등 각자의 고유한 오류 코드 체계를 갖는다. {인터페이스설계서}의 오류 코드 처리 설계가 해당 API 제공사의 공식 문서에 근거하여 정의되었는지 담당자 인터뷰로 확인하고, 임의 정의되었거나 실제 오류 코드와 불일치하는 경우 설계 보완을 권고한다.

3) AI 모델 입력 전문의 프롬프트 구조 별도 확인: AI 모델 API 입력 전문은 일반 API와 달리 프롬프트 구조(시스템 프롬프트 위치, 이용자 입력 삽입 방식, 컨텍스트 데이터 포맷)가 추가로 설계되어야 한다. 이 항목이 {인터페이스설계서}에 없으면 개발자마다 프롬프트 구조를 임의로 구성하여 AI 응답 품질이 일관되지 않는 문제가 발생하므로, 일반 인터페이스 점검 체크리스트에 없는 이 항목을 감리원이 별도로 확인한다.

4) 폴백(Fallback) 처리의 설계서 간 정합성 교차 확인: 외부 AI API 실패·이상 응답 시의 대체 처리(기본 응답 메시지 제공, 규칙 기반 처리 대체 등)는 {인터페이스설계서}의 예외 처리 흐름과 {응용시스템설계서}의 AI 추론 처리 로직에 일관되게 반영되어야 한다. 두 문서 중 하나에만 폴백이 설계된 경우 설계 불일치로 지적하고 정합성 확보를 권고한다.

5) NULL 허용 여부와 이상값 탐지 기준의 연계 확인: 응답 전문의 신뢰도 점수 필드가 NULL 허용(Nullable)이면 이상값 탐지 로직에 신뢰도 NULL 조건을 별도로 추가해야 하고, Non-null(필수값)이면 NULL 수신 자체가 스키마 위반이므로 즉시 오류 처리로 분기해야 한다. 전문 명세의 Nullable 설정과 이상값 탐지 기준이 서로 일관성 있게 설계되어 있는지 두 항목을 함께 확인하여 판단한다.

AG_SD42-2-15 | AI모델 아키텍처 점검

응용시스템 | 설계 | 인공지능 모델 개발 | LLM모델

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

AI 서비스 구축 사업에서, {요구사항정의서}에 정의된 AI 서비스 요건이 {AI모델설계서}에 빠짐없이 구체적으로 반영되었는지를 감리원의 관점에서 제3자적으로 점검하는 것이 목적이다.

구체적으로는 서비스 목적에 부합하는 AI 모델 유형(LLM, RAG, Fine-tuning 등)이 선정 근거와 함께 명시되었는지, 알고리즘 구조 및 데이터 처리 흐름이 다이어그램 수준으로 상세화되었는지, 그리고 환각(Hallucination)·편향(Bias) 등 AI 특유의 리스크에 대한 설계적 대응 방안이 문서화되었는지를 확인한다.

이를 통해 개발 단계 진입 전에 설계 미비로 인한 결함을 사전 식별하고, 이후 시험·검사 단계에서의 분쟁을 예방한다.

2. 필요성

▶ 생성형 AI 시스템은 모델 선택, 학습 방식, 프롬프트 전략 등의 핵심 설계 의사결정이 집중되며, 이 결정들은 이후 개발·시험·운영 전 단계의 품질에 연쇄적으로 영향을 미친다. 일반 정보시스템과 달리, AI 설계서에는 ① 모델 선정 근거 및 아키텍처 구조, ② 프롬프트 설계 전략, ③ 환각·편향 대응 방안이 반드시 포함되어야 한다.

▶ AI 모델 아키텍처 결정(기반 모델 종류, 파인튜닝 방식, 프롬프트 전략)은 시스템 성능·비용·확장성에 직접 영향을 미치며, 설계 단계에서 문서화되지 않으면 구현 단계에서 임의 변경이 발생하여 감리 추적이 불가능해진다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 유형 선정 근거 명시 여부	{AI모델설계서} 검토 및 담당자 인터뷰 (세부: LLM, RAG, Fine-tuning, 혼합 방식 등)
2	알고리즘 구조 및 AI 처리 흐름의 상세화 여부	{AI모델설계서} 내 아키텍처 다이어그램 및 흐름도 검토 (세부: 입력→처리→출력, 외부 API 연동 지점 포함)
3	프롬프트 설계(Prompt Engineering) 전략 반영 여부	{AI모델설계서} 또는 'AI 서비스 요건 정의서' 내 프롬프트 설계 전략 기술 여부 확인 (세부: 시스템 프롬프트 구조, 컨텍스트 관리, Few-shot 예시 등)

No	점검 항목	확인 방법
4	환각(Hallucination) 및 편향(Bias) 대응 설계 반영 여부	{AI모델설계서} 또는 'AI 위험성 평가서' 내 해당 항목 기술 여부 확인
5	외부 AI 모델 또는 API 활용 시 근거·제약사항 명시	{AI모델설계서} 및 계약·조달 관련 문서 교차 검토 (세부: 출처, 버전, 라이선스, 토큰 한도, SLA 등)
6	요구사항 ↔ 설계 항목 간 추적성 확보 여부	{요구사항정의서} ↔ {AI모델설계서} 간 추적성 매트릭스 교차 검토 (세부: 추적성 매트릭스)
7	AI 출력 품질 기준(측정 가능한 검증 지표) 설계 반영 여부	{AI모델설계서} 또는 'AI 서비스 요건 정의서' 내 품질 지표 기술 여부 확인 (세부: 정확도, F1 Score, BLEU, 응답 거부율 등)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	모델 유형 선정 근거, 처리 흐름 다이어그램, 프롬프트 전략, 환각·편향 대응 방안, 외부 API 제약사항, 측정 가능한 품질 지표가 {AI모델설계서}에 모두 구체적으로 반영되고 요구사항과의 추적성이 확보되어 있는 경우	현행 유지
보완권고 (△)	설계서는 작성되었으나 모델 선정 근거 불명확, 품질 지표의 수치화 미흡, 추적성 매트릭스 일부 누락 등 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{AI모델설계서}가 미작성되었거나 모델 아키텍처·리스크 대응 등 핵심 설계 항목이 반영되지 않아 요구사항의 설계 반영을 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 품질 요건의 수치화 유도: "AI는 정확한 답변을 제공해야 한다"와 같은 표현은 시험·검수 단계에서 분쟁의 원인이 된다. 감리원은 "응답 정확도 90% 이상", "금칙어 탐지율 95% 이상", "응답 거부(Refusal) 시나리오 N건 이상 반영" 등 측정 가능한 기준으로 구체화하도록 사업자에게 시정을 요구하여야 한다.

2) 외부 AI API 사용 시 계약 리스크 조기 확인: OpenAI, Google Gemini, Claude API 등 외부 AI 서비스를 활용하는 사업에서는, 해당 API의 이용약관상 공공기관 데이터 전송 허용 여부, 데이터 학습 활용 금지 조항, API 가용성(Uptime) SLA 등이 확인되었는지를 점검해야 한다. 계약 체결 후 API 정책 변경으로 인한 서비스 중단 리스크가 존재한다.

3) 설계서 ≠ 개발 명세서 구분: 설계서가 소스코드 수준의 상세 명세일 필요는 없으나, 모델 유형·처리 흐름·품질 기준 등 감리 검증이 가능한 수준의 구체성은 갖추어야 한다. 형식 충족 여부에 그치지 말고 실질적 검증 가능성을 기준으로 점검한다.

AG_SD42-2-16 | Agentic안전정지Loop통제점검

응용시스템 | 설계 | 인공지능 모델 개발 | 오케스트레이션

근거 | 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · NIST AI RMF

1. 목적

본 점검 가이드라인은 금융권 Agentic AI 서비스의 설계 단계에서, Agent의 자율 추론에 따른 무한 루프·권한 우회·시스템 오류 상황에서 프로세스를 강제 중단하고 안전 상태(Fail-safe)로 복귀시키는 안전정지·Loop 통제 메커니즘이 적절히 설계되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 'Agent 오케스트레이션 설계서'와 {응용시스템설계서}, 예외처리 설계 문서를 근거로 ① 툴 호출 Loop Lock(Max Iteration) 설계, ② userId와 caseId의 정합성 검증을 통한 권한 우회 차단 설계, ③ 오류·타임아웃 시 안전 상태 복귀(Safe Stop) 설계, ④ 반복 오류·루프 발생 시 관리자 Alert 설계를 확인한다.

감리원의 점검은 Agent 추론 알고리즘의 성능이 아니라 통제 메커니즘의 설계 완결성에 한정한다.

2. 필요성

▶ 무한 루프 차단: Agent가 동일 도구를 반복 호출하는 무한 루프는 컴퓨팅 자원 고갈과 외부 API 호출 비용 폭증, 서비스 장애로 직결된다. 프롬프트 수준의 지시가 아니라 오케스트레이션 엔진 레벨에서 호출 횟수를 강제 제한해야 통제로 인정된다.

▶ 권한 우회 방지: 프롬프트 인젝션으로 타인의 계좌·상담 이력에 접근하려는 시도는 Agent가 도구를 호출하는 시점에 차단되어야 한다. 세션의 userId와 요청 caseId의 정합성을 검증하는 설계가 없으면 LLM 추론만으로는 방어할 수 없다.

▶ 안전 상태 복귀: 외부 API 오류, 데이터 부재, 타임아웃 상황에서 Agent가 임의로 추론하여 응답을 생성하면 사실과 다른 정보가 고객에게 전달된다. 오류 유형별로 추론을 중단하고 표준 안내 메시지와 함께 안전 상태로 복귀하는 설계가 필요하다.

▶ 운영 대응 체계: 반복 오류나 루프 발생을 관리자가 사후에 로그로만 인지하는 구조는 금융 상담 서비스의 신뢰성을 훼손한다. 임계 초과 시 즉시 Alert가 발생하고 대응 담당자가 지정되어야 한다.

▶ 설계 단계 확정의 필요성: 안전정지 통제는 오케스트레이션 엔진의 실행 흐름에 내재되어야 하므로, 구현 이후에 추가하려면 전체 흐름의 재설계가 불가피하다. 통제 지점과 발동 조건을 확정해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	툴 호출 Max Iteration 제한 설계	'Agent 오케스트레이션 설계서'에서 툴 호출 최대 횟수(예: 3회)가 수치로 정의되고, 프롬프트 지시가 아닌 엔진 레벨에서 강제되도록 설계되었는지 확인
2	무의미 반복 호출 탐지·차단 설계	동일 파라미터의 반복 호출이나 진전 없는 호출 패턴을 루프로 판정하여 차단하는 조건이 설계서에 정의되었는지 확인하고, 판정 임계값을 면담으로 확인
3	도구 호출·세션 시간 한도 설계	개별 도구 호출 타임아웃과 세션 전체 시간 한도가 'Agent 오케스트레이션 설계서'에 수치로 정의되고, 초과 시 처리 방식이 설계되었는지 확인
4	userId-caseId 정합성 검증을 통한 권한 우회 차단	Agent가 도구를 호출할 때 세션의 userId와 요청 caseId의 소유 관계를 검증하여 불일치 시 호출을 차단하는 설계가 {응용시스템설계서}에 반영되었는지 확인
5	오류·타임아웃 시 안전 상태 복귀(Safe Stop) 설계	외부 API 오류, 데이터 부재, 타임아웃 등 오류 유형별로 임의 추론을 차단하고 표준 안내 메시지와 함께 안전 상태로 복귀하는 절차가 '예외처리 설계서'에 정의되었는지 확인
6	루프·반복 오류 발생 시 관리자 Alert 설계	루프 판정 또는 반복 오류 임계 초과 시 관리자에게 Alert가 발송되도록 'Agent 오케스트레이션 설계서'에 설계되었는지 확인하고, 수신자와 대응 절차가 지정되었는지 면담으로 확인
7	안전정지 발동 이력의 기록 설계	안전정지 발동 시각·발동 조건·중단된 도구 호출·복귀 결과가 로그 항목으로 설계되어 사후 원인 분석이 가능한지 {프로그램 명세서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	Max Iteration·무의미 반복 탐지·시간 한도의 Loop Lock, userId-caseId 정합성 검증을 통한 권한 우회 차단, 오류 유형별 Safe Stop 복귀, 관리자 Alert와 발동 이력 기록이 설계서에 구체적으로 명시된 경우	현행 유지
보완권고 (△)	Fail-safe 복귀 설계는 존재하나 Loop Lock이 프롬프트 지시 수준에 그치거나, 권한 우회 탐지·Alert·발동 이력 기록 중 일부가 누락된 경우	Agentic 통제 보강 요청 후 재점검
부적합 (X)	무한 루프 차단과 권한 우회 차단 메커니즘이 설계에 부재하거나, 오류 발생 시 안전 상태 복귀 절차가 정의되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 멈출 수 있는 Agent: Agent는 '멈출 줄 아는가'가 핵심이다. 무한 루프와 권한 우회를 엔진이 강제로 끊지 못하면 어떤 통제도 우회될 수 있다.

2) Max Iteration·시간 한도 권고: 툴 호출 Max Iteration은 3~10회, 세션 시간 한도 설정을 권고하며, 운영자에게 무한 루프 발생 시 강제 중단을 시연으로 확인한다.

3) 권한 우회 대응: 권한 우회 시도는 Deny 처리 후 보안 로그·Alert를 남기고, userId↔caseId 정합성 검증 없이 타인 데이터에 접근 가능한 설계가 없는지 개발자의 Context Mismatch 거부 동작으로 확인한다.

4) 표준 안전 메시지 강제: API 오류 시 LLM이 임의 값을 지어내 응답하지 않도록 '시스템 오류로 정보를 확인할 수 없습니다' 등 표준 메시지를 강제한다.

5) 연계 정합성 점검: 안전정지·루프 통제가 Orchestration 중단 분기, 가드레일 Tool Call 검증, Circuit Breaker(레거시 부하 차단)와 정합하게 연계되는지 함께 점검한다.

AG_SD43-2-01 | 튜토리얼 및 사용자가이드 작동 점검

응용시스템 | 구현 | 시스템 구현 | UI/UX, 접근성

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조 · 「인공지능기본법 이행 가이드라인」(과기정통부, 2026.1.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

AI 서비스의 종료단계(시험/전개)에서 최초 로그인 또는 서비스 최초 접속 시 제공되는 튜토리얼·안내 팝업이 {화면설계서}에 설계된 내용대로 {시스템화면}에 정확히 구현되어 있는지, 그리고 AI 서비스 이용에 필요한 사용자 가이드(사용 방법·주의사항·AI 응답 한계 안내 등)가 적정하게 제공되는지를 점검하는 것이 목적이다.

특히, 안내 팝업의 노출 조건, 내용 적정성, 재노출 통제 메커니즘이 설계 의도대로 구현·시험되었는지를 확인한다.

2. 필요성

▶ AI 응답 오인: AI 서비스는 응답의 불확실성, 환각(Hallucination) 가능성, 서비스 범위의 한계 등 사용자가 반드시 인지해야 할 특수한 이용 특성이 있으며, 안내가 없으면 사용자가 AI 응답을 사실로 오인하여 잘못된 의사결정을 내리는 피해가 발생할 수 있다.

▶ 서비스 범위 미인지: 안내 팝업 없이 서비스를 제공하면 사용자가 지원하지 않는 업무를 AI에 요청하고 응답 품질에 불만을 가지는 상황이 반복된다.

▶ 재노출 오작동: 안내 팝업이 매 로그인 시마다 반복 노출되거나 반대로 최초 접속 시에도 노출되지 않는 구현 오류는 사용자 불편과 서비스 신뢰도 저하로 이어진다.

▶ AI 고지 미흡: AI 서비스임을 명확히 알리는 안내가 누락되면 「인공지능 관련 법령·가이드라인」상 의무 고지 요건을 충족하지 못할 수 있다.

▶ 가이드 내용 부정확: 사용자 가이드가 실제 구현된 AI 기능과 불일치할 경우 사용자 혼란 및 민원이 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	최초 로그인·접속 시 안내 팝업이 설계서대로 노출되는지	{시스템화면}에서 최초 로그인·접속 시 안내 팝업이 {화면설계서}에 정의된 시점과 형태로 노출되는지 직접 확인
2	재로그인·재접속 시 팝업 노출 제어가 설계서대로 작동하는지	재로그인·재접속 시 팝업 노출 제어가 설계서대로 작동하는지 {테스트결과서}의 시험 결과로 확인

No	점검 항목	확인 방법
3	팝업 노출 조건 관리 메커니즘(플래그 저장 방식 등)이 설계서에 명시되어 있는지	팝업 노출 조건 관리 메커니즘(플래그 저장 위치·만료 조건)이 설계서에 명시되었는지 확인
4	팝업 내 AI 신원 고지 및 응답 한계 주의 문구가 포함되어 있는지	팝업 내 AI 서비스임을 알리는 신원 고지와 응답 한계·환각 가능성 주의 문구가 포함되었는지 화면에서 직접 확인
5	팝업 내 서비스 제공 범위 및 개인정보 입력 금지 안내가 포함되어 있는지	팝업 내 서비스 제공 범위 안내와 개인정보 입력 금지 안내가 포함되었는지 화면에서 직접 확인
6	팝업 내용이 {화면설계서}에 정의된 내용과 일치하는지	팝업의 문구·버튼·레이아웃이 {화면설계서}에 정의된 내용과 일치하는지 문서와 화면을 대조
7	버튼('확인', '다시 보지 않기' 등) 기능이 정상 작동하는지	'확인', '다시 보지 않기' 등 버튼의 기능이 설계대로 작동하는지 화면 조작과 {테스트결과서}로 확인
8	지원 디바이스·브라우저 환경별 팝업 표시가 정상인지	지원 대상 디바이스·브라우저 환경별 팝업 표시 결과가 {테스트결과서}에 환경별로 기재되었는지 확인
9	사용자 가이드 내용이 실제 시스템 기능과 일치하는지	사용자 가이드에 기술된 기능·화면이 실제 구현된 시스템과 일치하는지 가이드와 화면을 대조
10	사용자 가이드에 버전·작성일자가 표기되어 있고 최신 내용을 반영하는지	사용자 가이드에 버전과 작성일자가 표기되고 최신 배포 기능이 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	안내 팝업이 최초 로그인 시 설계대로 노출되고 팝업 노출 조건 관리 메커니즘이 설계서에 명시되며, AI 특성 고지 및 서비스 범위 안내가 포함되고, {테스트결과서}에 노출 조건·버튼 작동·환경별 시험이 합격 판정되며, 사용자 가이드가 시스템과 일치하고 버전·작성일자가 표기되어 최신 내용을 반영하는 경우	현행 유지
보완권고 (△)	팝업 내 AI 응답 한계 문구 누락, 재로그인 시 팝업 재노출 오류, 모바일 환경 시험 미수행, 가이드 내 일부 기능 설명 불일치 등 부분적 미흡이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	최초 로그인 시 안내 팝업이 전혀 노출되지 않거나 AI 신원 고지 문구가 전면 누락된 경우, {화면설계서}와 구현 화면 간 내용이 전면 불일치하거나 {테스트결과서}에 관련 시험이 전혀 수행되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 최초 로그인 시험은 '신규 계정'으로 반드시 확인: 피감리인이 기존 계정으로 시연하면 이미 팝업 확인 이력이 있어 팝업이 노출되지 않을 수 있으므로, 반드시 신규 생성 계정(또는 팝업 확인 이력이 초기화된 계정)으로 최초 로그인 시나리오를 시연받고, 가능하면 시연 전 테스트 계정 정보를 미리 요청한다.

2) 팝업 내 AI 특성 고지 문구의 구체성 확인: 'AI 서비스를 이용해 주셔서 감사합니다'와 같이 AI 응답의 불확실성에 대한 구체적 안내가 없는 팝업은 고지 요건을 충족하지 못하므로, '답변은 부정확할 수 있으며 중요한 사항은 담당자에게 확인하십시오'처럼 AI 한계를 명확히 전달하는 문구가 포함되어 있는지 확인한다.

3) 다시 보지 않기 기능의 지속성 확인: '다시 보지 않기' 버튼 클릭 후 브라우저를 완전히 닫고 재접속하거나 다른 기기에서 로그인했을 때도 팝업이 재노출되지 않는지 확인하고, 쿠키 기반 구현은 브라우저 초기화 시 재노출될 수 있으므로 서버 DB 기반 관리 여부를 설계서에서 확인한다.

4) 사용자 가이드의 AI 응답 특성 반영 여부 우선 확인: 일반 IT 서비스 가이드를 그대로 전용하여 AI 서비스 특성(환각, 응답 변동성, 최신 정보 미반영 등) 안내가 빠진 경우가 자주 발생하므로, AI 서비스 고유 항목이 별도 섹션이나 주의사항으로 포함되어 있는지를 먼저 확인한다. 아울러 사용자가 차별적·유해한 답변을 즉시 보고할 수 있는 신고·피드백 채널의 위치와 사용 방법이 가이드에 안내되어 있는지 점검한다.

5) 화면 캡처 또는 시연 동영상을 점검 증거자료로 확보: 팝업 작동 여부는 {테스트결과서}의 텍스트 기록만으로 사후 입증이 어려우므로, 최초 로그인 시 팝업 노출 화면, 버튼 클릭 후 화면 전환, 재로그인 시 미노출 화면을 캡처 이미지 또는 시연 동영상으로 확보하여 감리 증거자료로 활용한다.

AG_SD43-2-02 | 마스킹 비식별화 뷰 정상 조회 점검

응용시스템 | 구현 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」·「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF ·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 학습 데이터 및 AI 서비스 데이터에 포함된 개인정보·민감정보에 대한 마스킹(Masking) 및 비식별화(De-identification) 처리가 설계대로 구현되어 정상 동작하는지를 {시스템화면}을 통해 확인하고, 특히 권한이 없는 사용자가 실데이터를 조회할 때 마스킹이 적용되는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 대량의 개인정보를 학습·추론에 활용하므로, 일반 정보시스템 이상으로 엄격한 마스킹/비식별화 체계가 요구된다.

권한별 데이터 접근 통제가 시스템 화면 수준에서 정상 동작하는지, 마스킹 규칙이 설계서 정의와 일치하는지를 검증한다.

2. 필요성

▶ AI 서비스는 학습 데이터 수집·가공·활용 전 과정에서 개인정보를 처리하며, 모델 추론 결과에도 개인정보가 포함될 수 있다. 마스킹/비식별화가 적절히 구현되지 않으면 개인정보 유출 사고로 이어질 수 있으며, 이는 개인정보 보호법 위반 및 AI 시스템에 대한 사회적 신뢰 훼손을 초래한다.

▶ 개인정보 보호법 준수: AI 시스템에서 처리하는 개인정보(이름, 주민등록번호, 전화번호, 주소 등)는 접근 권한이 없는 사용자에게 노출되어서는 안 되며, 마스킹/비식별화는 이를 기술적으로 보장하는 핵심 통제 수단이다.

▶ AI 학습 데이터 보호: 학습에 사용된 원본 데이터에는 개인정보가 포함될 수 있으므로, 학습 데이터 조회·관리 화면에서도 권한에 따른 마스킹이 적용되어야 한다. 이는 TTA 개발안내서가 요구하는 프라이버시 보호 요건에 해당한다.

▶ 권한별 접근 통제 검증: 정의된 역할별 데이터 접근 권한(관리자/개발자/일반사용자/외부사용자 등)이 시스템 화면 수준에서 실제로 구현·동작하는지 시험을 통해 검증해야 한다.

▶ NIST AI RMF 프라이버시 요건 충족: NIST AI RMF는 프라이버시 보호(Privacy-Enhanced)를 AI 신뢰성의 핵심 특성으로 정의하고 있으며, 데이터 최소화(Minimization) 및 비식별화를 통한 프라이버시 리스크 관리를 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	마스킹 대상 항목 실제 적용 여부	이름·주민등록번호·전화번호·주소 등 설계서 정의 마스킹 대상 항목이 권한 없는 계정 로그인 시 마스킹된 상태로 조회되는지 {시스템화면}에서 직접 확인
2	마스킹 규칙 설계 일치 여부	마스킹 문자·위치·노출 자릿수 등 마스킹 규칙이 '개인정보 마스킹 설계서' 정의와 일치하는지 {시스템화면} 표시와 대조
3	권한별 차등 마스킹 적용 여부	관리자·일반사용자·외부사용자 등 권한별 마스킹 적용 수준이 차등 적용되는지 각 권한 계정으로 동일 데이터를 조회하여 비교
4	학습 데이터 화면 마스킹 적용 여부	학습 데이터 조회·관리 화면에서도 마스킹이 적용되는지 {시스템화면} 학습 데이터 관리 메뉴에서 확인
5	마스킹 우회 불가 여부	URL 직접 접근·API 호출 등 마스킹 우회(Bypass) 가능성이 없는지 {테스트결과서} 마스킹 우회 시험 결과에서 확인
6	AI 추론 결과 화면 마스킹 적용 여부	AI 모델 추론 결과 화면에 개인정보 포함 시 마스킹이 적용되는지 {시스템화면} 추론 결과 조회 화면에서 확인
7	마스킹·비식별화 요구사항 추적 가능 여부	{요구사항추적표}의 마스킹·비식별화 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 마스킹 대상 항목·규칙이 학습 데이터 및 추론 결과 화면을 포함하여 모두 적용되고, 권한별 차등 마스킹과 우회 차단이 시험으로 확인되는 경우	현행 유지
보완권고 (△)	마스킹은 구현되었으나 일부 항목·화면 누락, 권한별 차등 미적용, 우회(URL 직접 접근·API 호출) 시험 미수행 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	마스킹/비식별화가 미구현되어 권한 없는 사용자에게 개인정보 원문이 노출되는 경우 (개인정보 보호법 위반 소지)	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 애플리케이션 화면 확인에 더해 '데이터베이스 레벨의 동적 데이터 마스킹(DDM) 또는 뷰(View) DDL'을 직접 검증하라: 화면에서 마스킹되어 보인다고 해서 DB 레벨에서 안전한 것은 아니다.

감리원은 애플리케이션 소스 레벨의 마스킹에 그치지 말고, 데이터베이스 자체 기능(예: Oracle DDM, SQL Server Dynamic Data Masking) 또는 뷰(View) DDL로 마스킹이 구현되었는지 직접 검증한다.

2) 2개 이상 권한 계정으로 '동일 데이터'를 비교 조회하라: 마스킹의 권한별 차등 적용을 확인하는 가장 확실한 방법은, 관리자 계정과 일반사용자 계정으로 동일한 데이터를 조회하여 마스킹 수준의 차이를 직접 비교하는 것이다. 사업자에게 시연 계정 2개를 요청한다.

3) '다운로드/내보내기' 시 마스킹 유지 여부를 확인하라: 화면에서는 마스킹이 적용되더라도, 엑셀 다운로드·CSV 내보내기 시 마스킹이 해제되어 원본 데이터가 노출되는 사례가 있다. 데이터 내보내기 기능이 있는 화면에서는 반드시 내보내기 파일의 마스킹 유지 여부를 확인한다.

4) 시스템 접근이 불가한 경우 {테스트결과서} 화면 캡처를 활용하라: 현장감리 시 시스템에 직접 접근할 수 없는 경우, {테스트결과서}에 각 권한별 조회 화면 캡처가 포함되어 있는지 확인하고, 캡처 화면에서 마스킹 대상 항목·규칙을 대조한다. 캡처가 없으면 사업자에게 시연을 요청한다.

5) 마스킹 설계서가 없는 경우의 대응: '개인정보 마스킹 설계서'가 별도 산출물로 작성되지 않은 사업의 경우, '데이터베이스 설계서'의 뷰(View) 설계 항목이나 '보안 설계서'의 접근통제 항목에서 마스킹 규칙을 확인할 수 있다. 어떠한 설계 근거도 없이 마스킹이 구현되지 않은 경우, 개인정보 보호법 위반 가능성이 있으므로 부적합으로 판정할 수 있다.

AG_SD43-2-03 | 프롬프트 엔지니어링 관리체계 점검

응용시스템 | 구현 | 인공지능 모델 개발 | 프롬프트설계

근거 | OWASP LLM Top 10 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)

1. 목적

본 점검 가이드라인은 LLM 기반 공공 AI 서비스의 구현 단계에서, 프롬프트의 설계·버전 관리·보안·성능 검증이 체계적으로 운영되고 있는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 '프롬프트 관리 대장(인벤토리)', {프로그램 명세서}, 형상관리 저장소와 '프롬프트 평가 결과서'를 근거로 ① 시스템 프롬프트의 역할·제약·출력 형식 명시, ② 프롬프트 버전 관리와 변경 승인 절차, ③ 시스템 프롬프트 유출·인젝션 방어 가드레일, ④ 프롬프트 변경 시 성능 회귀 평가 체계를 확인한다.

감리원의 점검은 프롬프트 문구의 우수성을 평가하는 것이 아니라 관리 체계의 존재와 운영 증빙의 완결성에 한정한다.

2. 필요성

▶ LLM 서비스의 품질은 프롬프트 설계에 크게 좌우되며, 체계 없는 프롬프트 변경은 동일 질의에 대한 응답 품질을 예고 없이 바꾼다. 변경 이력과 승인 절차가 없으면 품질 저하의 원인을 사후에 규명할 수 없다.

▶ 시스템 프롬프트에는 운영 정보, API 키, 업무 처리 규칙이 포함되는 경우가 많다. 이 내용이 사용자 입력으로 유출되면 서비스의 통제 로직 전체가 공격자에게 노출된다.

▶ OWASP LLM Top 10은 프롬프트 인젝션(LLM01)과 시스템 프롬프트 유출(LLM07)을 주요 위험으로 명시하며, 입력 필터링과 출력 검증을 통한 다층 방어를 요구한다.

▶ TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 프롬프트 관리 체계의 수립과 문서화를 요구하며, 프롬프트를 형상관리 대상 산출물로 취급할 것을 권고한다.

▶ 프롬프트는 코드와 달리 배포 절차 없이 변경 가능한 경우가 많아 통제 밖에서 수정될 위험이 크다. 변경 시 성능 회귀 평가를 강제하지 않으면 운영 중 품질 저하가 누적된다.

3. 검토방법

No	점검 항목	확인 방법
1	시스템 프롬프트의 역할·제약·출력 형식 명시	'프롬프트 관리 대장'에 등록된 시스템 프롬프트에서 역할(Role)·금지 사항·출력 형식이 명시되었는지 확인하고, 담당자 면담으로 실제 운영 적용 버전과 대조
2	Few-shot 예시의 도메인 적합성	'프롬프트 관리 대장'에 등록된 Few-shot 예시가 도메인 특화 사례 3~5개를 포함하고 실제 업무 사례를 반영하는지 확인
3	구조화 출력(JSON·XML) 스키마 명세 및 검증	{프로그램 명세서}에서 구조화 출력 스키마가 정의되고, 응답이 스키마를 위반할 때의 검증·재시도 로직이 구현되었는지 확인
4	프롬프트 인벤토리 및 버전·변경 승인 관리	형상관리 저장소에서 프롬프트가 버전 태그와 함께 관리되는지 확인하고, 변경 시 검토·승인 이력이 남는지 최근 변경 건으로 확인
5	프롬프트 인젝션·시스템 프롬프트 유출 방어 가드레일	입력 필터링과 출력 검증이 다층으로 구현되었는지 확인하고, 시스템 프롬프트 노출 시도에 대한 차단 시험 결과를 확인
6	프롬프트 변경 시 성능 회귀 평가	프롬프트 변경 전후로 표준 평가셋 기반 회귀 평가가 수행되고, '프롬프트 평가 결과서'에 정확도·거절률 등 지표의 변화가 수치로 기재되었는지 확인
7	프롬프트 운영 거버넌스(권한·배포 절차)	프롬프트 수정·배포 권한이 지정된 담당자로 제한되고, 운영 반영 절차가 '운영 절차서'에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계·보안·평가·운영 4영역이 정량 기준과 함께 체계화되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	설계·평가는 수행했으나 보안 가드레일 또는 운영 거버넌스가 일부 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	프롬프트 인벤토리가 부재하고 보안·평가가 미수행되는 등 요건이 미반영되었거나 단순 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 다층 방어 적용: 프롬프트 인젝션은 '100% 방어 불가능'한 위험이므로 단일 가드레일에 의존하지 말고 다층 방어(Defense-in-Depth)를 적용해야 한다.

2) 민감정보 하드코딩 금지: 시스템 프롬프트에 API 키·내부 정책을 하드코딩하거나 Few-shot 예시에 실제 민감 정보를 포함하지 않도록 점검한다.

3) 프롬프트 길이 관리: 프롬프트 길이는 1000토큰 이하 압축을 권고하고, 운영 중 '프롬프트 길이 증가율' 모니터링으로 토큰 비용 비효율을 조기에 발견한다.

4) 예시 구성·구조화 출력 검증: Few-shot 예시는 3~5개로 다양성을 확보하고, 구조화 출력은 JSON Schema 검증을 필수로 적용한다.

5) 변경 절차 준수: 프롬프트 변경을 핫픽스로 운영에 즉시 적용하지 않도록 변경 승인 절차와 회귀 평가 체계(정확도·환각률·일관성·편향·거부율 정량 측정)를 확인한다.

AG_SD43-2-04 | 캐싱 히트율 및 속도개선

응용시스템 | 구현 | 시스템 구현 | RAG파이프라인

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

AI 서비스의 종료단계(시험/전개)에서 동일하거나 빈출되는 사용자 질의에 대해 캐시 서버(Redis 등)를 통한 응답이 설계서에 따라 적정하게 구현되었는지를 확인하는 것이 목적이다.

특히, 캐싱 히트율(Cache Hit Rate) 목표치 달성 여부와 캐싱 적용으로 인한 응답 속도 개선 효과가 {성능테스트결과서}에 측정·기록되어 있는지, 그리고 캐싱 대상 질의의 선정 기준과 갱신(TTL) 정책이 설계 의도대로 구현되었는지를 점검한다.

2. 필요성

▶ 응답 지연 심화: AI 서비스는 LLM 기반 추론 처리 특성상 캐시 미적용 시 동일 질의에도 매번 AI 추론 호출이 발생하여 응답 시간이 수 초에 달할 수 있다.

▶ 비용 과다 발생: LLM API 호출 기반 서비스의 경우 불필요한 추론 호출로 운영 비용이 급증할 가능성이 있다.

▶ 캐시 오염(Cache Pollution): 캐싱 대상 질의 선정 기준이 불명확하면 비빈출 질의가 캐시를 점유하여 히트율이 저하된다.

▶ TTL 정책 미흡: 유효기간 없이 캐시를 유지할 경우 오래된 정보가 사용자에게 제공되는 정확성 문제가 발생할 수 있다.

▶ 시험 항목 누락: 캐싱 효과를 측정하는 성능 시험이 {성능테스트결과서}에 포함되지 않으면 검수 근거가 부재하게 된다.

3. 검토방법

No	점검 항목	확인 방법
1	캐싱 대상 질의 선정 기준이 설계서에 수치로 정의되어 있는지	캐싱 대상 질의 선정 기준(빈출 횟수·집계 기간 등)이 설계서에 수치로 정의되었는지 확인. 기준 없이 '빈출 질의'로만 기술된 경우 캐시 오염 위험으로 미흡 판정
2	TTL(유효기간) 정책이 질의 유형별로 설계서에 정의되어 있는지	질의 유형별 TTL(유효기간) 정책이 설계서에 시간 단위로 정의되었는지 확인
3	캐시 히트율 목표값이 비기능 요구사항에 수치로 정의되어 있는지	캐시 히트율 목표값이 비기능 요구사항에 수치(%)로 정의되었는지 확인

No	점검 항목	확인 방법
4	{성능테스트결과서}에 캐시 히트율 측정값이 기록되어 있는지	{성능테스트결과서}에 캐시 히트율 실측값이 기록되고 목표값 충족 여부가 판정되었는지 확인
5	캐시 적중/미적중 시 응답 시간이 각각 측정·비교되어 있는지	캐시 적중 시와 미적중 시의 응답 시간이 각각 측정되어 개선 효과가 정량 비교되었는지 확인
6	캐시 서버 구성이 설계서와 일치하게 구현되었는지	캐시 서버 구성(노드 수·메모리·이중화)이 설계서와 일치하게 구현되었는지 설계서와 운영 모니터링 화면을 대조
7	캐시 키 설계 및 질의 정규화 방식이 설계서에 명시되어 있는지	캐시 키 설계와 질의 정규화 방식(대소문자·공백·동의어 처리)이 설계서에 명시되었는지 확인
8	캐시 무효화(Invalidation) 정책이 설계서에 정의되어 있는지	원본 데이터 변경 시 캐시 무효화(Invalidation) 시점과 방식이 설계서에 정의되었는지 확인
9	개인정보 포함 응답이 캐싱 제외 대상으로 설계서에 명시되어 있는지	개인정보·개인화 응답이 캐싱 제외 대상으로 설계서에 명시되었는지 확인. 미명시 시 타 사용자 응답 노출 위험으로 보완권고
10	{테스트결과서}에 캐싱 관련 시험 항목이 포함되어 있는지	{테스트결과서}에 캐싱 관련 시험 항목이 독립 항목으로 포함되고 합격 판정되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	캐싱 대상 질의 선정 기준·TTL 정책·히트율 목표값과 캐시 키 설계·질의 정규화 방식·무효화(Invalidation) 정책이 설계서에 명시되고, 개인정보·개인화 응답이 캐싱 제외 대상으로 지정되며, {성능테스트결과서}에 히트율 목표치 달성과 캐시 적중·미적중 응답시간의 정량 비교가 기재되고, {테스트결과서}에 캐싱 시험 항목이 독립 항목으로 포함되어 합격 판정되며, 캐시 서버 구성의 설계-구현이 일치하는 경우	현행 유지
보완권고 (△)	히트율 목표값이 수치로 미정의되거나, 캐시 적중/미적중 응답 시간 중 일부 미측정, TTL 정책이 일부 질의 유형에 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	캐싱 설계서 미작성, {성능테스트결과서}에 캐싱 관련 측정값 전면 누락, 개인정보 포함 응답이 캐시에 무분별하게 저장되는 구조인 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 캐시 히트율 목표값의 수치화 유도: '캐시 히트율을 최대한 높인다'와 같은 모호한 표현은 분쟁의 원인이 되므로, '빈출 질의 기준 캐시 히트율 70% 이상 달성', 'Top 100 질의 캐시 응답 시간 100ms 이내' 등 측정 가능한 수치로 정의하도록 현장에서 보완을 유도한다.

2) 시험 환경과 운영 환경의 차이 확인: {성능테스트결과서}의 시험 환경(서버 사양, 사용자 수, 질의 패턴)이 실제 운영 환경과 크게 다를 경우 시험 결과의 대표성이 낮아지므로, 시험 조건 기술 내용을 확인하고 차이가 크면 점검 소견에 '시험 환경 대표성 재검토 필요'로 명시한다.

3) 캐싱과 AI 응답 정확성 간 균형 점검: 캐시 히트율을 높이기 위해 TTL을 지나치게 길게 설정하면 AI 모델 업데이트 이후에도 구버전 응답이 계속 제공될 수 있으므로, TTL 설계 시 모델 업데이트 주기와 서비스 특성을 고려한 합리적 유효기간이 설정되었는지 확인한다.

4) 개인정보 캐싱 여부는 반드시 점검(보안 영역 연계): 사용자별 맞춤 응답이 캐시에 저장되면 다른 사용자에게 해당 응답이 제공되는 심각한 보안 사고로 이어질 수 있으므로, 개인화 응답의 캐싱 제외 여부를 반드시 확인하고 캐싱 전 개인정보 마스킹(비식별화) 등 보호 조치는 보안 영역 감리원과 연계 검토한다.

5) 피감리인에게 모니터링 캡처 자료 사전 요청: Redis 모니터링 화면 또는 APM 도구의 캐시 히트율 그래프를 점검 시작 전 사전 요청하여, 실시간 접근이 어려운 환경에서도 캡처 이미지를 {성능테스트결과서}와 함께 점검 증거자료로 확보하고 측정 방법론(시험 도구·조건·반복 횟수)을 통해 수치의 신뢰성을 검증한다.

AG_SD43-2-05 | 비정형 문서 처리 AI 요건 점검

응용시스템 | 구현 | 인공지능 모델 개발 | 멀티모달/에이전트

근거 | 「장애인차별금지 및 권리구제 등에 관한 법률」·「지능정보화 기본법」·「개인정보 보호법」· KW CAG

1. 목적

본 점검 가이드라인은 AI를 활용하여 PDF·공문서·계약서·스캔 문서 등 비정형 문서를 인식·추출·분류·요약하는 기능이 정확성·보안·접근성 요건을 충족하는지를 감리원이 구현 단계에서 제3자적 관점으로 점검하는 것을 목적으로 한다.

구체적으로는 'AI 문서처리 성능 평가 결과서', {테스트결과서}, {응용시스템설계서}를 근거로 ① OCR 인식률과 표·레이아웃 추출 정확도의 임계값 충족, ② 개인정보·민감정보의 자동 식별(NER)과 마스킹, ③ 업로드 문서의 보안 통제, ④ 추출 결과의 접근성 요건을 확인한다.

감리원의 점검은 인식 모델의 알고리즘 타당성이 아니라 측정 결과의 임계값 충족과 증빙의 완결성에 한정한다.

2. 필요성

▶ 공공기관 문서 디지털화 확산으로 AI 기반 비정형 문서 처리 수요가 급증하고 있으며, 처리 결과가 행정 처분의 근거 자료로 직접 사용되는 사례가 늘고 있다.

▶ OCR·문서 분류·정보 추출의 정확도가 낮으면 행정 처리 오류와 민원 증가로 직결된다. 정확도는 '우수함'과 같은 정성 표현이 아니라 CER·WER 등 표준 지표의 수치로 검증되어야 한다.

▶ 문서에는 주민등록번호·계좌번호·건강정보 등 개인정보와 민감정보가 포함될 가능성이 높다. 추출 과정에서 자동 식별(NER)과 마스킹이 적용되지 않으면 원문보다 넓은 범위로 정보가 확산된다.

▶ 표·이미지·수기 서명 등 복잡 요소는 일반 텍스트 대비 인식 오류율이 현저히 높다. 전체 평균 정확도만 제시되면 이러한 요소의 오류가 은폐된다.

▶ 취약계층의 문서 접근성을 위해 추출 결과의 대체 텍스트 제공과 화면낭독기 호환이 요구되며,「장애인차별금지 및 권리구제 등에 관한 법률」및 지능정보화 기본법상 공공서비스 접근성 의무와 연계된다.

3. 검토방법

No	점검 항목	확인 방법
1	문자 인식률(CER) 임계값 충족	'AI 문서처리 성능 평가 결과서'에서 Character Error Rate 측정값이 권고 임계값(5% 이하)을 충족하는지 확인하고, 평가 데이터셋의 구성과 표본 수를 함께 확인
2	단어 인식률(WER) 임계값 충족	'AI 문서처리 성능 평가 결과서'의 Word Error Rate 측정값이 권고 임계값(10% 이하)을 충족하는지 확인하고, 문서 유형별(인쇄·스캔·수기)로 구분 측정되었는지 확인
3	표 추출 정확도	'AI 문서처리 성능 평가 결과서'의 표 추출 셀 단위 일치율이 권고 임계값(90% 이상)을 충족하는지 확인하고, 병합 셀·다단 표 등 복잡 구조의 시험 포함 여부를 확인
4	레이아웃 분석 정확도	'AI 문서처리 성능 평가 결과서'의 레이아웃 영역 분류 일치율이 권고 임계값(85% 이상)을 충족하는지 확인하고, 이미지·서명·도장 영역의 처리 결과를 확인
5	개인정보·민감정보 자동 식별(NER) 및 마스킹	추출 결과에서 주민등록번호·계좌번호 등 개인정보가 자동 식별되어 마스킹되는지 {테스트결과서}의 시험 결과로 확인하고, 미탐·오탐률이 수치로 기재되었는지 확인
6	업로드 문서의 보안 통제	{응용시스템설계서}에서 업로드 문서의 저장 위치·암호화·보존 기간과 처리 후 파기 절차가 정의되었는지 확인
7	추출 결과의 접근성 요건	추출 결과에 대체 텍스트가 제공되고 화면낭독기 호환이 시험되었는지 {테스트결과서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	OCR·NER·보안·접근성 4영역이 정량 임계값과 함께 검증되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	주요 정확도는 측정했으나 보안 또는 접근성 요건이 일부 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	OCR 정확도만 측정하고 NER·보안이 미검증되는 등 요건이 미반영되었거나 단순 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 위험 가중치 반영 평가: OCR 정확도는 '평균 90%'가 '민감 정보는 100% 정탐 필요'와 충돌하므로, 위험 가중치를 반영한 평가가 필요하다.

2) 자체 평가 데이터셋 구축: 공급사 광고 정확도(98% 등)를 그대로 인용하지 말고, 발주기관 실문서 100~200건으로 자체 평가 데이터셋을 구축하여 검증한다.

3) 표 추출 별도 목표 설정: 행정 문서의 핵심인 표·차트는 별도로 '표 추출 정확도 95% 이상' 목표를 설정하고 측정 누락이 없는지 확인한다.

4) 미탐률 관리: 민감정보(PII) 마스킹은 정탐률뿐 아니라 '미탐률 1% 이하' 목표를 명시하고, Regex 단독이 아닌 Regex+NER 모델 하이브리드 마스킹과 정기 샘플 검수 절차를 확인한다.

5) 한국어 특화 OCR 우선 검토: 한국어 OCR은 Naver Clova OCR·Upstage Document AI 등 한국어 특화 솔루션을 우선 검토하고, 처리 실패 시 수동 입력 대체 절차를 점검한다.

AG_SD43-2-06 | AI 민원·상담 서비스 특화 요건 점검

응용시스템 | 구현 | 인공지능 모델 개발 | 챗봇

근거 | 「장애인차별금지 및 권리구제 등에 관한 법률」·「개인정보 보호법」·「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」제31조 ·「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 공공기관 AI 기반 민원 상담·챗봇 서비스가 정확한 정보 제공, 개인정보 보호, 민감 민원의 인간 상담원 연계, 취약계층 접근성 등 공공 서비스의 특수성을 충족하는지를 감리원이 구현 단계에서 제3자적 관점으로 점검하는 것을 목적으로 한다.

구체적으로는 'AI 상담 품질 평가 결과서', {테스트결과서}, {시스템화면}, '민원 라우팅 절차서'를 근거로 ① 민원 골든셋 기반 응답 정확도와 환각률, ② 근거 법령·문서의 표시와 정직한 거절 동작, ③ 민감 민원의 인간 상담원 연계, ④ 상담 이력의 개인정보 보호와 접근성 요건을 확인한다.

2. 필요성

▶ AI 챗봇의 24시간 민원 상담은 편의성을 제공하나, 오답이 행정 처리의 근거로 사용되면 민원인 피해와 기관 신뢰도 하락으로 직결된다. 정확도는 민원 골든셋 기반으로 정량 검증되어야 한다.

▶ 민원 내용에는 개인정보와 민감정보가 포함될 가능성이 높다. 대화 이력의 저장 범위와 마스킹, 보존 기간이 정의되지 않으면 상담 로그 자체가 최대 유출 지점이 된다.

▶ 복지·고충·재난 등 민감 민원은 AI의 응답만으로 종결되어서는 안 되며, 즉시 인간 상담원으로 연계되어야 한다. 연계 조건과 대기 시간이 정의되지 않으면 위기 상황 대응이 지연된다.

▶ AI가 확신이 없는 사안에 대해 그럴듯한 답을 생성하면 민원인은 이를 공식 안내로 오인한다. 근거 법령·문서를 표시하고, 확신이 없을 때는 정직하게 거절하는 동작이 요구된다.

▶ 취약계층의 이용 편의를 위해 음성 입출력, 다국어, 평이한 표현의 응답이 필요하며, 「장애인차별금지 및 권리구제 등에 관한 법률」상 공공서비스 접근성 의무와 연계된다.

3. 검토방법

No	점검 항목	확인 방법
1	민원 골든셋 기반 응답 정확도	'AI 상담 품질 평가 결과서'에서 민원 골든셋 정답률이 권고 임계값(85% 이상)을 충족하는지 확인하고, 골든셋의 구성·문항 수·갱신 이력을 함께 확인
2	환각률(잘못된 정보 제공률)	'AI 상담 품질 평가 결과서'의 잘못된 정보 제공률이 권고 임계값(3% 이하)을 충족하는지 확인하고, 환각 유형별(사실 왜곡·존재하지 않는 법령 인용 등) 측정 결과가 구분되었는지 확인
3	근거 법령·문서 표시율	{시스템화면}에서 응답과 함께 근거 법령·문서가 표시되는지 직접 확인하고, 표시율이 권고 임계값(90% 이상)을 충족하는지 평가 결과서로 확인

No	점검 항목	확인 방법
4	'모름' 응답의 적절성	검색된 근거가 없거나 확신이 낮을 때 임의 생성 대신 정직하게 거절하고 인간 상담원 연계를 안내하는지 {테스트결과서}의 시험 결과로 확인
5	민감 민원의 인간 상담원 연계	'민원 라우팅 절차서'에서 복지·고충·재난 등 민감 민원의 판별 기준과 연계 조건, 연계 대기 시간이 정의되었는지 확인하고 연계 동작을 {시스템화면}에서 확인
6	상담 이력의 개인정보 보호	대화 이력의 저장 범위와 개인정보 마스킹, 보존 기간·파기 절차가 정의되고 실제 적재 로그에 적용되는지 샘플로 확인
7	취약계층 접근성 지원	음성 입출력·다국어·평이한 표현의 응답 기능이 구현되고 화면낭독기 호환이 시험되었는지 {테스트결과서}와 {시스템화면}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	정확성·민감 민원 라우팅·개인정보 보호·접근성 4영역이 정량 기준과 함께 검증되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	정확성은 측정했으나 민감 민원 라우팅 또는 접근성 요건이 일부 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	위기 민원 라우팅이 부재하고 개인정보 보호가 미흡한 등 요건이 미반영되었거나 단순 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 정직한 '모름' 응답 우선: 민원 챗봇은 '답을 모르면 정직하게 모른다고 하는 것'이 '잘못된 답'보다 훨씬 낫다. 환각 방지·인간 연계 절차에 집중해야 한다.

2) 표준 안내 응답 설계: 응답 확신도가 낮으면 '담당 부서로 안내하겠다'는 표준 응답을 사용하도록 설계한다.

3) 위기 분류기 유지관리: 자살·자해·아동학대·가정폭력 등 위기 키워드 분류기는 정탐률뿐 아니라 미탐률도 측정하고, 분기 1회 재학습으로 정확도를 유지한다.

4) 에스컬레이션 인계 정보: 인간 상담원 전환 시 '민원 요약·이용자 의도·시도한 답변' 3종을 전달하여 민원인이 처음부터 다시 질문하지 않도록 한다.

5) 골든셋 지속 보강: 응답 만족도 조사 결과를 활용해 평가 골든셋을 지속 보강하고, 다국어 지원이 '번역기 통과' 수준에 그치지 않는지 확인한다.

AG_SD44-2-01 | 편향 모니터링 지표 정상 표출점검

응용시스템 | 시험 | 인공지능 모델 개발 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, {시스템화면}으로 제공되는 편향 모니터링 대시보드가 설계 의도에 따라 편향 정량 수치를 정확히 계산하고 정상적으로 표출하는지 여부를 제3자적 관점에서 검토하는 것을 목적으로 한다.

특히, 편향 지표가 화면에 표시되더라도 수치 계산 로직 오류·집단 분류 기준 불일치·갱신 주기 미반영 등으로 인해 실질적인 편향 감지 기능을 수행하지 못하는 경우가 있다.

감리원은 코드 분석이 아닌, 화면 표출 결과와 설계 산출물·시험 결과서 간 일치 여부를 중심으로 점검하여 배포 전 오류를 식별한다.

2. 필요성

▶ 편향 모니터링 지표 정상 표출 점검이 필요한 이유는 다음과 같다.

▶ 편향 지표의 시각적 표출과 실제 수치 정확성은 별개의 문제: 대시보드에 숫자가 표시된다고 해서 수치가 올바르게 계산되었음을 의미하지는 않는다. 집단 분류 오류, 분모 기준 불일치, 갱신 시점 불일치 등으로 인해 잘못된 값이 표시될 수 있으며, 이는 운영 단계에서 편향을 은폐하는 더 심각한 위험을 초래한다.

▶ 공공 AI 시스템의 법적·윤리적 책임 이행 확인 필요: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 배포 전 집단별 성능 차이(편향) 측정 및 모니터링 체계 구축을 필수 활동으로 규정하고 있다. 화면 표출 오류가 방치되면 발주기관이 편향 여부를 잘못 판단하는 의사결정 위험이 발생한다.

▶ 배포 후 시정 비용의 급격한 증가: 편향 모니터링 오류를 운영 단계에서 발견하면 대시보드 재설계, 데이터 소급 검증, 이해관계자 재보고 등 복합적인 비용이 발생한다. {시스템화면} 점검을 통해 조기에 발견하는 것이 비용 효율적이다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 설계 내용이 시스템으로 구현되었는지 여부를 점검 대상으로 규정하며, 편향 모니터링 대시보드의 {시스템화면} 구현 적합성은 이 점검 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	편향 지표 항목의 대시보드 표출 완전성	'편향 모니터링 기능 명세서'에 정의된 편향 지표 항목이 편향 모니터링 대시보드({시스템화면})에 누락 없이 표출되는지 '화면 설계서'와 대조
2	편향 수치 계산 방식 일치 여부	통계적 동등성 지수·균등화 오즈비 등 표출된 편향 수치가 '편향 모니터링 기능 명세서'의 계산 방식과 일치하는지 {통합테스트결과서}·'AI 시험 시나리오 및 결과서'에서 확인
3	집단 분류 기준 화면 반영 여부	성별·연령·지역 등 집단 분류 기준이 '편향 모니터링 기능 명세서'·'화면 설계서'와 동일하게 화면에 반영되었는지 {시스템화면}에서 대조
4	편향 수치 갱신 주기·갱신 시각 표시 여부	실시간·일별·주별 등 편향 수치 갱신 주기가 설계서 명시 주기와 일치하고 화면에 최종 갱신 시각이 표시되는지 {시스템화면}에서 확인
5	임계값 초과 시 경보 작동 여부	편향 임계값(허용 기준) 초과 시 경보(Alert) 또는 시각적 경고가 화면에 정상 작동하는지 {시스템화면} 테스트·{통합테스트결과서}에서 확인
6	편향 모니터링 AI 특화 시험 포함 여부	집단별 성능 차이 검증 등 편향 모니터링 AI 특화 시험 항목이 {단위시험 결과서}·{통합테스트결과서}에 포함되었는지 확인
7	편향 지표 데이터 원천 명시·표출 일치 여부	로그 DB·추론 결과 DB 등 편향 지표 데이터 원천이 '편향 모니터링 기능 명세서'에 명시되고 화면 표출 데이터와 일치하는지 {통합테스트결과서}·면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'편향 모니터링 기능 명세서'에 정의된 편향 지표가 {시스템화면}에 누락 없이 표출되고, 수치 계산 방식·집단 분류 기준·갱신 주기가 설계서와 일치하며, 임계값 초과 경보 동작과 시험 결과가 수치로 확인되는 경우	현행 유지
보완권고 (△)	편향 지표가 대체로 표출되나 일부 지표 누락, 집단 분류 레이블 불일치, 갱신 시각 미표시, 임계값이 정성 표현으로만 정의되는 등 부분적 미흡 사항이 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편향 모니터링 대시보드가 미구현되었거나, 표출 수치가 명세서의 계산 방식과 불일치하고 관련 시험이 수행되지 않아 편향 감지 기능을 신뢰할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "정상 표출"은 점검 완료가 아님을 명확히 인식: 대시보드에 숫자가 표시된다는 것만으로 편향 수치가 올바른 것은 아니다. {통합테스트결과서}에 기준값과 실측값의 수치 비교 결과가 명시되어 있는지를 반드시 확인하고, "정상", "이상 없음"만 기재된 경우 보완권고 의견을 제시한다.

2) 임계값 수치화 요구: "적절한 수준"은 부적합: '편향 모니터링 기능 명세서'에 편향 허용 기준이 "적절한 수준 유지" 등 정성 표현으로만 기재된 경우, 운영 단계에서 경보 기준 해석 분쟁의 원인이 된다. "통계적 동등성 차이 10%p 이하", "균등화 오즈비 0.8 이상" 등 측정 가능한 수치로 명시하도록 권고한다.

3) 집단 분류 기준의 설계-화면 일치 여부를 반드시 확인한다. 화면의 집단 분류 레이블이 명세서와 다르면 편향 수치 자체가 왜곡되어 해석 오류로 이어지므로, 불일치 발견 시 보완권고를 제시한다.

AG_SD44-2-02 | XAI 설명가능성 기능 동작확인 점검

응용시스템 | 시험 | 인공지능 모델 개발 | AI윤리/xAI

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, {시스템화면}으로 제공되는 XAI(설명가능성) 기능이 설계 의도에 따라 SHAP·LIME 등 설명 기법의 결과값을 사용자가 이해할 수 있는 형태로 정상 시각화·표출하는지를 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 설명 기법 내부 알고리즘 분석이 아니라, TTA 안내서(투명성 요건)에 따라 데이터 유형(텍스트·이미지·정형 등)에 맞는 적절한 XAI 기법(예: 이미지-GradCAM, 텍스트-LIME/SHAP)이 적용되었는지를 명세서와 시험결과서를 대조하여 확인하는 데 있다.

특히 XAI 시각화 화면이 단순히 존재하는 것을 넘어, 설명 결과값이 'XAI 기능 설계서'에 정의된 표출 방식과 일치하는지, 사용자 유형(일반 사용자·전문가·관리자)별 설명 수준이 적절히 구현되었는지, AI 판정 결과와 설명 내용이 논리적으로 일치하는지를 확인한다.

2. 필요성

▶ XAI 기능 동작 확인 점검이 필요한 이유는 다음과 같다.

▶ XAI 시각화 부재 또는 오류는 AI 투명성 요건의 형식적 이행을 의미: 설명가능성 기능이 요구사항에 포함되어 있더라도, 배포 단계에서 시각화 화면이 오작동하거나 설명 결과값이 AI 판정과 논리적으로 불일치하는 경우, AI 시스템의 투명성 요건을 실질적으로 이행하지 못한 것이다. 이는 발주기관의 법적·윤리적 책임 리스크로 직결된다.

▶ 공공 AI 시스템의 설명가능성 의무 이행 확인: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 AI 판정 결과에 대한 설명 제공을 필수 요건으로 규정하고 있으며, NIST AI RMF는 투명성(Transparency)을 핵심 리스크 관리 속성으로 정의한다. XAI 화면 동작 확인은 이 요건의 구현 완료 여부를 검증하는 필수 절차이다.

▶ 판정 결과와 설명 불일치 시 이용자 신뢰 훼손 위험: "승인"으로 판정되었지만 설명 화면에서는 주요 기여 요인이 부정적 방향으로 표시되는 경우, 이용자는 AI 시스템 전체를 신뢰할 수 없게 된다. 이러한 판정-설명 불일치는 발견 시 서비스 중단 사태로 이어질 수 있다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 설계 내용이 {시스템화면} 등 시스템으로 구현되었는지 여부를 점검 대상으로 규정하며, XAI 시각화 화면의 구현 적합성은 이 점검 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	XAI 표출 항목 구현 완전성	기여 요인·기여도 수치·시각화 차트 등 'XAI 기능 설계서'에 정의된 표출 항목이 {시스템화면}에 누락 없이 구현되었는지 '화면 설계서'와 대조
2	XAI 기법 결과값 표출 형식 일치 여부	채택된 XAI 기법(SHAP·LIME 등)의 결과값 표출 형식(막대·워터폴 차트·히트맵 등)이 'XAI 기능 설계서'·'화면 설계서'의 시각화 방식과 일치하는지 {시스템화면}에서 확인
3	AI 판정 결과와 XAI 설명 논리 일치 여부	AI 판정 결과(승인·거절·고위험 등)와 XAI 설명 내용(주요 기여 요인 방향·크기)이 논리적으로 일치하는지 {통합테스트결과서}·'AI 시험 시나리오 및 결과서'에서 확인
4	사용자 유형별 설명 수준 차등화 구현 여부	일반 사용자·전문가·관리자 등 사용자 유형별 설명 수준 차등화가 'XAI 기능 설계서' 정의대로 {시스템화면}에 구현되었는지 확인
5	XAI 응답 시간 내 표출·오류 메시지 표시 여부	XAI 결과값이 실시간 또는 설계된 응답 시간 내에 정상 표출되고 데이터 오류·누락 시 오류 메시지가 표시되는지 {시스템화면}·{통합테스트결과서}에서 확인
6	XAI 기능의 AI 특화 시험 포함 여부	판정-설명 일치 검증·특이 입력값 처리·사용자 유형별 표출 확인 등 XAI AI 특화 시험 항목이 {단위시험 결과서}·{통합테스트결과서}에 독립적으로 포함되었는지 확인
7	특이값·경계값 입력 시 정상 처리 여부	극단적 수치·결측값 포함 등 특이값·경계값 입력 시 XAI 설명 화면이 비정상 동작 없이 처리 결과를 표출하거나 안내 메시지를 제공하는지 'AI 시험 시나리오 및 결과서'와 테스트 환경 화면에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'XAI 기능 설계서'에 정의된 표출 항목·시각화 방식이 {시스템화면}에 누락 없이 구현되고, 판정-설명 일치성, 사용자 유형별 차등화, 설계된 응답 시간 내 표출과 데이터 오류·누락 시 오류 메시지 표시, 특이값 처리가 시험 결과로 확인되는 경우	현행 유지
보완권고 (△)	XAI 화면은 구현되었으나 일부 표출 항목 누락, 사용자 유형별 차등화 미구현, 시험 결과가 "정상 동작" 등 정성적 기재에 그치는 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	XAI 시각화 기능이 미구현되었거나 AI 판정 결과와 설명 내용이 논리적으로 불일치하는 등 설명가능성 요건을 실질적으로 이행하지 못한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) XAI 화면의 존재 여부와 기능 적정성은 별개로 구분하여 점검한다: (1) 표출 항목 완전성, (2) 판정-설명 일치성, (3) 사용자 유형별 차등화, (4) 특이값 처리 네 가지를 각각 독립적으로 점검하고, 항목별 결과를 분리하여 감리의견에 기재한다.

2) {통합테스트결과서}에 "정상 동작", "이상 없음"만 기재된 경우 시험 완료로 인정하지 않는다: "N개 케이스 중 판정-설명 방향 일치율 98%" 등 케이스 수·일치율·불일치 건수를 포함한 측정 가능한 형태로 재기재하도록 보완권고한다.

3) 설계서에 정의된 전문가 뷰 등 사용자 유형별 화면이 미구현된 경우 설계 미충족으로 처리한다: 형식적 충족 여부에 그치지 말고 실질적 이행 근거와 책임자 확인 여부까지 점검한다.

4) 판정-설명 일치성은 케이스 수와 일치율로 검증한다: "승인" 판정인데 설명 화면의 주요 기여 요인이 부정적 방향으로 표시되면 이용자는 시스템 전체를 신뢰할 수 없게 된다. 감리원은 판정-설명 방향 일치 검증이 독립 시험 항목으로 존재하고 불일치 건수가 기재되었는지 확인한다.

5) 특이값·경계값 입력 시의 화면 동작을 확인한다: 극단적 수치나 결측값이 포함된 입력에서 XAI 화면이 오류를 내거나 빈 화면을 표출하는 사례가 많다. 감리원은 특이값 시나리오의 시험 결과와, 데이터 오류·누락 시 오류 메시지가 정상 표시되는지를 확인한다.

AG_SD44-2-03 | 환각(Hallucination) 억제테스트 통과 점검

응용시스템 | 시험 | 인공지능 모델 개발 | 할루시네이션

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, RAG(검색 증강 생성) 방식으로 구현된 AI 서비스가 답변 생성 시 원본 출처를 명시하고 검색된 문서 범위 내에서만 응답하여 환각(Hallucination) 오류를 억제하는 기능이 설계 의도에 따라 정상 구현되었는지를 {테스트결과서}를 통해 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 RAG 구현 기술의 세부 알고리즘 분석이 아닌, 다음 세 가지 관점에 집중한다.

① 'AI 환각 억제 설계서' 및 'RAG 시스템 설계서'에 환각 억제 기준(출처 명시 방식, 근거 범위 제한 기준, 출처 불명 시 응답 거부 조건)이 정의되어 있으며, {테스트결과서}에서 해당 기준에 따른 시험이 수치로 확인되는가, ② 출처 명시 기능이 {시스템화면}에서 설계서에 정의된 형식과 일치하게 구현되어 있는가, ③ 검색된 문서에 없는 정보를 AI가 임의로 생성(환각)하는 경우를 탐지하는 시험이 충분히 수행되었는가. 이 세 가지 관점을 중심으로 설계 산출물과 {테스트결과서} 간 일치 여부 및 시험의 충분성을 확인한다.

2. 필요성

▶ 환각 억제 테스트 통과 여부 점검이 필요한 이유는 다음과 같다.

▶ 환각 오류는 AI 응답의 신뢰성을 근본적으로 훼손하는 핵심 위험: RAG 기반 AI 서비스에서 환각이 발생하면 AI가 검색된 문서에 없는 정보를 사실처럼 생성한다. 이는 이용자가 잘못된 정보를 사실로 오인하게 하여 금융·의료·법률 등 고위험 도메인에서 특히 심각한 피해를 초래하므로, 배포 전 환각 억제 시험 통과 여부를 확인해야 한다.

▶ 출처 미명시는 이용자의 응답 검증 불가 및 AI 신뢰성 저해: RAG 기반 AI가 출처를 명시하지 않으면 이용자는 AI 응답의 근거를 확인할 방법이 없다. 출처 명시는 이용자가 응답의 신뢰 여부를 스스로 판단할 수 있게 하는 최소한의 투명성 요건이며, TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 이를 응답 투명성의 핵심 요건으로 제시하고 있다.

▶ 환각 억제율 기준값 미검증은 운영 중 오류 응답 허용 수준 미보장: {테스트결과서}에서 환각 억제율이 수치로 검증되지 않으면, 운영 중 환각 오류가 어느 수준으로 발생하는지 배포 전 보증할 수 없다. 'AI 서비스 요구사항 정의서'에 환각 억제율 기준값이 정의되고 {테스트결과서}에서 기준 충족이 확인된 경우에만 운영 신뢰성을 보장할 수 있다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {테스트결과서} 등 산출물을 통해 AI 특화 기능이 적정하게 시험되었는지 여부를 주요 점검 항목으로 규정하며, RAG 기반 AI 서비스의 환각 억제 기능은 이 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	환각 억제 기준 정의·발주기관 승인 여부	검색 범위 제한 방식·출처 명시 형식·출처 불명 시 응답 거부 조건 등 환각 억제 기준이 'AI 환각 억제 설계서' 또는 'RAG 시스템 설계서'에 구체적으로 정의되고 발주기관 승인을 받았는지 확인
2	환각 유형별 억제율 시험·기재 여부	사실 왜곡·정보 추가·출처 혼동 등 환각 유형별 시험 시나리오가 {테스트결과서}에 독립 포함되고 각 유형의 환각 억제율(%)이 수치로 기재되었는지 확인
3	환각 억제율 기준값 충족 여부	환각 억제율이 'AI 서비스 요구사항 정의서'의 기준값을 충족하는지 {테스트결과서}와 대조
4	RAG 출처 표시 형식·출처 명시율 확인 여부	RAG 기반 답변 시 원본 출처(문서명·페이지·URL 등)가 'AI 환각 억제 설계서'의 형식·위치대로 {시스템화면}에 표시되고 출처 명시율(%)이 {테스트결과서}에서 수치로 확인되는지 확인
5	출처 불명 질의 응답 거부 시험 여부	검색 문서에 정보가 없는 출처 불명 질의에 대해 AI가 임의 생성 없이 '확인할 수 없습니다' 등 응답 거부 메시지를 출력하는지 {테스트결과서}에서 확인
6	출처 정확성 시험 수행 여부	출처 명시 링크·참조 위치가 실제 원본 문서 내용과 일치하는지(허위 출처 생성 여부) {테스트결과서}에서 출처 정확성 시험이 수행되었는지 확인
7	환각 발생 이력 기록 여부	발생 일시·질의 내용·환각 유형·탐지 방식 등 환각 발생 이력이 {프로그램 명세서} 정의대로 기록되는지 {테스트결과서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	환각 억제 기준이 설계서에 구체적으로 정의·승인되어 있고, {테스트결과서}에 환각 유형별 억제율·출처 명시율·출처 불명 질의 응답 거부·출처 정확성 시험 결과가 수치로 기재되어 요구사항 기준값을 충족하는 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	환각 억제 기능은 구현·시험되었으나 일부 환각 유형 시나리오 누락, 억제율 기준값 미정의, 출처 명시 형식의 설계서 불일치, 환각 발생 이력 기록 미비 등 부분적 미흡 사항이 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	환각 억제 기준이 정의되지 않았거나 환각 억제 시험이 전혀 수행되지 않아, 출처 명시·응답 거부 등 핵심 기능의 정상 동작을 보증할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "환각 없음 확인" 같은 정성 표현은 시험 완료로 인정하지 않는다: 사실 왜곡·정보 추가·출처 혼동·허위 출처 등 유형별 억제율을 확인할 수 없기 때문이다. 감리원은 유형별 시험 건수·환각 발생 건수·억제율(%)을 수치로 재기재하도록 보완권고를 제시한다.

2) 출처 명시율뿐 아니라 출처 정확성(허위 출처 여부) 시험을 반드시 확인한다: 출처가 표시되더라도 실제 원본 문서와 불일치하면 오히려 이용자의 오신뢰를 유발한다. 감리원은 출처 링크·참조 위치가 원본 내용과 일치하는지 시험되었는지 확인하고, 누락 시 보완권고를 제시한다.

3) 억제율 기준값이 요구사항에 사전 정의되어 있는지 먼저 확인한다: 기준값 없이 측정된 억제율만으로는 합격 여부를 판정할 수 없다. 감리원은 'AI 서비스 요구사항 정의서'의 기준값과 {테스트결과서}의 실측값을 대조한다.

4) 출처 불명 질의에 대한 응답 거부 시험이 포함되었는지 확인한다: 검색된 문서에 해당 정보가 없을 때 임의 생성 응답 대신 "확인할 수 없습니다"로 응답하는지가 환각 억제의 핵심이다. 감리원은 해당 시나리오가 독립 시험 항목으로 존재하는지 확인한다.

5) 환각 발생 이력의 기록 여부를 확인한다: 발생 일시, 질의 내용, 환각 유형, 탐지 방식이 기록되지 않으면 운영 단계에서 개선 근거를 확보할 수 없다. 감리원은 {프로그램 명세서}의 이력 항목 정의와 실제 이력 생성 결과를 대조한다.

AG_SD44-2-04 | 비동기 추론 프로그레스 Bar 상태 동작 점검

응용시스템 | 시험 | 시스템 구현 | UI/UX, 접근성

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · KWCAG

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 수초~수십 초 이상 소요되는 AI 추론·생성 작업이 비동기 방식으로 처리되는 동안, {시스템화면}에서 이용자에게 처리 진행 상태(로딩 중·단계별 진행률·예상 대기시간 등)와 완료 상태가 설계 의도에 따라 정상 표출되는지를 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 프로그레스 바의 구현 기술이나 UI 디자인 세부 분석이 아닌 ① 'AI 서비스 UX 설계서' 또는 '화면 설계서'에 비동기 추론 처리 중 표출할 상태 정보(로딩 표시 유형·단계별 안내 메시지·예상 소요시간 표시 여부·완료 전환 방식)가 정의되어 있으며 {시스템화면}에서 설계서와 일치하게 구현되어 있는가, ② 장시간 처리(설계서에 정의된 기준 시간 초과) 시 이용자가 처리 진행 여부를 인식할 수 있는 추가 안내가 표출되며 타임아웃 및 처리 실패 상황에서 적절한 안내가 제공되는가, ③ {테스트결과서}에 정상 처리·장시간 처리·타임아웃·처리 실패 시나리오가 각각 독립 항목으로 포함되어 있으며 시험 결과가 수치 또는 화면 상태로 확인되는가의 세 가지 관점에 집중한다.

2. 필요성

▶ AI 추론·생성 작업은 처리 시간이 길고 예측이 어려워 이용자 이탈과 중복 요청 문제를 유발한다. 처리 중 화면에 아무런 상태 표시가 없으면 이용자는 시스템이 멈춘 것으로 오인하여 페이지를 이탈하거나 동일 요청을 반복 전송하며, 중복 요청은 서버 과부하와 AI 추론 비용 증가로 이어진다.

▶ 프로그레스 바 고착(멈춤) 또는 완료 전환 미동작은 배포 후 이용자 불만의 주요 원인이다. 이 오류는 단위시험에서 발견되기 어렵고 {테스트결과서}에 명확한 시험 항목이 없으면 배포 후에야 발견된다.

▶ 비동기 처리 타임아웃 및 오류 처리 미구현은 이용자를 무한 대기 상태에 방치한다. AI 추론 서버 장애나 네트워크 지연 시 타임아웃 처리 없이 프로그레스 바만 계속 표시되면 이용자는 처리가 진행 중인지 실패했는지 알 수 없어, 실질적 서비스 불가 상태임에도 안내되지 않아 이용자 경험을 심각하게 훼손한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {시스템화면} 등 산출물이 설계 의도대로 구현되고 시험되었는지 여부를 점검 항목으로 규정한다.

3. 검토방법

No	점검 항목	확인 방법
1	비동기 추론 상태 정보 설계·승인 여부	로딩 표시 유형·단계별 안내 메시지·예상 소요시간·추가 안내 기준 시간·타임아웃 기준값·완료 전환 방식이 'AI 서비스 UX 설계서'에 정의되고 발주기관 승인을 받았는지 확인
2	추론 요청 직후 로딩 상태 즉시·지속 표출 여부	AI 추론 요청 직후 1~2초 이내에 설계된 로딩 상태(프로그레스 바·스피너·메시지)가 표출되고 처리 중 중단 없이 지속되는지 {시스템화면}에서 확인
3	장시간 처리 시 추가 안내·취소 수단 제공 여부	기준 시간 초과 시 단계별 진행 안내 또는 예상 대기시간이 추가 표출되고 이용자가 요청을 취소할 수 있는지 {시스템화면}·{테스트결과서}에서 확인
4	추론 완료 시 로딩 종료·완료 전환 여부	AI 추론 완료 후 로딩 상태가 즉시 종료되고 완료 상태(결과 표시·완료 메시지·아이콘 전환)로 전환되며 로딩 표시가 잔존하지 않는지 {시스템화면}·{테스트결과서}에서 확인
5	타임아웃 시 실패 안내·재시도 수단·기준값 정의 여부	타임아웃 발생 시 처리 실패 안내와 재시도·취소 수단이 제공되고 {프로그램 명세서}에 타임아웃 기준값·처리 로직이 정의되었는지 {시스템화면}·{테스트결과서}에서 확인
6	로딩 중 중복 요청 방지 여부	로딩 중 동일 작업의 중복 요청이 버튼 비활성화 또는 차단 메시지로 방지되는지 {시스템화면}·{테스트결과서}에서 확인
7	추론 처리 실패 시 오류 안내·상태 전환 여부	서버 오류·추론 결과 없음 등 처리 실패 시 오류 원인·조치 방법 안내 메시지가 표시되고 로딩 상태가 오류 상태로 전환되는지 {시스템화면}·{테스트결과서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	로딩 표시 즉시성·지속성, 설계 일치, 장시간 처리 시 추가 안내·취소 수단, 완료 즉시 전환, 타임아웃·처리 실패 안내와 재시도 수단, 중복 요청 방지, 오류 이력 관리가 모두 확인된 경우	현행 유지
보완권고 (△)	기본 로딩 표시는 동작하나 기준 시간 초과 후 추가 안내 미표출, 완료 후 로딩 표시 잔존, 타임아웃 기준값 미정의, 로딩 중 중복 클릭 시 추가 요청 전송, 오류 시 서버 내부 코드 노출, 오류 이력 항목 누락 등이 있는 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	로딩 상태 표시 자체가 미구현되어 이용자가 처리 진행 여부를 전혀 알 수 없거나, 타임아웃 처리가 없어 무한 로딩 상태가 지속되거나, 완료 후 로딩 상태가 전혀 해제되지 않거나 처리 실패 시 공백 화면이 표시되는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 정상 처리 시나리오만 시험하고 타임아웃·장시간 시나리오 누락 여부 반드시 확인: {테스트결과서}에 "AI 추론 정상 완료 확인"만 기재된 경우 장시간 처리·타임아웃·처리 실패 시의 화면 동작이 시험되지 않은 것으로 판단하고, 정상·장시간·타임아웃·실패 시나리오가 모두 독립 항목으로 포함되어 있지 않으면 시험 완료로 인정하지 않는다는 점을 보완권고에 명시한다.

2) 중복 요청 방지는 {시스템화면}에서 로딩 중 버튼 재클릭으로 반드시 현장 확인: {테스트결과서}만으로는 충분히 검증하기 어려우므로 현장에서 AI 추론 요청 후 로딩 중 동일 버튼을 3~5회 빠르게 재클릭하여 추가 요청 전송 여부를 직접 확인하고, 브라우저 개발자 도구 Network 탭으로 중복 요청 전송 여부를 확인한다.

3) 타임아웃 기준값이 정의되지 않으면 무한 로딩 위험 — 배포 전 필수 완료 조건: 'AI 서비스 요구사항 정의서'에 타임아웃 기준값이 수치로 정의되어 있지 않으면 AI 추론 서버 장애 시 이용자가 무한 로딩 상태에 방치될 수 있으므로, AI 기능 유형별(추론·이미지 생성·문서 생성) 타임아웃 기준값을 각각 수치로 정의하도록 배포 전 완료 조건으로 권고한다.

4) 완료 후 로딩 표시 잔존은 현장 직접 확인으로 발견: 완료 후 로딩 표시가 3초 이상 잔존하는 경우는 {테스트결과서}에 기재되지 않는 경우가 많으므로, 현장에서 AI 추론 완료 직후 화면 상태를 초 단위로 관찰하여 로딩 표시 해제 시점을 직접 확인하고, 직접 수행이 어려우면 사업자에게 시연을 요청한다.

5) 처리 실패 시 서버 오류 코드 노출 여부와 화면 이탈 후 복귀 시 상태 동기화 확인: "AI_ERROR_5032", "NullPointerException" 등 내부 오류 코드 노출은 이용자 경험과 보안 관점 모두에서 부적합이므로 실패 케이스 재현으로 안내 메시지의 적정성을 점검하고, 로딩 중 다른 메뉴 이동·탭 전환 후 복귀 시에도 비동기 처리가 끊기지 않고 진행·완료 상태가 정확히 재표출되는지 현장 확인한다.

AG_SD44-2-05 | 시스템 오작동 기본값 응답점검

응용시스템 | 시험 | 시스템 구현 | 가용성/이중화

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 추론 타임아웃·모델 오류·외부 API 연동 실패·입력값 이상 등 시스템 오작동 상황이 발생했을 때, 서비스가 이용자에게 안전한 기본값(Default) 응답 메시지를 반환하고 시스템이 안전 상태(Fail-safe)로 전환되는지를 {테스트결과서} 검토와 {시스템화면} 직접 확인을 통해 제3자적 관점에서 점검하는 것을 목적으로 한다.

감리원의 점검은 오류 처리 기술 구현 세부 사항의 분석이 아닌 ① 'AI 서비스 요구사항 정의서' 및 '오류 처리 설계서'에 오작동 유형별(타임아웃·추론 실패·외부 연동 오류·입력값 이상) 디폴트 응답 메시지와 시스템 전환 동작이 수치와 문구로 정의되어 있으며 발주기관 승인을 받았는가, ② {테스트결과서}에 오작동 유형별 시험 시나리오가 독립 항목으로 포함되어 있으며 디폴트 응답 메시지 반환과 시스템 안전 전환이 시험 결과로 확인되는가, ③ {시스템화면}에서 오작동 상황을 직접 또는 사업자 시연으로 재현했을 때 이용자에게 표출되는 메시지가 안전하고 이해 가능한 내용인가의 세 가지 관점에 집중한다.

2. 필요성

▶ AI 서비스는 일반 IT 시스템 대비 오작동 유형이 다양하고 예측이 어려워 배포 전 체계적인 시험이 필수적이다. AI 추론 엔진·외부 모델 API·벡터 데이터베이스(RAG 연동)·입력값 전처리 등 복수 구성 요소 중 하나라도 오작동하면 서비스 전체가 비정상 상태가 될 수 있으며, AI 서비스의 오류는 잘못된 추론 결과를 정상 응답으로 반환하는 형태로 나타날 수 있어 단순 오류 코드 표시보다 더 세밀한 대응이 필요하다.

▶ 디폴트 응답 메시지가 정의되어 있지 않으면 오작동 발생 시 내부 스택 트레이스·오류 코드·AI 모델명 등 시스템 내부 정보가 {시스템화면}에 그대로 표시될 수 있으며, 이는 이용자 경험을 훼손할 뿐 아니라 시스템 구조 정보를 노출하는 보안 위험을 유발한다.

▶ AI 추론 타임아웃 기준이 설정되지 않으면 AI 모델 응답이 없는 상황에서 서비스가 무한 대기 상태로 진입하여 이용자가 응답을 받지 못하고 서버 연결 자원이 계속 점유되며, 이 문제는 배포 후 부하 상황에서 복합적으로 발생하여 서비스 전체 장애로 이어질 수 있다.

▶ 감리기준 및 AI 안전성 요건 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {테스트결과서} 등 산출물이 요구사항에 정의된 기능(오류 처리 포함)을 설계 의도대로 시험했는지 점검을 요

구하며, TTA 신뢰할 수 있는 인공지능 개발안내서는 AI 시스템의 안전성(Safety) 요건으로 오작동 시 안전 상태 전환 체계 구축을 규정한다.

3. 검토방법

No	점검 항목	확인 방법
1	오작동 유형별 디폴트 응답·타임아웃 기준 정의·승인 여부	타임아웃·AI 추론 실패·외부 API 오류·입력값 이상·모델 로딩 실패 등 오작동 유형별 디폴트 응답 메시지 문구와 타임아웃 기준값이 '오류 처리 설계서'·'AI 서비스 요구사항 정의서'에 정의되고 발주기관 승인을 받았는지 확인
2	타임아웃 시나리오·디폴트 응답 반환 확인 여부	타임아웃 시나리오가 {테스트결과서}에 독립 항목으로 포함되고 기준값 경과 후 디폴트 응답이 반환되어 무한 대기가 발생하지 않는지 {시스템화면} 재현으로 확인
3	AI 추론 실패 시 디폴트 응답 반환 여부	모델 오류·추론 결과 없음·신뢰도 미달 등 AI 추론 실패 시나리오가 {테스트결과서}에 포함되고 잘못된 AI 응답 대신 디폴트 응답이 반환되는지 확인
4	외부 API 오류 시 서비스 지속·전환 여부	번역 API·외부 AI 모델·데이터 수집 API 등 연동 오류 시나리오가 {테스트결과서}에 포함되고 서비스 전체 중단 없이 디폴트 응답 또는 제한 서비스 모드로 전환되는지 확인
5	디폴트 메시지의 안전성·내부 정보 미노출 여부	오작동 재현 또는 사업자 시연 시 이용자에게 표출되는 디폴트 메시지가 이해 가능한 안전한 문구이고 내부 오류 코드·스택 트레이스·AI 모델명 등이 노출되지 않는지 {시스템화면}에서 확인
6	디폴트 전환 후 재시도·대체 안내 제공 여부	디폴트 응답 전환 후 재시도 수단 또는 대체 서비스 이용 안내가 제공되는지 {시스템화면}·{테스트결과서} 재시도 시험에서 확인
7	오작동 이력 기록·운영 대응 방안 정의 여부	발생 시각·오류 유형·처리 결과·디폴트 전환 여부 등 오작동 이력이 {프로그램 명세서} 정의대로 기록되고 배포 후 모니터링·개선 방안이 '운영 비상 대응 절차서'에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	오작동 유형별 디폴트 메시지 정의·승인, 타임아웃 기준값 수치 정의, {테스트결과서}에 오작동 유형별 시나리오 전체 포함, 디폴트 메시지 반환 결과 확인, 내부 정보 미노출, 재시도 수단 제공, 오작동 이력 기록, 운영 비상 대응 절차(Runbook) 수립이 모두 확인된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	기본 오류 처리는 구현되었으나 일부 오작동 유형 시나리오 누락, 타임아웃 기준값 미정의, 신뢰도 기준 미달 시험 미포함, 외부 API 오류 전파 차단 미시험, 오류 메시지 내부 코드 노출, 재시도 수단 미제공, 이력 항목 미정의, Runbook 미수립 등이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	타임아웃 처리 자체가 미구현되어 무한 대기 상태가 확인되거나, 모든 오작동 유형에서 디폴트 응답 없이 내부 오류 정보가 그대로 이용자에게 노출되거나, {테스트결과서}에 오류 처리 관련 시험이 전혀 포함되어 있지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) {테스트결과서}의 오작동 시나리오는 설계서와 1:1 대조로 누락 여부 반드시 확인: '오류 처리 설계서'에 5가지 오작동 유형이 정의되어 있는데 {테스트결과서}에 타임아웃 시나리오 1가지만 포함된 경우 나머지 4가지는 시험되지 않은 것이므로, 설계서의 오작동 유형 목록과 {테스트결과서}의 시나리오 목록을 나란히 놓고 1:1 매핑 체크리스트를 작성하여 누락 유형을 보완권고에 구체적으로 명시한다.

2) 신뢰도 기준 미달 시 디폴트 전환 시험은 반드시 독립 항목으로 확인: AI 서비스 특유의 오작동 유형인 "신뢰도 기준 미달 시 잘못된 결과 반환" 문제는 일반 IT 오류 처리 시험에서 다루지 않는 경우가 많으므로, 해당 시나리오가 없으면 낮은 신뢰도의 AI 응답이 그대로 이용자에게 전달될 위험이 있어 배포 전 완료 조건으로 보완권고를 제시한다.

3) 타임아웃 기준값은 AI 추론과 외부 API를 구분하여 정의되었는지 확인: AI 추론은 수십 초가 소요될 수 있어 일반 API 타임아웃(10초 내외)보다 긴 기준값이 필요하므로, 하나의 기준값만 정의된 경우 AI 추론과 외부 API 타임아웃을 각각 수치로 정의하도록 권고하고, 단계별(입력 처리·AI 추론·결과 반환) 타임아웃 구분 여부도 함께 확인한다.

4) 디폴트 메시지의 내부 오류 코드 노출 여부는 현장 재현으로 직접 확인: {테스트결과서}에 "정상 처리"로 기재되어 있어도 실제 화면에서는 내부 오류 코드가 노출되는 경우가 있으므로, 사업자에게 타임아웃 또는 추론 실패 재현 시연을 요청하여 API Key·내부 IP·DB 에러 코드·AI 시스템 프롬프트 원문 등 민감 정보 없이 정제된 메시지만 표출되는지 보안 관점에서 직접 확인한다.

5) '운영 비상 대응 절차서(Runbook)' 미수립은 종료단계 보완권고 대상: 배포 후 오작동 발생 시 신속한 디폴트 응답 전환과 서비스 복구를 위해 운영 절차가 사전에 정의되어야 하므로, Runbook 수립 여부를 종료단계에서 확인하고 미수립 시 배포 전 완료 조건으로 보완권고를 제시한다.

AG_SD44-2-06 | 안전모드 Fail-safe 전환테스트 점검

응용시스템 | 시험 | 시스템 구현 | 가용성/이중화

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템이 강제 임계값 초과 또는 장애 상황 발생 시 안전모드(Fail-safe, 기본모드)로 자동 전환되는 기능이 설계 의도에 따라 정상적으로 구현되고 충분히 시험되었는지 여부를 {테스트결과서}를 통해 제3자적 관점에서 검토하는 것을 목적으로 한다.

특히, 안전모드 전환 기능은 화면 표출 여부가 아닌 전환 조건 정의의 명확성, 전환 시험의 충분성(임계값별 시나리오 포함 여부), 전환 후 기본모드 동작의 적정성, 그리고 복구(원복) 절차의 완결성을 중심으로 점검한다.

감리원은 전환 메커니즘의 기술적 구현을 검토하는 것이 아니라, {테스트결과서} 및 설계 산출물 간의 일치 여부와 시험 충분성을 확인한다.

2. 필요성

▶ 안전모드(Fail-safe) 전환 테스트 점검이 필요한 이유는 다음과 같다.

▶ 안전모드 미구현 또는 전환 실패는 AI 시스템 장애의 직접적 피해 확산: AI 모델이 비정상 출력을 생성하거나 서비스 장애가 발생하는 상황에서 안전모드 전환이 이루어지지 않으면, 오류 판정 결과가 이용자에게 그대로 전달되거나 서비스 전체가 중단되는 연쇄 장애로 이어질 수 있다.

▶ 강제 임계값 초과 시나리오 시험 미수행은 배포 후 리스크를 배가: 안전모드 전환 기능이 코드에 구현되어 있더라도, 실제 임계값 초과·장애 시나리오를 적용한 시험이 수행되지 않으면 전환 기능의 정상 동작을 보증할 수 없다. TTA 신뢰할 수 있는 인공지능 개발안내서는 강건성(Robustness) 검증의 일환으로 이상 상황 시험 수행을 권고하고 있다.

▶ 기본모드 동작의 적정성 미검증 시 안전모드의 실질적 안전성 부재: 안전모드로 전환되더라도 기본모드에서 제공되는 서비스(규칙 기반 처리, 담당자 수동 처리 안내 등)가 적정하게 동작하지 않는다면, 안전모드 전환은 형식적 조치에 불과하다. 기본모드의 실제 동작 내용과 이용자 안내 방식도 함께 점검해야 한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 시험의 적정성 확인을 주요 점검 항목으로 규정하며, AI 시스템의 안전모드 전환 시험 수행 여부는 AI 특화 시험 항목으로서 이 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	안전모드 전환 트리거 조건의 유형별 수치 정의 여부	신뢰점수 임계값·응답 시간 초과·모델 오류율 임계값·외부 API 장애 등 안전모드 전환 트리거 조건이 'AI 안전모드 설계서'에 유형별로 수치를 포함하여 명확하게 정의되었는지 확인
2	안전모드 전환 시험 시나리오·전환 결과 기재 여부	강제 임계값 초과·장애 주입 등 안전모드 전환 시험 시나리오가 트리거 조건 유형별로 {테스트결과서}에 독립 포함되고 전환 성공 여부가 수치·결과로 기재되었는지 'AI 시험 시나리오 및 결과서'와 함께 확인
3	전환 후 기본모드 정상 동작 여부	규칙 기반 처리·담당자 수동 처리·서비스 최소화 등 안전모드 전환 후 기본모드가 설계서 정의대로 동작하는지 {테스트결과서}·{통합테스트결과서}에서 확인
4	안전모드 전환 응답 시간 기준 충족 여부	안전모드 전환 소요 시간이 'AI 서비스 수준 협약서' 또는 '운영 요구사항 정의서'의 기준을 충족하는지 {테스트결과서}에서 수치로 확인
5	안전모드 발동 시 이용자 안내 표출 여부	서비스 일시 변경 공지·담당자 연락 안내 등 안전모드 발동 시 이용자 화면 안내가 표출되는지 {테스트결과서}·{통합테스트결과서}에서 확인
6	안전모드 해제 조건·복구 절차·복구 시험 여부	안전모드 해제 조건과 AI 모드 복귀 절차가 'AI 안전모드 설계서'·'AI 운영 비상대응 절차서'에 명시되고 복구 시험 결과가 {테스트결과서}에 포함되었는지 확인
7	안전모드 발동 이력 로그 기록 여부	트리거 조건·전환 시각·기본모드 동작 내역·복구 시각 등 안전모드 발동 이력이 로그로 기록되도록 설계서에 정의되고 {프로그램 명세서}·{테스트결과서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'AI 안전모드 설계서'에 전환 트리거 조건이 유형별 수치로 명확히 정의되어 있고, {테스트결과서}에 트리거 유형별 전환 시험, 기본모드 동작, 전환 응답 시간, 복구 시험 및 로그 기록이 수치·결과로 기재되어 기준을 충족하는 경우	현행 유지
보완권고 (△)	안전모드 전환 기능은 구현·시험되었으나 일부 트리거 유형의 시험 누락, 전환 소요 시간 등 수치 미기재, 복구 절차 또는 발동 이력 로그 정의 미비 등 부분적 미흡 사항이 존재하는 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	안전모드 전환 기능이 미구현되었거나, 강제 임계값 초과·장애 시나리오 시험이 전혀 수행되지 않아 전환 기능의 정상 동작을 보증할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "구현했다"는 사업자 확인만으로는 부족하며 반드시 {테스트결과서}의 수치를 확인한다: 안전모드 전환은 평상시 발동되지 않는 기능이므로 구두 확인은 근거가 될 수 없다. 감리원은 트리거 조건 유형별 시험 결과가 전환 소요 시간 등 수치로 기재되어 있는지 확인하고, 미기재 시 보완권고를 제시한다.

2) 전환 시험뿐 아니라 복구(원복) 시험 결과까지 확인해야 점검이 완결된다: 안전모드 해제 조건과 AI 모드 복귀 절차가 시험되지 않으면 장애 복구 후 서비스 재개가 지연된다. 감리원은 복구 시험 결과와 'AI 운영 비상대응 절차서'의 정합성을 함께 확인한다.

3) 트리거 조건은 유형별로 독립 시험되어야 한다: 신뢰점수 임계값 초과, 응답 시간 초과, 모델 오류율 초과는 각각 다른 코드 경로를 거치므로 하나의 시나리오로 묶어 시험하면 미검증 경로가 남는다. 감리원은 트리거 유형별 독립 시나리오의 존재 여부를 확인한다.

4) 기본모드의 실제 동작을 확인하지 않으면 안전모드는 형식적 조치에 그친다: 전환 자체가 성공하더라도 규칙 기반 처리·담당자 수동 처리 안내가 설계서대로 동작하지 않으면 이용자 피해는 그대로 발생한다. 감리원은 기본모드 동작과 이용자 안내 화면 표출을 {통합테스트결과서}에서 확인한다.

5) 안전모드 발동 이력의 로그 기록 여부를 확인한다: 트리거 조건, 전환 시각, 기본모드 동작 내역, 복구 시각이 기록되지 않으면 사후 원인 분석이 불가능하다. 감리원은 {프로그램 명세서}의 이력 항목 정의와 {테스트결과서}의 실제 기록 생성 결과를 대조한다.

AG_SD44-2-07 | 실시간 지연시간 테스트 점검

응용시스템 | 시험 | 시스템 구현 | 성능

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 실제 운영 환경과 유사한 동시 접속 부하 조건에서 AI 서비스의 평균 응답시간(Latency)이 'AI 서비스 요구사항 정의서'에 정의된 목표 밀리초(ms)를 달성하는지를 {성능테스트결과서}를 통해 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 성능 최적화 기술이나 인프라 튜닝 방식의 분석이 아닌 다음 세 가지 관점에 집중한다.

① 'AI 서비스 요구사항 정의서'에 기능 유형별(일반 응답·AI 추론·RAG 검색 등) 응답시간 목표값과 동시 접속자 수 기준이 수치로 정의되어 있고 해당 조건의 시험이 수행되었는가, ② 평균 응답시간 외에 95·99 백분위수(Percentile) 응답시간과 오류율이 {성능테스트결과서}에 수치로 기재되어 있으며 최대 부하 조건에서도 목표값을 유지하는가, ③ 목표 미달 시 원인 분석과 개선 조치 결과가 {성능테스트결과서}에 기재되어 있는가. 이 세 가지 관점을 중심으로 성능 시험의 충분성과 요구사항 충족 여부를 확인한다.

2. 필요성

▶ 실시간 지연시간(Latency) 테스트 점검이 필요한 이유는 다음과 같다.

▶ AI 서비스는 일반 웹 서비스 대비 응답시간이 현저히 길어 이용자 경험에 직접적인 영향을 미침: AI 추론, RAG 검색, 대형 언어 모델(LLM) 응답 생성은 일반 데이터베이스 조회 대비 처리 시간이 수십~수백 배 길다. 배포 전 실제 동시 접속 조건에서 응답시간이 목표값을 달성하는지 {성능테스트결과서}에서 확인하지 않으면 운영 개시 후 응답 지연으로 인한 이용자 불만과 서비스 신뢰 저하가 발생한다.

▶ 단일 사용자 응답시간과 동시 접속 응답시간은 큰 차이를 보임: 개발 환경에서 단일 사용자 기준으로 응답시간이 빠르더라도, 실제 운영 환경과 유사한 동시 접속자 수 조건에서는 응답시간이 수 배 이상 증가하는 경우가 많다. {성능테스트결과서}에 동시 접속자 수를 명시한 부하 조건이 정의되어 있어야 시험 결과의 신뢰성을 확인할 수 있다.

▶ AI 추론 요청의 응답시간 변동성(Variance)은 일반 서비스보다 크고 예측이 어려움: 동일한 AI 서비스에서도 입력 데이터의 복잡도, 모델 캐시 히트 여부, GPU 자원 경합 상태에 따라 응답시간이 크게 달라진다. 평균 응답시간만으로는 이 변동성을 파악할 수 없으며, 95·99 백분위수 응답시간이 {성능테스트결과서}에 함께 기재되어야 한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {성능테스트결과서} 등 산출물을 통해 시스템 성능이 요구사항을 충족하는지 여부를 점검 항목으로 규정한다.

3. 검토방법

No	점검 항목	확인 방법
1	기능 유형별 응답시간 목표·동시 접속자 기준 정의·승인 여부	일반 조회·AI 추론·RAG 검색·스트리밍 등 기능 유형별 응답시간 목표값(ms)과 동시 접속자 수 기준이 'AI 서비스 요구사항 정의서'에 수치로 정의되고 발주기관 승인을 받았는지 확인
2	부하 조건별 응답시간·오류율 기재 여부	동시 접속자 수를 명시한 부하 조건별로 평균·P95·P99 응답시간과 오류율이 {성능테스트결과서}에 각각 수치로 기재되었는지 확인
3	응답시간 목표값 기능 유형별 충족 여부	평균·P95 응답시간이 'AI 서비스 요구사항 정의서'의 목표값을 기능 유형별로 충족하는지 {성능테스트결과서}와 대조
4	AI 특화 기능 응답시간 시험·목표 별도 정의 여부	AI 추론·RAG 검색 등 AI 특화 기능의 응답시간 시험이 일반 기능과 구분된 독립 항목으로 {성능테스트결과서}에 포함되고 목표값이 별도 정의되었는지 확인
5	최대 부하 조건 응답시간·오류율 확인 여부	목표 동시 접속자 수의 120~150% 이상 최대 부하 조건에서의 응답시간·오류율이 {성능테스트결과서}에 포함되고 허용 오류율 기준을 초과하지 않는지 확인
6	목표 미달 항목 원인 분석·재시험 결과 기재 여부	응답시간 목표 미달 항목의 병목 지점 원인 분석·개선 조치·재시험 결과가 '성능 시험 시나리오 및 결과서'에 기재되었는지 확인
7	운영 응답시간 모니터링 기준 정의 여부	운영 환경 응답시간 지속 모니터링 기준(임계값·알림 발송 조건)이 'AI 성능 모니터링 운영 절차서'에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	기능 유형별 응답시간 목표값과 동시 접속자 수 기준이 수치로 정의되고, 부하 조건별 평균·P95·P99 응답시간과 오류율이 {성능테스트결과서}에 기재되어 목표값을 모두 충족하며, AI 추론·RAG 검색 등 AI 특화 기능의 응답시간 시험이 일반 기능과 구분된 독립 항목으로 수행되고, 최대 부하 조건(목표 동시 접속자 수의 120~150% 이상)의 응답시간·오류율이 허용 기준을 충족하며, 목표 미달 항목의 원인 분석·개선·재시험 결과와 운영 환경의 응답시간 모니터링 임계값·알림 조건이 정의된 경우	현행 유지
보완권고 (△)	성능 시험은 수행되었으나 부하 조건 미명시, P95·P99 등 일부 지표 누락, AI 특화 기능 미구분, 목표 미달 항목에 대한 원인 분석·재시험 결과 미기재 등이 확인된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	응답시간 목표값이 정의되지 않았거나 동시 접속 부하 시험이 수행되지 않아 성능 요구사항 충족 여부를 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 단일 평균 응답시간만으로는 성능 충분성을 판단할 수 없다: 기능 유형별로 평균·P95·P99 응답시간과 오류율이 각각 수치로 기재되어 있지 않으면 충분한 성능 시험으로 인정하지 않는다는 점을 보완권고에 명시한다.

2) 부하 조건(동시 접속자 수)이 명시되지 않은 성능 시험 결과는 근거로 인정하지 않는다: 단일 사용자 또는 비현실적으로 낮은 부하 조건의 결과일 가능성이 있으므로, 'AI 서비스 요구사항 정의서'에 정의된 목표 동시 접속자 수 조건에서 시험이 수행되었는지 반드시 확인한다.

3) AI 특화 기능(AI 추론·RAG 검색·스트리밍 응답)은 일반 조회 기능과 반드시 구분하여 별도의 목표값과 시험 결과를 점검한다: 하나의 평균값으로 합산 제시된 결과는 AI 기능의 실제 지연을 은폐할 수 있다.

4) 최대 부하 조건에서의 오류율까지 함께 확인한다: 목표 동시 접속자의 120~150% 조건에서 응답시간만 확인하고 오류율을 누락하면 실제 한계 지점을 알 수 없다. 감리원은 최대 부하 시험의 응답시간과 오류율이 함께 기재되었는지 확인한다.

5) 목표 미달 항목의 원인 분석과 재시험 결과를 확인한다: 병목 지점 식별 없이 "개선 예정"으로 종결된 항목은 조치로 인정하지 않는다. 감리원은 개선 조치 후 재시험 결과가 목표값을 충족했는지, 운영 단계 응답시간 모니터링의 임계값·알림 조건이 절차서에 정의되었는지 함께 확인한다.

AG_SD44-2-08 | 프롬프트 인젝션 방어필터 동작점검

응용시스템 | 시험 | 인공지능 모델 개발 | AI특화보안

근거 | OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템에 대한 프롬프트 인젝션(Prompt Injection) 공격 시도 시 방어 필터가 설계 의도에 따라 악의적 입력을 차단하고 안전한 에러 메시지를 반환하는 기능이 정상 구현·시험되었는지 {테스트결과서}를 통해 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원은 방어 필터의 기술적 구현 방식이나 공격 코드 분석이 아닌, 다음 세 가지 관점에서 점검한다.

① '프롬프트 인젝션 방어 설계서'에 정의된 차단 대상 공격 유형·패턴이 {테스트결과서}의 시험 시나리오와 빠짐없이 대응되는가, ② 차단율·오탐률 등 측정 가능한 기준이 수치로 기재되어 충족 여부를 판단할 수 있는가, ③ 차단 시 반환되는 에러 메시지가 시스템 내부 정보를 노출하지 않으면서 이용자에게 적절히 안내하는가. 이 세 가지 관점을 중심으로 설계 산출물과 {테스트결과서}의 일치 여부, 시험의 충분성을 확인한다.

2. 필요성

▶ 프롬프트 인젝션 방어 필터 동작 점검이 필요한 이유는 다음과 같다.

▶ 프롬프트 인젝션은 생성형 AI 서비스의 가장 현실적인 보안 위협으로, 방어 필터 미구현 또는 우회 가능한 필터는 서비스 안전성 전체를 무력화: 악의적 사용자가 텍스트 입력만으로 AI의 시스템 지시사항을 덮어쓰거나 금지 콘텐츠를 생성하도록 유도할 수 있어, 방어 필터의 정상 동작 시험은 배포 전 필수 확인 사항이다.

▶ 다양한 우회 기법에 대한 시험 미수행은 배포 후 보안 사고로 직결: 방어 필터가 코드에 구현되어 있더라도 직접 인젝션·역할극 방식·다국어·인코딩 변환 등 공격 유형별 시험이 수행되지 않으면, 단일 기법만 차단하고 나머지 우회 경로는 열린 상태로 배포되는 위험이 상존한다. TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)도 공격 유형별 보안 검증 수행을 권고하고 있다.

▶ 에러 메시지 부적절 설계는 공격자에게 우회 단서 제공이라는 역효과 발생: 차단 시 반환되는 에러 메시지가 시스템 프롬프트 내용·필터 로직·내부 오류 코드를 노출하면, 오히려 공격자에게 정교한 우회 방법을 제공하는 부작용이 발생한다. 에러 메시지 내용의 적정성은 구현 점검과 별개로 반드시 확인해야 한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 시험의 적정성을 주요 점검 항목으로 규정하며, 생성형 AI 서비스의 프롬프트 인젝션 방어 시험은 AI 특화 시험 항목으로서 이 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	차단 대상 공격 유형의 설계서 정의 여부	직접·간접 인젝션·역할극·가상 시나리오·다국어·인코딩 우회·탈옥 시도 등 차단 대상 공격 유형이 '프롬프트 인젝션 방어 설계서'에 유형별·구체적 예시와 함께 정의되었는지 확인
2	공격 유형별 방어 시험·차단율 기재 여부	공격 유형별 방어 시험 시나리오가 {테스트결과서}에 독립 항목으로 포함되고 차단 건수·차단율이 수치로 기재되었는지 'AI 보안 시험 시나리오 및 결과서'와 함께 확인
3	차단율·오탐률 기준 정의·충족 여부	차단율과 오탐률(정상 입력 오차단 비율)이 'AI 서비스 보안 요구사항 정의서'에 수치 기준으로 정의되고 {테스트결과서}에서 기준을 충족하는지 대조
4	차단 에러 메시지의 내부 정보 미노출·안내 형식 여부	차단 시 반환 에러 메시지가 시스템 프롬프트·필터 로직·오류 코드 등 내부 구조를 노출하지 않고 이용자 이해 가능한 안내 형식으로 구현되었는지 {테스트결과서} 샘플과 설계서 기준을 대조
5	알려진 탈옥 패턴 방어 시험·결과 기재 여부	DAN 모드·역할 부여·분산 입력 등 알려진 탈옥(Jailbreak) 패턴 방어 시험이 {테스트결과서}에 유형별로 포함되고 차단 결과가 수치로 기재되었는지 확인
6	정상 입력 오탐 시험 포함·확인 여부	정상 서비스 이용 시나리오에서 방어 필터가 정상 입력을 잘못 차단하지 않는지 {테스트결과서}에 오탐 시험이 독립 항목으로 포함되어 수치로 확인되는지 확인
7	인젝션 시도 이력 로그 기록 여부	입력 패턴 해시·차단 시각·공격 유형 등 프롬프트 인젝션 시도 이력이 로그로 기록되는 기능이 {프로그램 명세서}에 정의되고 {테스트결과서}에서 로그 생성이 확인되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'프롬프트 인젝션 방어 설계서'에 공격 유형이 구체적으로 정의되어 있고, {테스트결과서}에 유형별 방어 시험·탈옥 시험·오탐 시험이 독립 항목으로 포함되어 차단율·오탐률이 수치 기준을 충족하며, 에러 메시지의 내부 정보 미노출이 확인되는 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	방어 필터가 구현·시험되었으나 일부 공격 유형 시나리오 누락, 차단율·오탐률의 수치 기준 미정의 또는 정성 표현 기재, 에러 메시지 기준 미비 등 부분적 미흡 사항이 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	프롬프트 인젝션 방어 필터가 미구현되었거나, 방어 시험이 전혀 수행되지 않아 공격 차단 여부를 보증할 수 없거나, 에러 메시지가 시스템 내부 정보를 노출하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "차단 정상 동작"은 시험 완료로 인정하지 않음 — 유형별 수치 결과 요구: {테스트결과서}에 프롬프트 인젝션 방어 시험 결과가 "정상 동작 확인", "차단됨" 등 정성 표현으로만 기재된 경우, 어떤 공격 유형을 시험했는지, 차단율이 얼마인지 확인할 수 없다. 공격 유형별 시나리오, 시험 건수, 차단 건수, 차단율을 수치로 재기재하도록 보완권고를 제시한다.

2) 오탐 시험 누락은 서비스 이용성 위험 — 반드시 독립 항목으로 확인: 공격 차단 시험만 수행하고 정상 입력에 대한 오탐 시험이 누락된 경우, 방어 필터가 지나치게 민감하게 설정되어 이용자의 정상 질문이 차단되는 서비스 품질 문제가 발생할 수 있다. 오탐 시험 항목 추가와 오탐률 기준값 정의를 보완권고에 명시한다.

3) 에러 메시지 내용을 직접 검토하여 시스템 내부 정보 노출 여부를 확인한다. 시스템 프롬프트·필터 로직·오류 코드가 노출되면 공격자에게 우회 단서를 제공하므로 즉시 보완을 요구한다.

AG_SD44-2-09 | 데이터 파이프라인 재학습용 동작 점검

응용시스템 | 시험 | 데이터 수집·처리 | 모니터링/재학습

근거 | 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · NIST AI RMF · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 서비스 운영 중 생성되는 추론 결과 데이터 중 이용자 피드백이 긍정적인 데이터(좋아요·높은 평점·수락된 추천·재이용 행동 등)를 자동으로 선별하여 재학습용 데이터 마트에 이관하는 파이프라인이, '데이터 파이프라인 설계서'에 정의된 선별 기준·이관 방식·이관 주기대로 정상 동작하는지를 {통합테스트결과서} 검토와 데이터 마트 직접 확인을 통해 제3자적 관점에서 점검하는 것을 목적으로 한다.

감리원의 점검은 파이프라인 기술 구현의 세부 분석이 아닌 ① '데이터 파이프라인 설계서'에 긍정 피드백 데이터 선별 기준(피드백 유형·임계값·제외 조건), 이관 대상 데이터 항목, 이관 주기(배치·실시간), 이관 후 데이터 마트의 스키마 정의가 명확히 기술되어 있으며 발주기관 승인을 받았는가, ② {통합테스트결과서}에 파이프라인 동작 전 구간(추론 결과 생성 → 피드백 수집 → 선별 → 데이터 마트 이관)에 대한 시험 시나리오와 이관 결과 검증이 포함되어 있는가, ③ 데이터 마트에 이관된 데이터가 설계서에 정의된 스키마·형식·수량과 일치하며 이관 이력이 추적 가능한 형태로 관리되는가의 세 가지 관점에 집중한다.

2. 필요성

▶ AI 모델의 지속 개선은 재학습 데이터 파이프라인의 정상 동작에 직접 의존한다. 배포 후 성능 저하(모델 드리프트)를 감지하고 모델을 재학습하려면 품질 좋은 학습 데이터가 지속적으로 데이터 마트에 공급되어야 하며, 파이프라인이 동작하지 않거나 잘못 구현되면 재학습 데이터가 누락되거나 오염되어 배포 후 AI 성능 개선 자체가 불가능해진다.

▶ 피드백 선별 기준이 명확히 정의되지 않으면 부정적인 피드백 데이터나 이상 데이터(봇 트래픽·스팸 입력)가 재학습 데이터 마트에 포함될 수 있으며, 저품질 데이터로 AI 모델을 재학습하면 모델 성능이 저하되는 역효과가 발생한다.

▶ 이관 이력 미관리 시 재학습 후 AI 성능이 저하된 경우 어느 시점에 어떤 데이터가 이관되었는지 추적할 수 없어 원인 분석이 불가능하다. 이관 이력 관리는 AI 시스템의 책무성(Accountability) 요건으로 감사 추적의 기반이 된다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {통합테스트결과서} 등 산출물이 요구사항에 정의된 기능(데이터 파이프라인 포함)을 설계 의도대로 시험했는지 점검을 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	재학습 파이프라인 설계·발주기관 승인 여부	긍정 피드백 선별 기준(피드백 유형·수치 임계값·제외 조건)·이관 대상 항목·이관 주기·데이터 마트 스키마가 '데이터 파이프라인 설계서'에 정의되고 발주기관 승인을 받았는지 확인
2	파이프라인 전 구간 동작·이관 결과 일치 여부	추론 결과 생성→긍정 피드백→선별 로직→데이터 마트 이관 전 구간 시나리오가 {통합테스트결과서}에 포함되고 이관 건수·항목·형식이 예상 결과와 일치하는지 확인
3	경계값·이상 데이터 제외 시나리오 포함 여부	임계값 직상·직하 데이터의 포함/제외와 봇 트래픽·중복 피드백·스팸 등 이상 데이터 제외 시나리오가 {통합테스트결과서}·{테스트결과서}에 포함되었는지 확인
4	이관 데이터의 스키마 일치 검증 여부	데이터 마트 이관 데이터가 {데이터베이스설계서}의 스키마(테이블 구조·컬럼·형식·NOT NULL·파티션)와 일치하는지 {통합테스트결과서} 이관 결과와 데이터 마트 직접 조회로 검증
5	이관 오류 재처리·중복 방지 시나리오 포함 여부	DB 연결 실패·프로세스 중단 등 이관 오류 시 재처리(Retry) 시나리오가 {통합테스트결과서}에 포함되고 재처리 후 중복 이관 없이 정상 적재되는지 확인
6	이관 이력 기록·추적 가능 여부	이관 시각·건수·선별 기준 버전·처리 결과 코드 등 이관 이력이 {프로그램 명세서} 정의대로 기록되어 특정 시점 이관 결과를 추적할 수 있는지 이력 테이블 조회로 확인
7	이관 데이터 품질 기준·비식별화 방침 정의 여부	이관 데이터 품질 기준(완전성·정확성·일관성)·배포 후 모니터링 방안과 개인정보 비식별화 처리 방침이 '데이터 품질 관리 계획서'·{데이터베이스설계서}에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	선별 기준 수치 정의·승인, 전 구간 통합 시험 포함·이관 결과 일치, 경계값·이상 데이터 시험 포함, 이관 스키마 일치, 오류 재처리 시나리오 포함·중복 이관 없음, 이관 이력 기록·조회 가능, 데이터 품질 기준 정의, 개인정보 비식별화 방침 정의가 모두 확인된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	기본 파이프라인 동작은 확인되나 선별 기준 수치 미정의, 경계값·이상 데이터 제외 시험 누락, 단위 기능만 시험하고 전 구간 통합 시험 미포함, 이관 이력 항목 누락, 재처리 시나리오 미포함, 품질 기준 미정의, 개인정보 비식별화 방침 미정의 등이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	파이프라인 자체가 미구현되어 {통합테스트결과서}에 파이프라인 관련 시험이 전혀 없거나, 데이터 마트에 이관된 데이터가 없거나, 이관 데이터에 이름·이메일·주민번호 등 개인정보가 비식별화 없이 원문 저장된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선별 기준이 수치 없이 정의된 경우 반드시 수치화 요구: "품질 좋은 데이터", "이용자가 만족한 결과"와 같이 수치 임계값 없는 선별 기준은 구현 단계에서 임의로 해석될 수 있으므로 "평점 4점 이상", "체류 시간 30초 이상", "추천 수락 클릭 후 5분 이내 구매" 등 측정 가능한 수치 기준으로 재정의하도록 권고한다. 수치화된 기준은 기준 변경 시 버전 관리가 가능하여 이관 이력 추적에도 활용된다.

2) 단위 기능 시험과 전 구간 통합 시험을 반드시 구분하여 확인: 선별 로직 시험과 이관 로직 시험이 각각 있다고 해서 전 구간 통합 시험이 수행된 것은 아니며, 데이터 형식 불일치·타임스탬프 시간대 불일치·트랜잭션 처리 문제 등은 전 구간 통합 시험에서만 발견되므로 {통합테스트결과서}에 파이프라인 전 구간 시나리오가 반드시 포함되어야 함을 보완권고에 명시한다.

3) 실제 사용자의 누적 데이터가 부족할 수 있으므로, 사업자가 긍정·부정 피드백이 혼재된 모의 데이터(Mock Data) 수만 건을 임의로 주입하여 파이프라인 전 구간의 자동 이관 여부와 선별의 정확성, 형식 불일치 등을 통합 테스트했는지 그 결과를 중점적으로 확인한다.

4) AI 질의 내용 로그의 데이터 마트 원문 저장 여부를 개인정보 관점에서 반드시 확인: 이용자가 입력한 질의 내용이 재학습 데이터로 원문 저장되면 개인정보 보호법 위반 소지가 있으므로, 현장에서 데이터 마트를 직접 조회하여 개인식별 정보 포함 여부를 확인하고, 이관 파이프라인 내에서 주민번호·전화번호·이름 등이 자동 마스킹·삭제(비식별화)되어 적재되는지 시험 결과로 검증한다.

5) 이관 이력에 선별 기준 버전이 없으면 재학습 후 성능 저하 원인 파악 불가: 어느 시점에 어떤 선별 기준으로 이관된 데이터로 학습했는지 추적할 수 있어야 원인 분석이 가능하므로, 이관 이력 항목에 "선별 기준 버전" 필드 추가를 보완권고에 명시한다. 아울러 데이터 마트 읽기 전용 계정 발급 등 접근 권한은 감리 사전조사 단계에서 서면으로 요청한다.

AG_SD44-2-10 | 실시간 성능지표 로깅 점검

응용시스템 | 시험 | 시스템 구현 | 모니터링/재학습

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「전자정부법」·「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 서비스 이용 중 발생하는 이용자 행동 지표(클릭·체류시간·페이지 이동·AI 추론 요청·추천 항목 반응 등)가 '로그 설계서'에 정의된 항목과 형식에 따라 정확히 수집되어 ELK 스택 등 로그 적재 플랫폼에 빠짐없이 저장되는지를 {시스템로그} 검토와 수집 기능 직접 시험을 통해 제3자적 관점에서 점검하는 것을 목적으로 한다.

감리원의 점검은 로그 수집 기술 구현의 세부 분석이 아닌 다음 세 가지 관점에 집중한다.

① '로그 설계서'에 수집할 이용자 행동 지표 항목(이벤트 유형·수집 필드·형식·보존 기간·개인정보 처리 방침)이 명확히 정의되어 있으며 발주기관 승인을 받았는가, ② {시스템로그}에서 실제 수집·적재된 로그 데이터가 '로그 설계서'에 정의된 항목과 일치하고 클릭·체류 등 주요 행동 이벤트가 누락 없이 적재되어 있는가, ③ {테스트결과서}에 로그 수집 기능의 시험(항목 완전성·정확성·적시성·개인정보 비식별화) 결과가 수치 근거와 함께 기재되어 있는가.

2. 필요성

▶ 실시간 성능(행동) 지표 로깅 검증 점검이 필요한 이유는 다음과 같다.

▶ AI 서비스의 지속적 개선은 이용자 행동 로그 데이터 품질에 의존함: AI 추천·검색·생성 기능의 성능은 배포 후 수집된 이용자 행동 데이터를 기반으로 모델을 재학습하거나 알고리즘을 조정하는 방식으로 지속 개선된다. 클릭·체류·이탈 로그가 누락되거나 부정확하게 수집되면 AI 모델 재학습 데이터의 품질이 저하되어 서비스 개선이 지연되거나 왜곡된다.

▶ 로그 항목 누락 또는 형식 불일치는 배포 후 데이터 분석 파이프라인 전체에 영향을 미침: 로그 수집 단계에서 특정 필드가 누락되거나 데이터 형식이 설계와 다르면, 이후 ELK 스택의 인덱스 매핑 오류·집계 쿼리 실패·대시보드 지표 오류로 연쇄적으로 문제가 전파된다. 이 문제는 데이터가 충분히 쌓인 후에야 발견되는 경우가 많아 배포 전 검증이 필수적이다.

▶ 이용자 행동 로그에 개인정보가 포함될 경우 개인정보 보호법 위반 위험이 있음: 클릭·체류·검색 로그에 이용자 식별 정보(이름·이메일·IP 주소 원문 등)가 비식별화 처리 없이 로그 저장소에 그대로 적재되면 개인정보 보호법 위반 소지가 있다. 특히 ELK 스택과 같이 다수의 운영자가 접근 가능한 로그 플랫폼에서 이 문제는 더욱 심각한 유출 위험으로 이어진다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {시스템로그} 등 산출물이 요구사항에 정의된 기능을 설계 의도대로 구현하고 시험했는지 점검을 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	로그 수집 이벤트·필드·처리 방침 설계·승인 여부	수집 이벤트 유형·이벤트별 필드·데이터 형식·샘플·적재 지연 허용 기준·개인정보 처리 방침이 '로그 설계서'에 정의되고 발주기관 승인을 받았는지 확인
2	이벤트 로그의 정의 필드 적재·형식 일치 여부	클릭·체류·AI 추론 요청 이벤트 로그가 '로그 설계서'의 모든 필드를 포함하여 {시스템로그}에 적재되고 필드별 형식이 명세와 일치하는지 ELK Kibana 등에서 샘플 확인
3	로그 적재 지연 허용 기준 이내 여부	이벤트 발생 후 로그 적재까지의 지연 시간이 'AI 서비스 요구사항 정의서'의 허용 기준 이내인지 {테스트결과서}와 {시스템로그} 시각 대조로 확인
4	로그 중복·누락 방지 시험 여부	동일 이벤트 중복 적재·이벤트 누락 여부를 {시스템로그} 샘플로 대조하고 중복 방지·누락 시나리오(네트워크·클라이언트 오류) 시험이 {테스트결과서}에 포함되었는지 확인
5	행동 로그 개인정보 비식별화 여부	이용자 행동 로그에 이름·이메일·IP 원문·기기 식별자 원문 등 개인정보가 비식별화 없이 저장되지 않고 {프로그램 명세서}에 비식별화 로직이 정의되었는지 {시스템로그} 샘플로 확인
6	행동 로그·추론 로그 상관관계 추적 가능 여부	AI 추론 로그(요청·응답 시각·소요시간·기능명·결과 코드)가 {시스템로그}에 적재되고 이용자 행동 로그와 동일 이벤트 ID 또는 세션 ID로 연결되는지 확인
7	성능 지표 조회·집계 대시보드 구성 여부	행동 로그 기반 주요 AI 성능 지표(클릭률·체류시간·AI 추천 수락률·응답 만족도)를 조회·집계할 수 있는 대시보드 또는 조회 기능이 구성되었는지 {시스템로그} 플랫폼에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	'로그 설계서'에 정의된 이벤트·필드가 {시스템로그}에 누락·중복 없이 적재되고, 적재 지연이 허용 기준 이내이며, AI 추론 로그와 이용자 행동 로그가 동일 이벤트 ID·세션 ID로 연결되어 상관관계 추적이 가능하고, 개인정보 비식별화와 지표 집계 기능이 모두 확인되는 경우	현행 유지
보완권고 (△)	로그 수집 체계는 구현되었으나 일부 필드 누락·형식 불일치, 적재 지연·누락 시나리오 시험 미수행, 지표 대시보드 미구성 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	이용자 행동 로그가 수집·적재되지 않거나 개인정보가 비식별화 없이 원문 적재되는 등 로깅 체계가 실질적으로 미구현된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) ELK 로그 조회 접근 계정은 현장감리 전에 반드시 사전 확보한다. 현장 당일 계정 발급 요청 시 시간이 지체되어 점검이 불가능해질 수 있으므로, 사전조사 단계에서 읽기 전용 계정과 주요 인덱스 목록 제공을 서면으로 요청하고, ELK 버전별 조회 방법도 미리 확인한다.

2) 개인정보 비식별화 미처리 확인 시 보완권고가 아닌 배포 전 반드시 완료해야 할 조건으로 명시한다. 특히 AI 질의 내용 로그에 이용자가 입력한 개인정보(이름·주민번호 등)가 원문 저장되는 경우는 심각한 개인정보 위험으로 판단하고 별도로 강조하여 기재한다.

3) 로그 형식 불일치(타임스탬프 형식 등)는 인덱스 매핑 오류와 집계 오류로 연결되므로, 필드별 데이터 형식이 '로그 설계서'의 명세와 일치하는지 샘플 로그를 통해 반드시 확인한다.

AG_SD44-2-11 | 외부 연계시스템 통합시험 점검

응용시스템 | 시험 | 시스템 구현 | 외부연계

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업의 종료(시험/전개) 단계에서 {통합테스트결과서}를 중심으로 외부 연계 시스템의 통합 시험이 적정하게 수행되었는지를 점검하는 데 목적이 있다.

구체적으로는 AI 특화 연계 항목(환각 통제 시나리오, 외부 API 장애대응, 이상응답 처리, 데이터 송수신 정합성)이 통합시험에 반영되었는지, 발견된 결함이 적정히 관리·조치되었는지, 그리고 배포·전개 방안이 AI 서비스 특성을 고려하여 수립되었는지를 문서 검토 관점에서 확인하는 것이 목적이다.

2. 필요성

▶ AI 서비스는 일반 IT 시스템과 달리 외부 AI API, 벡터 데이터베이스, 언어 모델 추론 엔진 등 복수의 비결정론적(Non-deterministic) 외부 연계 구성 요소를 포함한다. 이러한 구성 요소는 동일 입력에도 출력이 달라질 수 있어, 전통적인 통합시험 방식만으로는 AI 서비스의 품질을 보장하기 어렵다.

▶ 또한 생성형 AI 서비스는 환각(사실과 다른 응답 생성), 프롬프트 인젝션(악의적 입력에 의한 시스템 탈출), 민감정보 외부 유출 등 AI 고유의 위험 시나리오를 통합시험 단계에서 반드시 검증해야 한다. 이를 문서로 증빙하지 않을 경우, 운영 전환 이후 중대한 품질 문제로 이어질 수 있다.

▶ 감리원은 기술적 구현의 적정성보다는 계획 대비 시험 수행 여부, AI 특화 시나리오의 반영 여부, 결함 조치 이력의 완결성, 배포 계획의 구체성을 산출물을 통해 확인하는 역할을 수행한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 연계 구간 시나리오 반영 여부	• {통합테스트결과서}에 외부 AI API(LLM, RAG, 외부 데이터소스 등) 연계 구간에 대한 테스트 시나리오가 별도로 구성되어 있는지 확인
2	환각 및 이상응답 처리 확인	• 생성형 AI 출력값 중 근거 없는 사실을 생성(환각)하거나 유해·부적절 응답을 반환하는 시나리오에 대해 통합시험이 수행되었는지 {통합테스트결과서}를 통해 확인

No	점검 항목	확인 방법
3	외부 API 장애·타임아웃 대응 시험	• 외부 AI API 서비스(예: LLM API 공급사, 검색 API, 벡터 DB 등)가 응답 불가·타임아웃 상태일 때 시스템이 정상적으로 폴백(Fallback) 처리하거나 사용자에게 장애 안내를 제공하는지 시험되었는지 확인
4	데이터 송수신 정합성 및 형식 검증	• AI 연계 구간에서 송수신되는 데이터(입력 프롬프트, 출력 응답, 메타데이터)의 인코딩·형식·필드 누락 등이 {통합테스트결과서}에서 검증되었는지 확인
5	RTM 연계 및 요구사항 충족 여부 추적	• {통합테스트결과서}의 테스트 케이스가 {RTM(요구사항추적매트릭스)}의 AI 관련 요구사항 항목과 1:N으로 매핑되어 있는지 확인
6	결함 관리 및 조치 이력 적정성	• 통합시험에서 발견된 결함이 결함관리대장(또는 이슈관리시스템)에 등록되고, 중결함·경결함으로 분류되어 조치 우선순위가 설정되어 있는지 확인
7	배포·전개 방안 수립 적정성	• AI 서비스 배포 계획서(또는 '전개 계획서')에 AI 모델 버전 관리, 롤백(Rollback) 절차, 운영 전환 체크리스트가 포함되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	통합테스트 수행 계획과 결과가 충실히 문서화되어 있고, AI 특화 연계 항목(환각 통제·장애대응·이상응답 기록)이 {통합테스트결과서}에 명확히 반영되며, AI 연계 구간의 데이터 송수신 정합성·형식 검증이 수행되고, 테스트 케이스가 {RTM}의 AI 관련 요구사항과 매핑되며, AI 모델 버전 관리·롤백 절차·운영 전환 체크리스트를 포함한 배포·전개 방안이 수립되어 주요 결함이 해소된 경우	현행 유지
보완권고 (△)	통합테스트 결과가 존재하나 AI 특화 시나리오(외부 API 장애 대응, 이상응답 처리)가 누락되거나, 결함 내역이 불명확하고 조치 계획만 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{통합테스트결과서}가 미비하거나, AI 연계 구간 테스트가 수행되지 않았거나, 중결함 이상이 미해소 상태로 종료 단계에 진입한 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 고유 이상응답(환각·부적절 출력) 시나리오가 통합시험 케이스에 포함되었는지 확인한다: 일반 SW 통합시험과 달리 AI는 예측 불가한 출력이 발생할 수 있으므로 해당 상황의 처리 로직이 시험 대상이어야 한다.

2) 외부 API 응답 지연·장애 시 AI 서비스의 동작(graceful degradation) 시험이 포함되었는지 점검한다: 연계 장애 시 AI가 이전 결과를 반환하는지, 오류 안내를 하는지 등이 명확히 시험되어야 한다.

3) 통합 시험 결과를 검토할 때 '통과' 여부만 확인하지 말고 Pass 기준의 타당성을 검토한다: AI 정확도 기준이 비현실적으로 낮게 설정되어 통과 처리되는 경우가 있다.

4) 결함 조치 이력의 완결성을 확인한다: 중결함이 미해소 상태로 남아 있거나 조치 결과에 대한 재시험 기록이 없으면 통합시험은 완료된 것이 아니다. 감리원은 결함관리대장의 등급 분류·조치·재시험 이력이 {통합테스트결과서}와 일치하는지 확인한다.

5) 배포·전개 방안의 구체성을 함께 점검한다: AI 서비스는 모델 버전 교체가 곧 서비스 변경이므로 모델 버전 관리와 롤백 절차가 없으면 장애 시 복구 수단이 없다. 감리원은 배포 계획서에 모델 버전 관리, 롤백 절차, 전개 후 모니터링 방안이 정의되었는지 확인한다.

AG_SD44-2-12 | API 연계데이터 정합성 점검

응용시스템 | 시험 | 시스템 구현 | 외부연계

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 서비스와 연계 응용시스템 간 API를 통해 송수신되는 데이터가 포맷 오류 및 필드 누락 없이 설계 명세에 따라 정확하게 전달되는지를 {통합테스트결과서}를 통해 제3자적 관점에서 검토하는 것을 목적으로 한다.

감리원의 점검은 API 구현 방식이나 통신 기술의 세부 분석이 아닌 ① 'API 연계 설계서' 또는 '인터페이스 정의서'에 연계 대상 시스템별 API 명세(요청·응답 필드, 데이터 타입, 필수 여부)가 정의되어 있으며 {통합테스트결과서}에서 각 API의 정합성 시험이 수치로 확인되는가, ② 포맷 오류(데이터 타입 불일치·값 범위 초과) 및 필드 누락(필수 필드 미전송·응답 필드 미수신)에 대한 시험이 정상 케이스와 오류 케이스로 구분되어 수행되었는가, ③ API 연계 오류 발생 시 재처리·재시도 로직과 이용자 안내 방식이 {프로그램 명세서}에 정의되어 있으며 {통합테스트결과서}에서 시험이 확인되는가의 세 가지 관점에 집중하여, 설계 산출물과 {통합테스트결과서}의 일치 여부와 시험의 충분성을 확인한다.

2. 필요성

▶ AI 서비스는 다수 연계 시스템에 의존하여 데이터 정합성 오류의 연쇄 영향이 크다. 회원정보 시스템, 상품 데이터베이스, 로그 수집 시스템, 외부 AI 모델 API 등 어느 한 연계 지점에서 포맷 오류나 필드 누락이 발생하면 AI 모델의 입력 데이터가 오염되어 잘못된 추론 결과가 생성되며, 이 오류가 다음 연계 시스템으로 전파된다.

▶ 배포 후 API 연계 오류 발견은 수정 비용이 배포 전보다 현저히 높다. API 연계 데이터 포맷 오류는 단위시험에서 발견되지 않고 두 시스템이 실제로 연동될 때 비로소 나타나는 경우가 많으며, {통합테스트결과서}에서 연계 데이터 정합성이 확인되지 않은 상태로 배포되면 운영 중 오류 발생 시 서비스 중단과 데이터 정정에 상당한 비용과 시간이 소요된다.

▶ AI 특화 연계 데이터(모델 입력 벡터·추론 결과·신뢰점수 등)는 부동소수점 정밀도, 배열 차원, Null 값 처리 방식 등 일반 업무 데이터보다 훨씬 정밀한 포맷 준수가 요구되므로, AI 서비스 연계 시험에서 AI 특화 데이터 타입 시험을 별도로 수행해야 한다.

▶ 감리기준 준수: 정보시스템 감리기준 제8조(종료단계 감리)는 {통합테스트결과서} 등 산출물을 통해 시스템 간 연계 기능이 적정하게 시험되었는지 여부를 주요 점검 항목으로 규정한다.

3. 검토방법

No	점검 항목	확인 방법
1	API 명세 및 AI 특화 필드 정의 여부	연계 대상 시스템별 API 명세(요청·응답 필드명·데이터 타입·필수 여부·값 범위)와 추론 결과·신뢰점수·모델 버전 등 AI 특화 데이터 필드가 'API 연계 설계서'에 정의되었는지 확인
2	정상·오류 케이스 시험 및 통과율 기재 여부	정상 케이스와 오류 케이스(포맷 오류·필수 필드 누락·값 범위 초과·Null 처리) 시험이 각각 독립 항목으로 {통합테스트결과서}에 포함되고 케이스별 통과율(%)이 수치로 기재되었는지 확인
3	포맷 오류율·필드 누락률 허용 기준 충족 여부	포맷 오류율·필드 누락률이 'AI 서비스 요구사항 정의서'의 허용 기준값(예: 오류율 1% 이하)을 충족하는지 {통합테스트결과서}와 대조
4	AI 특화 데이터 필드 시험 포함 여부	신뢰점수 소수점 정밀도·배열 차원·Null 값 허용 여부 등 AI 모델 API 관련 특화 데이터 필드 시험이 {통합테스트결과서}에 별도 항목으로 포함되었는지 확인
5	연계 오류 재시도·대체 처리 로직 정의·시험 여부	포맷 오류·타임아웃·응답 없음 등 API 연계 오류 시 재시도 로직·재시도 횟수·간격·최대 초과 시 대체 처리 방안이 {프로그램 명세서}에 정의되고 {통합테스트결과서}에서 시험되었는지 확인
6	연계 오류 안내 메시지 설계 일치 여부	API 연계 오류 시 이용자에게 표시되는 안내 메시지가 'API 연계 설계서'에 정의된 내용과 일치하게 구현되었는지 {시스템화면}·{통합테스트결과서}에서 확인
7	연계 오류 이력 기록 여부	오류 발생 일시·API 경로·오류 코드·재시도 횟수·처리 결과 등 API 연계 오류 이력이 {프로그램 명세서} 정의대로 기록되는지 {통합테스트결과서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 특화 필드를 포함한 API 명세의 완전성, 정상·오류 케이스 구분 수치 시험과 허용 기준값 충족, AI 특화 데이터 타입 시험 포함, 재시도·폴백 등 오류 처리 로직 정의 및 시험, 오류 메시지 설계 일치, 연계 오류 이력 기록이 모두 확인된 경우	현행 유지
보완권고 (△)	기본 API 통신은 동작하나 "연계 정상 확인" 등 정성 표현만 기재, AI 특화 데이터 시험 미수행, 오류 케이스 시험 미구분·수치 미기재, 재시도 횟수·간격 미정의, 오류 메시지 내부 정보 노출, 이력 항목 누락 등이 있는 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	API 연계 자체가 미구현되었거나, 정상 케이스 시험에서 포맷 오류 또는 필수 필드 누락이 다수 발생하거나, 오류율이 요구사항 기준값을 현저히 초과하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "연계 정상 확인" 정성 표현은 시험 완료 불인정 — 케이스별 수치 필수: {통합테스트결과서}에 "API 연계 정상", "데이터 전송 확인" 등 정성 표현만 있는 경우 오류 케이스별 통과율을 확인할 수 없으므로, 정상·오류 케이스를 구분하고 케이스별 시험 건수·통과율(%)을 수치로 재기재하도록 보완권고를 제시하며 단순 정성 확인은 시험 완료로 인정하지 않는다.

2) AI 특화 데이터 시험은 일반 API 시험과 반드시 구분하여 점검: 신뢰점수 소수점 정밀도, 추론 결과 배열 크기, Null 값 처리 등 AI 특화 데이터 타입 시험이 일반 API 시험과 동일 기준으로만 수행된 경우 AI 연계 데이터의 고유 위험이 검증되지 않은 것으로 판단하고, AI 모델 API 시험 항목이 {통합테스트결과서}에 별도로 포함되어 있는지 반드시 확인한다.

3) 재시도 로직 미정의는 연계 오류 시 서비스 완전 중단 위험: {프로그램 명세서}에 재시도 횟수·간격이 정의되지 않으면 연계 시스템의 일시적 장애 시 AI 서비스가 즉시 오류 응답을 반환하거나 무한 대기 상태에 빠질 수 있으므로, 재시도 횟수·간격·타임아웃 수치 정의와 폴백 처리 방안을 배포 전 완료 조건으로 보완권고에 포함한다.

4) 오류 메시지 내부 정보 노출은 보안 관점 부적합 — 반드시 현장 확인: API 오류 메시지에 API 경로, HTTP 오류 코드, 스택 트레이스 등이 이용자 화면에 그대로 노출되면 시스템 구조 정보가 노출되는 보안 취약점이 되므로, {통합테스트결과서} 검토와 함께 반드시 {시스템화면}에서 직접 오류를 유발하여 표시되는 메시지를 현장 확인한다.

5) 연계 대상이 많을 경우 AI 모델 API를 최우선 중점 점검 대상으로 선정: AI 추론 요청·결과 수신 API, 개인정보 포함 데이터 연계 API, 금융·의료 등 고위험 도메인 연계 API를 최우선 중점 대상으로 선정하고 나머지 연계 API는 대표 항목 기준으로 점검 범위를 조정한다. 특히 AI 모델 연동 API는 처리 시간이 길므로 네트워크 타임아웃 발생 시 데이터 유실 없이 재처리되거나 적절한 오류를 반환하는지 예외 처리(Resilience) 정합성 시험 여부를 확인한다.

4.4. 데이터베이스

학습·운영 데이터의 품질, 비식별화, 계보·형상관리 점검 가이드를 수록한다. (총 45건)

요구정의·분석 (11건)

AG_SD41-3-01 특성분석 및 선정기준 요건

AG_SD41-3-02 공통코드 명세요건 및 AI모델 공통분류 코드체계 요구 여부

AG_SD41-3-03 데이터계보 추적요건 점검

AG_SD41-3-04 메타데이터 확보 관리 요건

AG_SD41-3-05 AI 의사결정 감사 추적(Audit Trail) 요건 점검

AG_SD41-3-06 민감정보 및 개인 식별번호 비식별화 마스킹 및 가명-익명처리요건

AG_SD41-3-07 기존데이터 마이그레이션 및 레거시 데이터 이관 검증 요건

AG_SD41-3-08 학습-검증 테스트 분할비율 및 데이터셋 분할기준 명시 여부

AG_SD41-3-09 AI 학습데이터 품질기준 명시

AG_SD41-3-10 벡터 DB 및 RAG 파이프라인 요구정의 적정성 점검

AG_SD41-3-11 데이터 분할 계획 및 검증용(Hold-out) 데이터 확보

설계 (11건)

AG_SD42-3-01 편향방지 피처엔지니어링 점검

AG_SD42-3-02 학습데이터 버전-모델 매핑점검

AG_SD42-3-03 비식별화-가명화 뷰 설계점검

AG_SD42-3-04 데이터 정합성 검증 배치 설계점검

AG_SD42-3-05 개인정보 암호화 컬럼 설계점검

AG_SD42-3-06 테스트용 데이터관리 점검

AG_SD42-3-07 공공데이터 연계·활용 요건 점검

AG_SD42-3-08 DB 성능 최적화 실행계획 설계점검

AG_SD42-3-09 EDB-벡터 DB-하이브리드 데이터모델 정합성

AG_SD42-3-10 성능지표 데이터마트 설계점검

AG_SD42-3-11 정제 전후 적재공간 분리점검

구현 (10건)

AG_SD43-3-01 편견금칙어 사전 적용점검

AG_SD43-3-02 공통코드 데이터정합성 점검

AG_SD43-3-03 데이터계보 이력저장 확인점검

AG_SD43-3-04 레거시데이터 이행 ETL 검증 점검

AG_SD43-3-05 이상데이터 실제 제거율점검

AG_SD43-3-06 데이터정합성 Null형식 배치검증 점검

AG_SD43-3-07 외부크롤링 적재 무결성점검

AG_SD43-3-08 모델 추론결과 프롬프트 응답 로깅점검

AG_SD43-3-09 정제 전후DB분리 환경점검

AG_SD43-3-10 데이터 소스변경 모니터링 알림구현 적정성 점검

시험 (13건)

AG_SD44-3-01 보호변수 배제 분리 보관 확인점검

AG_SD44-3-02 학습데이터 버전관리 점검

AG_SD44-3-03 개인정보 비식별화 및 파기 점검

AG_SD44-3-04 데이터 옵트아웃 동의 처리점검

AG_SD44-3-05 대용량백업 스케줄러 실행점검

AG_SD44-3-06 악성노이즈-중독데이터 필터점검

AG_SD44-3-07 관리자-일반사용자 DB권한 통제점검

AG_SD44-3-08 파일-스토리지 민감컬럼 암호화점검

AG_SD44-3-09 성능평가 마트 OLAP 집계검증 대시보드 통계 정확성점검

AG_SD44-3-10 샘플링클래스 불균형 해소점검

AG_SD44-3-11 데이터 모델ERD물리DB 일치점검

AG_SD44-3-12 실시간 스트리밍적재 파이프라인점검

AG_SD44-3-13 인덱스-파티셔닝 효율성 점검

AG_SD41-3-01 | 특성분석 및 선정기준 요건

데이터베이스	요구정의·분석	데이터 수집·처리	편향/공정성
근거	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP		

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용할 특성(Feature) 변수의 선정 기준과 선정 과정의 객관성·투명성·문서화 요건이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 특성 변수 선정의 자의성을 방지하고, 선정 근거가 도메인 지식·통계적 기준·윤리적 기준에 기반하여 문서화되었는지 확인한다.

2. 필요성

▶ AI 모델의 예측 품질은 학습에 사용하는 특성 변수의 적합성에 직접 의존한다. 특성 변수가 도메인 맥락과 무관하게 임의로 선정되거나, 차별을 유발하는 보호 속성(성별·나이·인종 등)이 직간접적으로 포함되거나, 미래 데이터에서 구할 수 없는 변수(데이터 누수)가 포함되면 모델은 잘못된 패턴을 학습하여 서비스 운영에서 심각한 오류를 유발한다.

▶ 자의적 특성 선정으로 인한 모델 과적합·과소적합 위험 방지: 특성 선정 기준 없이 개발자 직관에만 의존하면 타깃 변수와 무관한 노이즈 변수를 포함하거나 핵심 변수를 누락시켜 모델이 학습 데이터에만 최적화되고 실제 서비스 데이터에서 성능이 급락하는 과적합이 발생할 수 있다.

▶ 데이터 누수(Data Leakage)에 의한 허위 성능 지표 방지: 모델 평가 시점에는 알 수 없는 미래 정보나 타깃 변수 자체와 직결된 파생 변수가 특성으로 포함되면 학습·검증 단계에서 비현실적으로 높은 성능이 측정되지만 실제 서비스에서는 해당 정보를 사용할 수 없어 급격한 성능 저하가 발생한다. 데이터 누수 방지 요건이 명시되어야 이를 근원적으로 차단할 수 있다.

▶ 차별 유발 보호 속성 포함 위험 차단: 성별·나이·국적·장애 여부 등 법적으로 보호되는 속성이나 이와 높은 상관관계를 가진 프록시 변수(Proxy Variable)가 특성으로 포함되면 모델이 특정 집단에 불리한 결과를 체계적으로 생성하는 알고리즘 차별이 발생한다. 보호 속성 배제 요건이 {요구사항정의서}에 명시되어야 한다.

▶ TTA 신뢰 AI 및 NIST AI RMF 공정성·투명성 요건 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI·일반 분야)는 학습 특성 변수 선정의 공정성·투명성 확보를 신뢰 AI 요건으로 명시하며, NIST AI RMF MAP 2.3은 특성 선정 과정의 위험 요소 식별과 문서화를 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	후보 특성 변수 목록 및 선정 기준 문서화 요건	알고리즘 학습에 사용할 후보 특성 변수 전체 목록과 각 변수의 원천 테이블·컬럼 매핑, 선정 근거(도메인 전문가 지식·통계적 상관관계·비즈니스 규칙 등)가 {요구사항정의서} 또는 연계 산출물에 정의되었는지 확인한다. '적절한 특성을 선정한다'는 선언적 서술은 미정의로 판단한다.
2	통계적 특성 중요도 평가 방법 및 기준 수치 정의	특성 변수의 타깃 변수 예측 기여도를 평가하는 통계적 방법(상관계수·정보 이득·SHAP 값·분산 임계값 등)과 선정/제외 판단에 사용하는 기준 수치(예: 분산 0.01 미만 제거, 상관계수 0.05 이상 유지)가 {요구사항정의서}에 정의되었는지 확인
3	데이터 누수(Data Leakage) 방지 요건 정의	모델 예측 시점에 실제로 사용할 수 없는 미래 정보·타깃 파생 변수·시험 데이터 기반 생성 변수가 학습 특성에 포함되지 않도록 하는 데이터 누수 검토 절차와 배제 기준이 비기능 요건으로 정의되었는지 확인
4	보호 속성 및 프록시 변수 배제 요건 정의	성별·나이·인종·장애 여부 등 법적 보호 속성과 이와 높은 상관관계를 가진 프록시 변수(우편번호·이름 등)의 특성 포함 배제 요건, 배제 여부 검토 방법(보호 속성과 특성 간 상관관계 분석 등)이 {요구사항정의서}에 정의되었는지 확인
5	임베딩 모델·벡터 특성 선정 요건	정형 특성 외에 RAG·벡터 검색에 사용하는 임베딩 모델·차원·거리척도(cosine/dot/L2)의 선정 근거(도메인·언어 적합성)가 문서화 요건으로 정의되었는지 확인
6	RFP 대비 특성 선정 요건 완전성 추적	RFP에 명시된 특성 변수 선정·데이터 누수·보호 속성 배제 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	후보 특성 목록·원천 컬럼 매핑·선정 근거, 통계적 중요도 평가 방법·기준 수치, 데이터 누수 유형별 검토 절차, 보호 속성 및 프록시 변수 배제 요건, RAG·벡터 검색 활용 시 임베딩 모델·차원·거리척도(cosine/dot/L2)의 선정 근거 문서화 요건이 {요구사항정의서}에 모두 수치 기반으로 정의됨	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	후보 특성 목록은 있으나 선정 근거 미문서화, 통계적 기준 수치 미정의, 데이터 누수 검토 절차 또는 보호 속성 배제 요건 중 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	특성 변수 선정 기준·데이터 누수 방지·보호 속성 배제 요건이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'AI 전문가가 알아서 선정한다'는 답변은 가장 위험한 신호: 특성 변수 선정 기준을 '개발팀 내부적으로 결정한다', 'AI 전문가가 경험으로 판단한다'고 답변하는 경우가 가장 많다. 이는 선정 기준이 문서화되지 않아 추후 모델 오류 발생 시 원인 분석과 책임 소재 파악이 불가능하다는 의미이다. 감리원은 특성 선정 근거 문서화가 {요구사항정의서}에 비기능 요건으로 반드시 포함되도록 유도해야 한다.

2) 데이터 누수는 모델 성능이 '너무 좋을 때' 의심해야: 학습·검증 단계에서 모델 정확도가 비현실적으로 높게(95% 이상) 측정되는 경우 데이터 누수 가능성을 최우선으로 의심해야 한다. 감리원은 {요구사항정의서}에 데이터 누수 검토 절차가 포함되었는지 확인하고, 특히 시계열 데이터나 사기 탐지·이상 탐지 모델에서 시간적 누수 방지 요건이 별도로 정의되었는지 집중 점검해야 한다.

3) 우편번호·지역명·이름은 프록시 변수 위험 최우선 점검 대상: 우편번호·행정구역·이름·학교명·직업명 등은 인종·소득·성별 등 보호 속성과 높은 상관관계를 가지는 대표적 프록시 변수이다. 이 변수들이 후보 특성 목록에 포함되어 있다면 보호 속성과의 상관관계 분석 결과를 'EDA 보고서'에서 확인하고, 상관관계 분석 없이 포함된 경우 보완 권고로 기록한다.

4) 다중공선성 방치는 모델 해석 불가능으로 이어짐: 서로 높은 상관관계를 가진 특성 변수들이 동시에 포함되면 개별 변수의 모델 기여도 해석이 불가능해진다. 특히 감리·규제 기관에 모델 설명을 요구받는 고위험 도메인(금융·의료·채용)에서는 다중공선성 탐지 요건(VIF 분석 등)이 {요구사항정의서}에 명시되도록 유도해야 하며, 누락 시 보완 권고로 기록한다.

5) 전처리(정규화·인코딩) 기준의 훈련/검증 분리 적용 누락은 누수 원인: 정규화(Min-Max·표준화)나 인코딩을 전체 데이터 기준으로 먼저 수행한 뒤 훈련/검증 데이터로 분리하면 검증 데이터의 통계 정보가 훈련에 이미 반영되어 데이터 누수가 발생한다. 훈련 데이터 기준으로 전처리 파라미터를 산출하고 검증·테스트 데이터에 동일 파라미터를 적용하는 절차가 {요구사항정의서}에 명시되었는지 확인해야 한다.

AG_SD41-3-02 | 공통코드 명세요건 및 AI모델 공통분류 코드체계 요구 여부

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · 「전자정부법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델이 학습·추론·결과 저장에 활용하는 공통 분류 코드 체계(AI 모델 출력 분류 코드·입력 범주 코드·상태 코드 등)의 코드 집합 정의·코드 값 명세·코드 관리 체계가 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

공통 코드 미정의로 인한 AI 모델 분류 결과 해석 오류·시스템 간 코드 불일치·운영 단계 코드 증식을 사전에 차단하는 요건 정의 여부를 확인한다.

2. 필요성

▶ AI 서비스 구축 사업에서 공통 코드는 AI 모델의 입력 데이터 전처리·라벨 부여·추론 결과 해석·후속 업무 시스템 연계 전 과정에 걸쳐 사용되며, 요구정의 단계에서 공통 코드 체계가 명확히 정의되지 않으면 AI 모델이 생성하는 분류 결과 코드가 업무 시스템에서 인식되지 않거나 운영 단계에서 코드 항목이 무분별하게 추가·증식되는 위험이 있다.

▶ AI 모델 분류 결과 코드 미정의로 인한 후속 업무 연계 오류 방지: AI 모델이 출력하는 분류 결과(예: 이상 거래 탐지 결과 코드 '01=정상/02=의심/03=이상')가 {요구사항정의서}에 정의되지 않으면 후속 업무 시스템이 코드 값 의미를 임의 해석하거나 연계 인터페이스를 잘못 설계하여 운영 단계에서 업무 처리 오류가 발생할 수 있다.

▶ AI 학습 데이터 라벨 코드 불일치로 인한 모델 편향 방지: 라벨 코드가 공통 코드로 정의되지 않으면 라벨링 작업자마다 코드 값을 다르게 해석하거나 다른 코드 값을 동일 의미로 혼용하여 학습 데이터 라벨이 오염되므로, 공통 코드 명세가 요건으로 확정되어야 라벨링 품질을 보장할 수 있다.

▶ 운영 단계 코드 증식(Code Proliferation)으로 인한 모델 재학습 품질 저하 방지: 공통 코드 관리 체계 없이 운영하면 업무 변경 시마다 유사한 코드 항목이 추가·중복 생성되어 AI 모델 재학습 시 코드 의미 혼재로 인한 학습 데이터 오염이 누적되므로, 요구정의 단계에서 코드 추가·변경·폐기 관리 절차가 요건으로 정의되어야 한다.

▶ 다중 AI 모델 간 공통 분류 코드 불일치로 인한 앙상블·연계 오류 방지: 여러 AI 모델이 연계·앙상블로 구성된 경우 각 모델의 출력 코드 체계가 통일된 공통 코드 기준으로 정의되지 않으면 모델 간 출력 코드 불일치로 앙상블 집계 및 결과 해석 오류가 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 입출력 공통 코드 집합 식별 및 목록 정의	AI 모델 학습 라벨 코드·입력 범주형 특성 코드·추론 결과 분류 코드·상태 코드 등 AI 시스템에서 공통으로 사용되는 코드 집합 목록이 {요구사항정의서} 또는 '공통 코드 정의서'에 식별·정의되었는지 확인
2	코드 집합별 코드 값·코드 의미·유효 기간 명세 완전성	각 코드 집합의 코드 값(코드 키)·코드 의미(한글 설명)·유효 기간(사용 시작일·종료일)·사용 가능 여부(활성/비활성)·코드 길이·코드 순서와 예외 코드(미분류·기타·오류 코드 등)의 정의가 '공통 코드 정의서'에 완전히 명세되었는지 확인
3	AI 모델 분류 결과 코드와 업무 시스템 연계 코드 일치 여부 요건 정의	AI 모델이 출력하는 분류 결과 코드가 후속 업무 시스템(ERP·CRM·심사 시스템 등)에서 사용하는 코드 체계와 일치하는지 또는 변환 매핑이 필요한지를 {요구사항정의서}에 요건으로 정의하였는지 확인
4	공통 코드 추가·변경·폐기 관리 절차 요건 정의	운영 중 업무 변경으로 인한 코드 추가·변경·폐기 시 AI 모델 재학습 영향 분석과 코드 관리 승인 절차(요청-검토-승인-배포)가 {요구사항정의서}에 요건으로 정의되었는지 확인
5	벡터 검색 메타·필터 코드 체계 요건	벡터 검색 필터링에 사용하는 문서 분류·접근권한·출처 코드 체계가 업무 시스템 공통 코드와 일치하도록 정의되었는지 확인
6	RFP 대비 공통 코드 명세 요건 완전성 추적	RFP에 명시된 공통 코드 정의·AI 모델 분류 코드 체계·코드 관리 절차 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 모델 관련 공통 코드 집합 목록, 코드 값·코드 의미·유효 기간 명세, 예외 코드 정의, AI 모델 출력-업무 시스템 코드 일치 또는 매핑 요건, 벡터 검색 필터링에 사용하는 문서 분류·접근권한·출처 코드 체계의 업무 시스템 공통 코드 정합 요건, 코드 관리 절차(추가·변경·폐기 시 재학습 영향 분석)가 {요구사항정의서} 및 관련 산출물에 모두 정의됨	현행 유지
보완권고 (△)	공통 코드 집합 목록은 식별되었으나 코드 의미 명세 일부 누락, 예외 코드 미정의, 코드 관리 절차 미정의, 후속 업무 시스템 코드 매핑 요건 누락 등 일부 요건 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 모델 관련 공통 코드 집합이 {요구사항정의서}에 전혀 식별·정의되지 않거나 '공통 코드를 정의하여 사용한다'는 선언적 서술만 존재	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 모델 추론 결과 코드는 요구정의 단계에서 확정되어야 하는 핵심 설계 요소: 추론 결과 분류 코드는 후속 업무 시스템 설계·업무 규칙 정의·화면 UI 설계에 직접 의존하므로, 'AI 모델을 구현하면서 코드를 정하겠다'는 접근은 전체 사업 일정 지연과 시스템 간 불일치를 초래한다. 감리원은 추론 결과 분류 코드 집합이 확정되도록 유도해야 한다.

2) 예외 코드(미분류·판정불가) 미정의는 운영 단계 오류 처리 불가 사태로 이어진다: AI 모델은 운영 환경에서 기존 코드 집합으로 분류할 수 없는 입력을 반드시 마주치며, 예외 코드(99=미분류, 88=판정불가, 00=해당없음)가 정의되지 않으면 처리 불가 케이스를 임의 코드로 강제 매핑하거나 오류를 발생시킨다. 감리원은 AI 서비스 도메인의 예외 케이스를 파악하고 예외 코드가 공통 코드 정의서에 포함되었는지 확인해야 한다.

3) 코드 의미 미기재는 라벨링 작업자의 주관적 해석을 허용하는 것과 같다: 코드 값 목록만 있고 한글 의미 설명이 없으면 동일 코드에 다른 의미의 라벨이 부여되므로, 감리원은 코드 값별 의미 설명이 '01=정상' 수준의 단순 용어가 아니라 경계 사례에 대한 판단 기준(예: 거래 금액 5만원 미만은 01 코드 부여)까지 기술되었는지 확인해야 한다.

4) 코드 추가 무분별 허용은 AI 모델 재학습 품질 저하의 잠재적 원인: 코드 항목이 검토 없이 추가되면 신규 코드로 쌓인 운영 데이터를 기존 학습 데이터와 혼합 재학습할 때 코드 체계 불일치로 재학습 효과가 저하되므로, 코드 추가 요청 시 AI 모델 재학습 영향 분석 수행 여부를 판단하는 기준(예: 분류 카테고리 수 변경 시 필수 재학습)이 {요구사항정의서}에 포함되었는지 확인해야 한다. 아울러 정의된 공통 코드가 물리 DB 적재 시 Foreign Key 또는 Check Constraint 등 제약조건으로 무결성이 강제되도록 요건에 명시되었는지 점검한다.

5) 기존 업무 시스템 공통 코드와 AI 코드 체계 간 불일치 여부를 현장에서 직접 확인: 발주기관이 기존 공통 코드 관리 시스템을 운영 중이면 신규 AI 시스템 코드 체계가 기존 공통 코드와 연계·호환되는지 또는 별도 코드 체계로 운영되는지 {요구사항정의서}에 명확히 정의되었는지 확인하고, 별도 운영 시 두 코드 체계 간 변환 매핑 테이블과 동기화 관리 요건이 정의되었는지 반드시 점검해야 한다.

AG_SD41-3-03 | 데이터계보 추적요건 점검

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템에 활용되는 데이터의 계보(Lineage) 추적 요건이 {요구사항정의서}에 구체적으로 명시되었는지, 데이터 수집·변환·모델링 파이프라인 요건이 AI 생명주기 전 단계에 걸쳐 추적 가능하도록 정의되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템의 성능·신뢰성·공정성은 데이터의 출처, 변환 과정, 모델 입력까지의 흐름 전체에 의해 결정된다. 요구정의 단계에서 데이터 계보 및 파이프라인 요건을 명확히 정의하지 않으면, 이후 개발·감리 단계에서 데이터 품질 결함의 원인 추적이 불가능해지고, 모델 편향(Bias) 발생 시 책임 소재도 불분명해진다.

▶ 데이터 추적 가능성(Traceability) 확보: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 학습·파인튜닝·RAG에 활용되는 데이터셋의 출처, 수집 방법, 전처리 이력이 추적 가능하도록 요건을 정의할 것을 요구한다.

▶ 파이프라인 복잡성 관리: AI 데이터는 수집 → ETL 변환 → 정제 → 특징 추출(Feature Engineering) → 모델 입력까지 다단계 파이프라인을 거치므로, 각 단계의 요건이 요구사항에 미리 정의되어야 이후 설계·개발 감리에서 이행 여부를 점검할 수 있다.

▶ 편향 감지 및 공정성 근거 확보: NIST AI RMF MAP 2.3은 데이터 수집·선택 과정의 과학적 무결성과 대표성 검증을 문서화할 것을 요구하며, 계보 추적 요건은 이러한 감사 근거 자료의 기반이 된다.

▶ 감리 실효성 확보: 데이터 계보 요건이 요구사항 단계에서 수치화·구체화되지 않으면, 개발 이후 감리에서 설계 적정성 판단이 불가능하며 '적합'·'부적합' 판정 기준 자체가 모호해진다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 계보 추적 요건 명시	데이터 출처(Source), 수집 방법, 이력 기록 기준, 버전 관리 방식이 {요구사항정의서}에 구체적으로 명시되었는지 확인

No	점검 항목	확인 방법
2	파이프라인 단계별 요건 정의	수집 → ETL 변환 → 정제 → 특징 추출 → 모델 입력 각 단계의 처리 기준, 책임자, 품질 임계값이 요건으로 정의되었는지 확인
3	데이터 추적 범위 및 granularity 정의	추적 단위(레코드/배치/버전 단위)와 메타데이터 항목이 명시되었는지, RAG 지식 베이스 포함 여부가 명확히 정의되었는지 확인
4	파이프라인 오류·변경 발생 시 추적 요건	파이프라인 중 이상 발생 시 롤백(Rollback) 기준, 재처리 이력 기록, 경보 요건이 사전 정의되었는지 확인
5	임베딩 데이터 계보 추적 요건 (벡터 DB)	RAG 사용 시 원문서→청크→임베딩→벡터 인덱스 변환의 각 단계 식별자(doc_id·chunk_id·embedding_version)와 사용 임베딩 모델·파라미터가 추적 가능하게 요건으로 정의되었는지 확인
6	RFP 대비 요건 완전성 추적	RFP에 명시된 데이터 계보·파이프라인 요건이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}를 통해 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	데이터 계보 추적 요건, 파이프라인 단계별 처리·책임 요건, 추적 범위 및 granularity, 오류·변경 발생 시 추적·재처리 요건, 임베딩 데이터 계보 추적 요건이 {요구사항정의서}에 구체적·수치적으로 명시되어 있고, RFP 대비 누락이 없는 경우	현행 유지
보완권고 (△)	요건이 부분적으로 명시되었으나, 추적 범위 및 granularity, 파이프라인 단계별 책임자, 오류·변경 시 재처리 이력 요건, 임베딩 데이터 계보 추적 요건 중 일부가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	데이터 계보·파이프라인 요건이 {요구사항정의서}에 전혀 정의되지 않음	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요건의 수치화 유도: "데이터는 충분히 수집한다"는 모호한 표현은 부적합 의견의 근거가 된다. 감리원은 "수집 데이터의 출처 시스템 수, 이력 기록 보존 기간(예: 최소 3년), 계보 추적 항목 수(예: 5종 이상)"처럼 측정 가능한 수치를 요구사항에 반영하도록 사업자를 유도해야 한다.

2) 파이프라인 요건과 DB 설계 요건의 구분: 파이프라인 요건(수집·변환·적재 흐름)은 DB 스키마 설계 요건과 다르다. 감리원은 '{요구사항정의서}'에서 두 요건이 혼재되거나 파이프라인 요건이 누락된 경우를 별도로 식별하여 보완을 권고해야 한다.

3) RAG 구조 확인 시 별도 점검: RAG(Retrieval-Augmented Generation) 방식의 생성형 AI 서비스는 외부 지식 베이스가 지속 갱신되므로, 지식 베이스 갱신 이력과 버전 관리 요건이 별도 항목으로 {요구사항정의서}에 정의되었는지 반드시 확인해야 한다.

4) 특징 저장소(Feature Store) 미도입 시 대안 요건 확인: Feature Store를 도입하지 않는 경우에도 특징 데이터의 버전 관리·재현 가능성 요건이 대안적 방법으로 정의되어 있어야 한다. 단순 '로컬 파일 저장' 서술만 있는 경우 보완을 권고한다.

5) 인터뷰 활용 시 주의사항: 사업자 담당자가 "파이프라인은 개발팀이 알아서 한다"고 답변하는 경우, 이는 요구사항 정의 미흡으로 판단할 수 있다. 감리원은 인터뷰 결과를 감리 보고서에 근거로 기록해야 한다.

AG_SD41-3-04 | 메타데이터 확보 관리 요건

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP · ISO 19115 · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용되는 데이터에 대한 메타데이터 확보·관리 요건과 학습데이터와 메타데이터의 분리 명세 요건이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템에서 메타데이터는 학습 데이터셋의 출처, 품질, 수집 조건, 라벨링 기준 등을 설명하는 '데이터에 관한 데이터'로서 모델 성능의 재현 가능성과 설명 가능성의 핵심 기반이며, 요구정의 단계에서 메타데이터 요건이 학습데이터 요건과 혼재되거나 누락되면 이후 개발·검사 단계에서 데이터 품질 감사와 AI 모델 책임성 확보가 곤란해진다.

▶ 모델 재현 가능성 및 감사 가능성 확보: 동일 조건으로 학습을 재현하거나 문제 발생 시 원인 데이터를 소급 추적하기 위해서는 학습 데이터셋의 메타데이터가 별도로 관리되어야 하며, 이를 위한 요건이 요구정의 단계에서 먼저 정의되어야 한다.

▶ 학습데이터-메타데이터 혼재로 인한 품질 오염 방지: 학습데이터와 메타데이터가 동일 테이블·파일에 혼재될 경우 전처리 과정에서 메타데이터가 학습 입력값으로 오입력되어 모델 편향(Bias)이 발생하는 구조적 위험이 있으므로, 분리 명세 요건을 요구사항으로 정의하여 사전 차단해야 한다.

▶ 신뢰할 수 있는 AI 구현 요건 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 학습데이터의 수집 출처, 품질 기준, 레이블 정보 등을 메타데이터로 문서화할 것을 요구하며, NIST AI RMF MAP 2.3은 데이터 수집·선택의 대표성 및 적합성 검증 문서화를 요건으로 명시하고 있다.

▶ 감리 점검 기준 확보: 메타데이터 항목과 학습데이터 항목이 {요구사항정의서}에서 명확히 분리·정의되지 않으면 이후 DB 설계 감리에서 설계 적정성 판단 기준이 부재하게 된다.

3. 검토방법

No	점검 항목	확인 방법
1	메타데이터 항목 및 확보 요건 명시	수집 출처, 수집 일시, 데이터 형식, 레이블 기준, 품질 등급 등 메타데이터 필수 항목이 {요구사항정의서}에 열거·정의되었는지 확인

No	점검 항목	확인 방법
2	학습데이터-메타데이터 분리 명세 요건 정의	학습데이터(입력값·레이블)와 메타데이터(설명 정보)가 요구사항에서 구조적으로 분리 정의되었는지, 물리적 분리(별도 테이블·파일) 방안이 요건으로 명시되었는지 확인
3	메타데이터 표준 및 형식 정의	메타데이터 표현 형식(JSON-LD, Dublin Core, ISO 19115 등), 필드명 명명 규칙, 데이터 타입이 요건에 정의되었는지 확인
4	메타데이터 갱신·버전 관리 요건 정의	학습데이터 추가·변경 시 메타데이터의 자동 갱신 또는 수동 갱신 기준, 버전 식별자 부여 방식, 이력 보존 기간이 요건으로 명시되었는지 확인
5	벡터 데이터 메타데이터 확보 요건	각 벡터별 원본 매핑(doc_id·chunk_id), 임베딩 모델명·버전·차원·거리척도·생성일시 등 벡터 DB 메타데이터 확보·관리 요건이 {요구사항정의서}에 정의되었는지 확인
6	RFP 대비 메타데이터 요건 완전성 추적	RFP에 명시된 메타데이터 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}를 통해 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	메타데이터 필수 항목 및 확보 요건, 학습데이터-메타데이터 분리 명세, 메타데이터 표준·형식, 갱신·버전 관리 요건, 벡터 데이터 메타데이터 확보 요건이 {요구사항정의서}에 구체적으로 명시되어 있고, RFP 대비 누락이 없는 경우	현행 유지
보완권고 (△)	메타데이터 항목은 정의되었으나, 학습데이터 분리 명세, 표준·형식, 갱신·버전 관리, 벡터 데이터 메타데이터 요건 중 일부가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	메타데이터 요건이 {요구사항정의서}에 전혀 정의되지 않거나, 학습데이터와 혼재 기술됨	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 메타데이터 항목의 수치화 유도: "메타데이터를 관리한다"는 서술만으로는 부적합 판정이 가능하므로, "메타데이터 필수 항목 최소 7종 이상 정의(출처·수집일시·형식·언어·레이블기준·품질등급·라이선스)", "버전 식별자 부여 기준 명시" 등 측정 가능한 요건으로 구체화하도록 사업자를 유도해야 한다.

2) 학습데이터 요건과 메타데이터 요건의 혼재 여부 주의: {요구사항정의서}에서 두 항목이 같은 섹션에 혼재 기술된 경우 '분리 명세 미흡'으로 판단하고 별도 섹션 또는 별도 명세서로 분리 작

성하도록 보완 권고해야 하며, 단순 관리 선언을 넘어 TTA 가이드라인 등에 명시된 필수 메타데이터 항목(데이터셋 버전, 품질 등급, 수집 일자, 라이선스 등)이 구체적인 속성(Attribute) 레벨로 도출되도록 가이드해야 한다.

3) 데이터셋 카드와 요구사항정의서의 관계 명확화: 사업자가 'AI 학습데이터 구축 계획서'나 '데이터셋 카드'를 별도 제출하더라도 해당 내용이 {요구사항정의서}에 요건으로 명시되지 않았다면 감리 기준상 요건 정의로 인정되지 않으므로, 별도 산출물 내용의 {요구사항정의서} 반영 여부를 반드시 확인해야 한다.

4) RAG 지식 베이스 메타데이터 별도 점검: RAG 방식의 생성형 AI 서비스는 외부 지식 문서가 학습데이터와 다른 성격을 가지므로, 지식 문서에 대한 메타데이터(출처 URL, 갱신 일시, 문서 버전, 신뢰도 등급) 요건이 학습데이터 메타데이터 요건과 별도로 정의되어 있는지 반드시 확인해야 한다.

5) 인터뷰 시 분리 구현 계획 구체적 질의: 사업자가 '메타데이터도 같이 관리할 계획'이라고 답변하는 경우 분리 명세 미이행 위험 신호이므로, '메타데이터 전용 테이블 또는 데이터 카탈로그를 별도로 구축할 계획인지', '연계 키(ID) 설계는 어떻게 할 것인지'를 구체적으로 질의하고 답변 내용을 감리 보고서의 인터뷰 근거로 기록해야 한다.

AG_SD41-3-05 | AI 의사결정 감사 추적(Audit Trail) 요건 점검

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | EU AI Act · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 AI 시스템의 신뢰성과 책임성을 제도적으로 뒷받침하기 위해 마련되었다.

구체적으로는 AI 모델의 추론 과정, 자동화된 의사결정 결과, 모델 입력·출력·신뢰도 점수, 예외처리 이력 등 AI 운영 감사 추적에 필요한 데이터 항목과 로그 저장 요건이 {요구사항정의서} 및 {데이터 요구사항}에 사전에 정의되었는지를 감리원이 점검하는 것을 목적으로 한다.

AI 시스템이 행정 결정이나 공공 서비스에 활용될 경우, 해당 시스템의 판단 근거를 사후에 재현하고 설명할 수 있어야 한다.

이를 위해서는 감사 추적 데이터의 항목, 보존 기간, 저장 구조, 검색 성능이 요구정의 단계부터 체계적으로 명시되어야 하며, 본 가이드라인은 그 이행 여부를 감리원이 점검하는 기준을 제시한다.

2. 필요성

▶ AI 의사결정의 감사 추적(Audit Trail) 요건이 명확히 정의되어야 하는 이유는 다음과 같이 다층적으로 존재한다.

▶ 현행 DB 영역 점검 가이드라인은 주로 설계·종료 단계에서의 DB 구조 및 성능을 다루고 있어 AI 의사결정 추적용 데이터 정의는 누락되기 쉽다. 요구정의 단계에서 감사 항목이 누락되면, 설계·개발 완료 후 이를 보완하기 위해 시스템 전반의 재설계가 불가피해질 수 있으며, 이는 사업 일정과 예산에 상당한 부담을 초래한다.

▶ AI 자동 결정에 대한 행정심판 또는 소송이 발생할 경우 의사결정의 근거 자료 보존은 법적 의무이자 기관 방어의 핵심 수단이 된다. 특히 복지급여 결정, 허가·인허가, 행정처분 등 고영향 AI 결정의 경우, 감사 추적 자료의 부재는 기관의 법적 책임 리스크를 직접적으로 높인다.

▶ EU AI Act 제12조는 고위험 AI 시스템에 대한 자동 로그 기록 의무를 명문화하고 있으며, 국내 AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법)(고영향 AI 조항) 역시 이에 준하는 의무를 규정하고 있다. 공공 AI 서비스를 구축하는 기관은 이러한 국내외 법령 요건을 요구사항에 반드시 반영해야 한다.

▶ DB 설계 단계에 진입하기 전에 '무엇을 저장할지'를 요구정의 단계에서 명확히 확정해야 한다. 데이터 항목, 보존 기간, 저장 구조가 미리 정해지지 않으면 설계 단계에서 임의로 결정되거나, 운영 중 필요에 따라 파편적으로 추가되어 데이터 일관성과 추적 완전성이 훼손될 위험이 있다.

3. 검토방법

No	점검 항목	확인 방법
1	감사 추적 데이터 항목 정의 완전성	{요구사항정의서}와 {데이터 요구사항}에서 모델 입력·출력, 신뢰도 점수, 모델·프롬프트 버전, 예외 처리 이력이 저장 항목으로 열거되었는지 항목 단위로 확인
2	보존 기간 법령 매핑	감사 로그 보존 기간이 업무 관련 법령·기록물 관리 규정과 매핑되어 근거 조항과 함께 기재되었는지 확인하고, 만료 후 파기 절차의 적법성을 법무 담당자 면담으로 확인
3	DB 구조 사전 설계 적정성	감사 추적 테이블의 분리 여부와 인덱스·파티셔닝 전략이 {요구사항정의서} 및 {데이터 요구사항} 수준에서 정의되어 임의 결정 여지가 없는지 확인
4	조회 성능 SLA 정의 여부	수년치 누적을 전제로 한 감사 로그 조회 응답시간 목표(SLA)와 조회 조건(기간·건수)이 요구사항에 수치로 정의되었는지 확인
5	RAG 검색 출처·근거 감사 추적 요건	RAG 기반 서비스인 경우 최종 응답 외에 인용 문서·청크 식별자 등 근거 데이터가 저장 항목으로 {데이터 요구사항}에 포함되었는지 확인
6	권한 분리·접근 통제 정의	감사 로그의 열람·수정·삭제 권한이 운영자 권한과 분리되어 정의되었는지, 위·변조 방지 요건이 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	감사 추적 데이터 항목(모델 입력·출력·신뢰도·예외 이력), 보존 기간의 법령 매핑, 저장 구조(테이블 분리·인덱스·파티셔닝), 조회 성능 SLA, RAG 출처 추적, 권한 분리·접근 통제 요건이 {요구사항정의서} 및 {데이터 요구사항}에 구체적으로 정의된 경우	현행 유지
보완권고 (△)	감사 추적 항목은 정의되었으나 보존 기간의 법령 근거, 조회 성능 SLA, 인덱스·파티셔닝 전략, RAG 출처 추적 등 일부 요건이 미흡하거나 선언적 수준에 그친 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 의사결정 감사 추적 요건이 요구사항에 정의되지 않았거나, 로그 데이터 항목·보존·저장 구조 요건이 전혀 명시되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 감사 추적은 요구정의에서 누락되면 사후 보강이 사실상 불가능한 영역이다: 테이블 하나를 추가하는 문제가 아니라 데이터 항목·저장 구조·보존 정책·접근 권한 체계가 시스템 전반에 걸쳐 설계되어야 하기 때문이다. 감리원은 이 네 가지가 함께 확정되었는지 확인한다.

2) '무엇을 저장할지'를 항목 단위로 확인한다: 모델 입력·출력만 기록하고 신뢰도 점수, 모델·프롬프트 버전, 예외 처리 이력이 빠지면 사후 재현과 책임 규명이 불가능하다. 감리원은 감사 추적 데이터 항목이 요구사항에 열거되어 있는지 확인한다.

3) 보존 기간은 반드시 법령 근거와 매핑되어야 한다: '3년 보관'처럼 근거 없는 기간은 행정심판·소송 대응 시 방어력이 없다. 감리원은 보존 기간이 업무 관련 법령·기록물 관리 규정과 매핑되었는지, 만료 후 파기 절차가 함께 정의되었는지 확인한다.

4) RAG 기반 서비스는 검색 출처·근거의 추적 요건을 별도로 확인한다: 최종 응답만 저장하고 인용된 문서·청크 식별자를 저장하지 않으면 "왜 그 답이 나왔는가"를 사후에 설명할 수 없다. 감리원은 RAG 출처 추적 항목이 요구사항에 포함되었는지 확인한다.

5) 조회 성능 SLA와 권한 분리를 요구정의에서 확정한다: 감사 로그는 수년치가 누적되어 인덱스·파티셔닝 없이는 조회가 불가능해진다. 감리원은 조회 성능 목표(SLA)가 정의되었는지, 로그 열람·삭제 권한이 운영자와 분리되어 정의되었는지 확인한다.

AG_SD41-3-06 | 민감정보 및 개인 식별번호 비식별화 마스킹 및 가명-익명처리요건

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」 · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF GOVERN

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 사용되는 데이터에 포함된 개인정보(PII, Personally Identifiable Information)를 가명처리·익명처리·마스킹 등의 비식별화 방법으로 처리하는 요건이 {요구사항정의서}에 법적 근거와 함께 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 모델은 학습 데이터를 반복적으로 처리하고 모델 가중치에 데이터 패턴을 내재화하는 특성 때문에, 학습 데이터에 포함된 개인정보(이름·주민등록번호·전화번호·이메일·의료정보 등)가 모델 출력을 통해 재현되거나 역추론될 위험이 있다. 개인정보 보호법과「개인정보 보호법」제28조의2(가명정보의 처리)에 따라 학습 데이터 구축 전 비식별화 처리가 법적 의무로 요구된다.

▶ 개인정보 침해로 인한 법적 제재 위험 차단: 개인정보 보호법 위반 시 과태료(최대 3,000만원)·과징금(위반행위 관련 매출액의 3% 이하)·형사처벌이 부과된다. 학습 데이터 구축 전 비식별화 처리 요건이 {요구사항정의서}에 명시되어야 법적 의무 이행의 설계 단계 기반이 마련된다.

▶ AI 모델의 개인정보 재현·역추론 위험 예방: 생성형 AI 모델은 학습 데이터의 특정 패턴을 암기(Memorization)하여 프롬프트 입력 시 실제 개인정보(실명·전화번호 등)를 그대로 출력하는 '학습 데이터 추출 공격(Training Data Extraction Attack)' 위험에 노출될 수 있다. 학습 전 완전한 비식별화 처리가 선행되어야 이 위험을 예방할 수 있다.

▶ TTA 신뢰 AI 및 NIST AI RMF 프라이버시 요건 충족: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 학습 데이터의 개인정보·민감정보 비식별화를 신뢰성 확보 필수 요건으로 명시하며, NIST AI RMF GOVERN 6.2는 AI 시스템의 프라이버시 위험을 사전 식별하고 설계 단계부터 통제 방안을 마련할 것을 요구한다.

▶ 감리원의 개인정보 처리 적정성 독립 검증 기반 확보: 비식별화 대상 항목·처리 방법·검증 절차가 {요구사항정의서}에 명시되어야 감리원이 이후 데이터 실제 비식별화 처리가 요건대로 수행되었는지 독립적으로 검증할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	PII 식별 대상 항목 목록 및 처리 법적 근거 정의	학습 데이터에 포함된 PII 항목(이름·주민등록번호·전화번호·이메일·주소·의료정보·금융정보 등)을 식별한 목록과 각 항목의 처리 법적 근거(개인정보 보호법·의료법·신용정보법 등)가 {요구사항정의서}에 정의되었는지 확인

No	점검 항목	확인 방법
2	가명처리 방법 및 적용 기준 정의	개인정보 보호법 제28조의2에 따른 가명처리 방법(일방향 암호화·양방향 암호화·토큰화·데이터 범주화 등)과 처리 대상 항목, 처리 후 재식별 방지를 위한 추가 정보 분리 보관 요건이 정의되었는지 확인
3	익명처리 방법 및 완전성 검증 요건 정의	개인 재식별이 원천적으로 불가능한 익명처리 방법(데이터 집계·일반화·노이즈 추가·데이터 스와핑 등)과, 처리 후 재식별 가능성이 없음을 검증하는 방법 및 기준이 {요구사항정의서}에 정의되었는지 확인
4	마스킹 방법·적용 범위 및 비식별화 후 검증 절차 정의	특정 항목을 부분 또는 전체 마스킹하는 방법(문자 대체·해시처리·널 처리 등)·적용 범위(전체 마스킹·부분 마스킹 기준)·마스킹 후 재식별 가능성 검증 절차가 비기능 요건으로 정의되었는지 확인
5	임베딩 내 PII 비식별화 요건	원문 비식별화가 임베딩 생성 이전에 수행되는지, 임베딩 역추적(Inversion)·멤버십 추론에 의한 PII 노출 위험 평가와 벡터 저장소 접근통제·동일 마스킹 적용 요건이 정의되었는지 확인
6	비식별화 처리 이력 기록 요건 정의	비식별화 처리 일시·대상 항목·적용 방법(가명/익명/마스킹)·수행자·검증 결과가 이력으로 기록되도록 {요구사항정의서}에 요건이 정의되었는지 확인
7	RFP 대비 PII 비식별화 요건 완전성 추적	RFP에 명시된 개인정보 비식별화·가명처리·익명처리·마스킹 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	PII 식별 대상 항목 목록·법적 근거, 가명처리 방법·추가 정보 분리 보관, 익명처리 방법·재식별 검증 기준, 마스킹 방법·항목별 적용 기준, 임베딩 생성 이전의 원문 비식별화와 임베딩 역추적·멤버십 추론 위험 평가 및 벡터 저장소 접근통제 요건, 비식별화 처리 이력 기록 요건이 {요구사항정의서}에 모두 정의됨	현행 유지
보완권고 (△)	비식별화 방법은 정의되었으나, PII 항목 목록 미정의 또는 가명처리 추가 정보 분리 보관 요건·재식별 검증 기준 수치·처리 이력 기록 요건 중 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	PII 비식별화/마스킹 요건 및 가명/익명 처리 비기능 요건이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '개인정보 보호법 준수'는 요건이 아니고 의무 선언일 뿐: '개인정보 보호법을 준수하여 데이터를 처리한다'는 서술은 비기능 요건이 아니라 법적 의무 준수 선언에 불과하다. 감리원은 어떤 PII 항목을, 어떤 방법(가명처리·익명처리·마스킹)으로, 어떤 기준(k-익명성 k값 등)으로 처리하는지가 {요구사항정의서}에 항목별로 명시되도록 유도해야 한다.

2) 가명처리와 익명처리의 법적 차이를 명확히 구분하여 점검: 가명처리된 데이터는 개인정보 보호법 적용 대상으로 추가 정보 분리 보관 의무가 있으며, 익명처리된 데이터만 법 적용에서 제외된다. 사업자가 '비식별화를 했다'고 주장할 때 그 방법이 가명처리인지 익명처리인지를 구분하고, 가명처리라면 추가 정보 분리 보관 요건이 {요구사항정의서}에 별도로 정의되었는지 확인해야 한다.

3) 간접 식별자 조합에 의한 재식별 위험 별도 점검: 이름·주민등록번호 등 직접 식별자만 제거하고 간접 식별자(나이·직업·거주 지역·직장명 등)를 그대로 남겨두면, 간접 식별자 조합으로 재식별이 가능하다. 감리원은 직접 식별자 처리 요건뿐만 아니라 간접 식별자 조합을 통한 재식별 가능성 검증 요건(k-익명성 등)이 {요구사항정의서}에 정의되었는지 확인해야 한다.

4) 생성형 AI 학습 데이터의 텍스트 내 PII 탐지 자동화 요건 확인: 자유 텍스트(고객 상담 이력·SNS 게시물·의료 기록 등)에서 PII를 탐지하는 것은 수작업으로 불가능하며, NER(Named Entity Recognition) 기반 PII 탐지 도구 등 자동화 도구 활용이 필요하다. 텍스트 데이터가 포함된 경우 자동화 PII 탐지 도구 적용 요건과 탐지 결과 검증 절차가 정의되었는지 확인해야 한다.

5) 비식별화 처리 이력 산출물은 감사·분쟁 대비 핵심 증거: 비식별화 처리 이력(처리 대상 항목·방법·일시·담당자·검증 결과)이 산출물로 기록되지 않으면 개인정보 보호법 위반 시 조치 이행 입증이 불가능하다. '비식별화 처리 결과 보고서'를 공식 산출물로 정의하고 감리원이 열람할 수 있도록 제출 요건을 {요구사항정의서}에 포함하도록 반드시 유도해야 한다.

AG_SD41-3-07 | 기존데이터 마이그레이션 및 레거시 데이터 이관 검증 요건

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 정제(클렌징)/변환

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」 · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 기존 레거시 시스템에 축적된 데이터를 신규 AI 시스템으로 이관할 때 필요한 데이터 전환·마이그레이션 범위, 변환 규칙, 이관 검증 기준이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 모델 학습에 활용되는 레거시 데이터의 품질·완전성·정합성 확보 요건이 명확히 정의되었는지 확인한다.

2. 필요성

▶ 레거시 시스템의 기존 데이터는 AI 모델 학습의 핵심 자원이자 서비스 연속성을 위한 필수 자산이나, 코드 체계 불일치·데이터 형식 차이·누락 값·중복 레코드·문자 인코딩 오류 등 다양한 품질 문제를 내포하고 있어 이관 과정에서 데이터 손실·변조·정합성 오류가 발생할 위험이 크다.

▶ 이관 누락·데이터 손실에 의한 AI 모델 학습 품질 저하 방지: 이관 대상 범위가 명확히 정의되지 않으면 AI 모델 학습에 필요한 역사적 데이터 일부가 누락되어 모델이 특정 기간·구간의 패턴을 학습하지 못하는 편향이 발생할 수 있으므로, 이관 대상 테이블·컬럼·기간 범위를 요구정의 단계에서 명확히 정의해야 한다.

▶ 데이터 변환 오류에 의한 AI 학습 데이터 정합성 훼손 방지: 레거시와 신규 시스템 간 코드 체계·데이터 타입·단위·문자 인코딩이 다를 경우 변환 규칙 없이 이관하면 데이터가 잘못 변환되거나 손상되어 AI 모델이 오염된 데이터를 학습하게 되므로, 항목별 변환 매핑 규칙이 요건으로 정의되어야 한다.

▶ 이관 완료 여부 검증 기준 부재에 의한 사후 분쟁 위험 예방: 이관 건수·합계·체크섬 등 정량적 검증 기준이 사전에 정의되지 않으면 이관 완료 후 발주기관과 사업자 간 이관 완전성에 대한 분쟁이 발생할 수 있으므로, 검수 기준이 {요구사항정의서}에 수치로 명시되어야 한다.

▶ 감리원의 데이터 이관 적정성 독립 검증 기반 확보: 이관 범위·변환 규칙·검증 기준이 {요구사항정의서}에 명시되어야 감리원이 이후 실제 이관 결과가 요건대로 수행되었는지 독립적으로 검증할 수 있는 기준이 마련된다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 이관 대상 범위 정의 (테이블·컬럼·기간)	이관 대상 레거시 DB의 테이블 목록, 각 테이블의 이관 대상 컬럼 목록, 이관 대상 데이터 기간(시작~종료 일자), 이관 제외 대상(임시 데이터·테스트 데이터·삭제된 데이터 등)이 {요구사항정의서}에 구체적으로 정의되었는지 확인
2	데이터 항목별 변환 규칙 매핑 정의	레거시 시스템과 신규 시스템 간 코드 체계·데이터 타입·단위·문자 인코딩·날짜 형식 차이에 대한 항목별 변환 규칙(코드 매핑 테이블·타입 변환 기준·NULL 처리 규칙 등)이 '데이터 전환 계획서' 등 관련 산출물에 정의되었는지 확인
3	이관 완전성·정합성 검증 방법 및 기준 수치 정의	이관 완료 후 레거시-신규 DB 간 건수 일치 검증·합계 검증·샘플 데이터 비교 검증·참조 무결성 검증 방법과 이관 완료 판정 기준 수치(예: 건수 불일치율 0%(완전 일치), 샘플 검증 일치율 기준 등)가 정의되었는지 확인
4	이관 실패·오류 발생 시 롤백 절차 및 재이관 기준 정의	이관 중 오류·실패 발생 시 레거시 시스템으로 복구하는 롤백 절차, 오류 허용 임계값(이 값 초과 시 이관 중단), 재이관 조건 및 승인 절차가 '데이터 전환 계획서'에 정의되었는지 확인
5	재임베딩·벡터 재구축 마이그레이션 검증 요건	레거시 이관에 더해 임베딩 모델·차원 변경 시 전체 재임베딩·벡터 인덱스 재구축 범위와 구·신 버전 병행 및 이관 정합성 검증 기준이 정의되었는지 확인
6	RFP 대비 데이터 이관 요건 완전성 추적	RFP에 명시된 레거시 데이터 이관 범위·변환 규칙·검증 기준·롤백 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	이관 대상 테이블·컬럼·기간 범위, 항목별 코드 매핑·타입·인코딩 변환 규칙, 건수·합계·샘플 검증 방법 및 판정 기준 수치, 롤백 절차·오류 허용 임계값, 임베딩 모델·차원 변경 시 재임베딩 및 벡터 인덱스 재구축 범위와 구·신 버전 병행·이관 정합성 검증 기준이 {요구사항정의서} 및 관련 산출물에 모두 정의됨	현행 유지
보완권고 (△)	이관 범위는 정의되었으나 코드 매핑 규칙 일부 누락, 검증 기준 수치 미정의, 롤백 절차 미정의 등 일부 요건 미흡	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	데이터 이관 범위·변환 규칙·검증 기준·롤백 계획이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '전체 데이터를 이관한다'는 요건은 가장 위험한 선언: '모든 기존 데이터를 신규 시스템으로 이관한다'는 서술은 이관 대상이 전혀 정의되지 않은 상태로 이관 완료 여부를 둘러싼 분쟁의 씨앗이 되므로, 감리원은 이관 대상 테이블·컬럼 목록, 이관 기간, 이관 제외 대상 및 제외 사유가 {요구사항정의서}에 구체적으로 명시되도록 반드시 유도해야 한다.

2) 코드 체계 불일치는 AI 학습 데이터 오염의 주요 원인: 레거시 코드 값(성별 1/2, 상태 코드 01~09 등)이 신규 코드 체계와 다를 때 매핑 규칙 없이 이관하면 AI 모델이 오염된 분류 레이블을 학습하게 되므로, 코드 매핑 테이블의 항목별 작성 여부와 미매핑 코드 값 처리 기준(예외 처리·NULL 처리·기본값 지정)을 반드시 확인하고, "AS-IS vs TO-BE 데이터 매핑 정의서 작성"이 산출물 요구사항으로 명시되어 있는지 점검 기준을 강화해야 한다.

3) 검증 기준 없는 이관 완료 선언은 감리 의견 대상: '이관을 완료하였음'이라는 선언만 있고 건수·합계·샘플 검증 결과가 없으면 이관 완전성을 검증할 수 없으므로, 이관 검증 방법·판정 기준 수치·검증 결과 산출물이 {요구사항정의서}에 사전 정의되도록 유도하고 미정의 시 보완 권고를 명확히 기록해야 한다. 아울러 이관 전 레거시 데이터의 프로파일링(결측치·이상치 비율 등)을 수행하고 AI 학습 허용 기준과 비교하는 요건이 포함되었는지 점검한다.

4) EUC-KR → UTF-8 인코딩 변환 오류는 가장 빈번한 이관 장애: 국내 레거시 시스템은 EUC-KR 인코딩을 사용하는 경우가 많아 UTF-8 환경으로 이관 시 한글이 깨지거나 물음표(?)로 변환되는 오류가 빈번하므로, 레거시 DB 인코딩 현황과 신규 DB 인코딩 기준, 변환 방법 및 변환 후 한글 정합성 검증 요건이 '데이터 전환 계획서'에 명시되었는지 집중 점검해야 한다.

5) 롤백 계획 없는 이관은 원점 복구 불가 위험을 내포: 이관 중 심각한 오류 발생 시 레거시로 되돌아갈 롤백 계획이 없으면 양쪽 시스템 모두 불완전한 상태가 되어 서비스가 중단될 수 있으므로, 이관 오류 허용 임계값, 자동 중단 조건, 롤백 수행 주체·절차·소요 시간, 발주기관 승인 절차가 '데이터 전환 계획서'에 정의되었는지 반드시 확인해야 한다.

AG_SD41-3-08 | 학습-검증 테스트 분할비율 및 데이터셋 분할기준 명시 여부

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 정제(클렌징)/변환

근거 | NIST AI RMF MEASURE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 활용될 전체 데이터셋을 학습(Training)·검증(Validation)·테스트(Test) 세트로 분할하는 비율 기준, 분할 방법, 데이터 누수 방지 요건이 {요구사항정의서}에 구체적으로 반영되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

분할 기준 미정의로 인한 모델 과적합·성능 허위 보고·데이터 누수 위험을 사전에 차단하는 요건 정의 여부를 확인한다.

2. 필요성

▶ AI 모델의 예측 성능은 학습·검증·테스트 세트의 분할 방법과 비율에 의해 직접적으로 영향을 받는다. 분할 기준이 명확히 정의되지 않으면 전체 데이터를 학습에 사용하거나, 테스트 데이터가 학습 과정에 노출되거나, 시계열 데이터를 무작위 분할하는 등의 오류가 발생하여 모델 성능이 실제보다 과대 평가되는 허위 성능 보고로 이어질 수 있다.

▶ 테스트 세트 부재 또는 학습 데이터 혼용에 의한 허위 성능 보고 방지: 독립적인 테스트 세트 없이 학습 데이터로 모델 성능을 측정하거나, 테스트 데이터 일부가 학습 과정에 사용되면(데이터 누수) 실제 서비스 환경에서는 달성 불가한 허위 성능 지표가 산출된다. 이는 사업 완료 후 심각한 분쟁으로 이어질 수 있다.

▶ 시계열·그룹 데이터의 무작위 분할 오류로 인한 데이터 누수 위험 방지: 시간 순서가 있는 시계열 데이터(금융 거래·로그·센서 데이터 등)를 단순 무작위(Random) 분할하면 미래 데이터가 학습 세트에, 과거 데이터가 테스트 세트에 포함되어 실제 운영 환경에서 존재할 수 없는 미래 정보로 학습한 모델이 만들어진다. 시계열 데이터에는 시간 기준 순차 분할 요건이 정의되어야 한다.

▶ 분할 비율 재현 가능성 보장을 위한 랜덤 시드 고정 요건 확보: 동일한 분할 비율을 적용하더라도 랜덤 시드(Random Seed)가 고정되지 않으면 실행마다 다른 분할 결과가 생성되어 실험 재현이 불가능하다. 감리 시 모델 성능의 독립 재현 검증이 불가해지므로 랜덤 시드 고정 요건이 {요구사항정의서}에 포함되어야 한다.

▶ 데이터 규모·도메인별 적정 분할 비율 근거 확보: 분할 비율은 전체 데이터 규모·클래스 분포·도메인 특성에 따라 달라진다. 단순히 '70:20:10 비율로 분할한다'는 선언이 아니라 해당 비율을

선택한 근거(데이터 규모·클래스 불균형 고려 여부 등)가 {요구사항정의서}에 기술되어야 감리원이 적정성을 판단할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	학습/검증/테스트 세트 분할 비율 및 각 세트의 역할 정의	전체 데이터셋의 학습·검증·테스트 세트 분할 비율(예: 70:15:15 또는 80:10:10)과 각 세트의 역할(학습: 모델 학습, 검증: 하이퍼파라미터 튜닝, 테스트: 최종 성능 평가)이 {요구사항정의서}에 정의되었는지 확인
2	데이터 특성에 맞는 분할 방법 정의 (무작위·층화·시계열 순차)	전체 데이터의 분포 특성(클래스 불균형·시계열·동일 그룹 내 데이터 존재 여부)에 따라 적합한 분할 방법(단순 무작위·층화 무작위·시계열 순차·그룹 기반)이 선택되었고, 선택 근거가 기술되었는지 확인
3	테스트 세트 독립성 보장 및 데이터 누수 방지 요건 정의	테스트 세트가 학습·검증 과정에서 완전히 격리되어 모델 튜닝에 사용되지 않는다는 독립성 요건과, 전처리(정규화·스케일링·결측값 처리) 파라미터를 학습 세트에서만 계산하고 검증·테스트 세트에는 적용만 하는 누수 방지 절차가 정의되었는지 확인
4	랜덤 시드 고정 및 분할 결과 재현 가능성 요건 정의	무작위 분할 시 사용하는 랜덤 시드(Random Seed) 값을 고정하여 동일 분할 결과를 재현할 수 있는 요건이 {요구사항정의서} 또는 'AI 모델 설계서'에 정의되었는지 확인
5	RAG 검색 평가셋 분리 요건	RAG 검색 품질 평가셋(질의-정답쌍)이 색인 대상 문서와 분리되어 평가 데이터 누수가 방지되도록 분할·관리 기준이 정의되었는지 확인
6	RFP 대비 데이터셋 분할 기준 요건 완전성 추적	RFP에 명시된 학습/검증/테스트 분할 비율·분할 방법·성능 평가 기준 관련 요구사항이 {요구사항정의서}에 누락 없이 반영되었는지 {과업대비표} 및 {요구사항추적표}로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	분할 비율·각 세트 역할, 데이터 특성에 맞는 분할 방법·선택 근거, 테스트 세트 독립성 요건, 전처리 누수 방지 절차, 랜덤 시드 고정 요건, RAG 활용 시 검색 품질 평가셋(질의-정답쌍)을 색인 대상 문서와 분리하여 평가 데이터 누수를 방지하는 분할·관리 기준이 {요구사항정의서} 및 관련 산출물에 모두 정의됨	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	분할 비율은 정의되었으나 분할 방법 선택 근거 미기술, 데이터 누수 방지 절차 미정의, 랜덤 시드 미고정, 시계열 데이터 무작위 분할 등 일부 요건 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	학습/검증/테스트 분할 비율과 분할 방법이 {요구사항정의서}에 전혀 정의되지 않거나 선언적 서술만 존재	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '전체 데이터로 학습한다'는 요건은 성능 측정 자체가 불가한 설계: 전체 데이터를 학습에 사용하고 학습 데이터로 성능을 측정하는 것은 일반화 성능이 아니라 암기 수준을 측정하는 것으로, 운영 환경 성능과 무관한 허위 지표다. 감리원은 독립적인 테스트 세트 확보 요건이 정의되도록 보완을 권고해야 한다.

2) 시계열 데이터의 무작위 분할은 심각한 데이터 누수의 원인: 금융 거래·로그·센서 데이터 등 시간 순서가 있는 데이터를 단순 무작위 분할하면 미래 정보를 활용한 모델이 만들어진다. 감리원은 AI 서비스 데이터가 시계열 특성을 갖는지 확인하고, 해당 시 시간 기준 순차 분할 요건이 정의되었는지 점검해야 한다.

3) 전처리 파라미터 계산 시점이 데이터 누수의 핵심 점검 포인트: '전체 데이터로 정규화 후 분할' 방식은 테스트 세트의 통계 정보가 학습에 반영되는 전처리 누수를 유발한다. '전처리 파라미터(평균·표준편차·인코딩 맵)를 학습 세트에서만 계산하고 검증/테스트 세트에는 적용(transform)만 한다'는 절차가 정의되었는지 확인해야 한다.

4) 랜덤 시드 미고정은 감리원의 독립 재현 검증을 불가능하게 만든다: 시드가 고정되지 않으면 동일 코드를 실행해도 다른 분할 결과가 나와 성능 수치를 독립적으로 재현 검증할 수 없다. 특히 성능 미달 분쟁 시 결정적 약점이 되므로, 랜덤 시드 고정 요건 명시를 요구해야 한다.

5) 분할 비율 근거 없는 관행적 70:30 적용은 보완 권고 대상: 클래스 불균형이 심각하거나(소수 클래스 1% 미만) 데이터 규모가 매우 작은(1만 건 미만) 경우 관행적 분할은 소수 클래스가 테스트 세트에 수 건 미만으로 배정되어 유의미한 성능 평가가 불가능하다. 감리원은 데이터 특성과 분할 비율의 근거가 일치하는지 확인해야 한다.

AG_SD41-3-09 | AI 학습데이터 품질기준 명시

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 데이터 품질

근거 | 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」·「신뢰할 수 있는 인공지능 개발안내서」(TTA)·NIST AI RMF MAP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 학습에 투입되는 데이터의 완전성(Completeness)·정확성(Accuracy)·정합성(Consistency)·유일성(Uniqueness)·적시성(Timeliness)·유효성(Validity) 등 품질 특성별 측정 기준과 허용 임계값이 {요구사항정의서}에 수치로 정의되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 학습 데이터 품질 기준 미정의로 인해 불량 데이터가 AI 모델에 학습되어 성능 미달·편향·검수 단계 분쟁이 발생하는 상황을 사전에 차단한다.

2. 필요성

▶ AI 모델의 성능과 신뢰성은 학습 데이터의 품질에 직접 의존한다. '좋은 데이터로 학습시키겠다'는 선언만 있고 품질 특성별 측정 기준과 허용 임계값이 {요구사항정의서}에 수치로 정의되지 않으면, 사업자·발주기관·감리원 각각이 '충분한 데이터 품질'을 다르게 해석하여 검수 단계에서 심각한 분쟁이 발생한다.

▶ AI 학습 데이터 품질 기준 미정의로 인한 검수 단계 분쟁 방지: AI 서비스 구축 사업에서 학습 데이터 품질 기준이 {요구사항정의서}에 수치로 정의되지 않으면 사업자는 최소 품질 기준으로 데이터를 구축하고 발주기관은 최고 품질을 기대하는 인식 차이가 발생하여 검수 단계에서 '데이터 품질 미달'과 '사업 요건 충족' 주장이 충돌하는 분쟁이 발생한다.

▶ 결측값·오류값 과다로 인한 AI 모델 학습 성능 미달 방지: 완전성 기준(결측값 허용 비율)이 정의되지 않으면 중요 특성(Feature) 컬럼의 결측값 비율이 30~50%에 달하는 데이터가 AI 학습에 투입되어 모델이 부정확한 패턴을 학습하고 요구 성능 기준을 충족하지 못하는 상황이 발생한다.

▶ 정합성 기준 미정의로 인한 복수 소스 데이터 통합 시 오류 누적 방지: 여러 소스에서 수집한 데이터를 통합할 때 동일 개체에 대한 값이 소스마다 다른 경우(예: 동일 고객의 나이가 A소스에서 35세, B소스에서 38세) 정합성 기준이 없으면 어떤 값을 학습에 투입할지 판단 기준이 없어 오류 데이터가 누적되고 AI 모델이 잘못된 패턴을 학습하게 된다.

▶ AI 데이터 품질 측정 도구 없이 개발자 육안 검사에 의존하는 구조 방지: {요구사항정의서}에 측정 가능한 품질 기준과 측정 방법이 정의되지 않으면 사업자가 '개발자 육안 검사'나 '샘플 기반 경험적 판단'으로 데이터 품질을 관리하게 되어 대용량 AI 학습 데이터의 품질 이상을 체계적으로 탐지하지 못하는 구조가 만들어진다.

3. 검토방법

No	점검 항목	확인 방법
1	완전성·유효성 기준: 결측값 허용 비율·값 범위 및 형식 기준 수치 정의	AI 모델 학습에 투입되는 데이터의 완전성(중요 컬럼별 결측값 허용 비율 %, 예: 필수 특성 컬럼 결측값 2% 이하)과 유효성(숫자 컬럼 값 범위·날짜 형식·코드값 허용 목록) 기준이 {요구사항정의서}에 수치로 정의되었는지 확인
2	정확성·정합성 기준: 오류값 허용 비율·소스 간 불일치 처리 기준 수치 정의	실제값과 다른 오류값의 허용 비율(예: 정확성 오류율 1% 이하)과 복수 소스에서 수집한 데이터 통합 시 동일 개체에 대한 소스 간 값 불일치 처리 기준(우선 소스 지정·최신값 우선 등)이 수치와 규칙으로 정의되었는지 확인
3	유일성·중복 기준: 중복 레코드 허용 비율·중복 판단 기준 키 정의	AI 학습 데이터셋 내 중복 레코드의 허용 비율(예: 중복률 0.5% 이하)과 중복 여부를 판단하는 기준 키(복합 키 포함)·중복 발견 시 처리 방법(최신 레코드 유지·랜덤 선택 등)이 정의되었는지 확인
4	적시성 기준: 시계열·동적 데이터의 수집 시점 및 최신성 기준 수치 정의	시계열 데이터·실시간 데이터·주기적으로 변화하는 데이터를 학습에 활용하는 경우 데이터의 최대 허용 지연 시간(예: 수집 후 24시간 이내 데이터만 학습 투입) 또는 최신성 기준(예: 최근 1년 이내 데이터만 활용)이 수치로 정의되었는지 확인
5	벡터 데이터 품질·검색 정확도 기준 요건	검색 정확도(Recall@k·MRR 등) 목표 수치, 중복 벡터 허용 비율, 임베딩 결손·차원 불일치 검출, EDB 원본-벡터 정합성(1:1 매핑) 검증 기준이 수치로 정의되었는지 확인
6	품질 측정 방법·측정 도구·측정 주기 및 품질 임계값 미달 시 처리 절차 정의	품질 특성별 측정에 사용할 도구(Great Expectations·Deequ·SQL 쿼리·커스텀 스크립트 등)·측정 주기(배치 수집 후·학습 투입 전·일별)·측정 결과 이력 기록 방법과 품질 임계값 미달 시 처리 절차(제외·보정·재수집)가 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	완전성·정확성·정합성·유일성·적시성·유효성 각 특성별 측정 기준 수치·허용 임계값, 벡터 데이터 품질·검색 정확도 기준(Recall@k·MRR 목표 수치, 중복 벡터 허용 비율, 임베딩 결손·차원 불일치 검출, 원본-벡터 1:1 정합성), 측정 도구·측정 주기, 품질 임계값 미달 시 처리 절차(제외·보정·재수집), 측정 결과 이력 관리 방법이 {요구사항정의서}에 모두 수치 또는 규칙으로 정의됨	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	일부 품질 특성의 기준은 수치로 정의되었으나 나머지 특성 기준 누락, 측정 방법·주기 미정의, 임계값 미달 처리 절차 누락, 이력 관리 계획 미수립 등 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 학습 데이터 품질 기준이 {요구사항정의서}에 전혀 없거나 '고품질 데이터를 확보하겠다', '데이터 품질을 관리하겠다'는 선언만 존재하고 품질 특성별 측정 기준 수치 전무	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 품질 기준은 반드시 수치(%)로 정의되어야 한다 - '최소화', '충분히', '적절히'는 무효: '결측값을 최소화하겠다', '오류 데이터를 충분히 제거하겠다'는 서술은 측정 기준이 아니라 의도의 표명이다. 감리원은 {요구사항정의서}의 품질 기준에 '결측값 허용 비율 2% 이하', '오류율 1% 이하', '중복률 0.5% 이하'처럼 측정 가능한 수치(%)로 정의되어 있는지 확인해야 한다.

2) AI 모델 유형에 따라 가장 중요한 품질 특성이 다르다: 감리원은 6가지 품질 특성 모두를 동일한 강도로 점검하기보다 해당 AI 모델 유형에서 가장 영향력 있는 품질 특성에 집중해야 한다. 분류 모델은 레이블 정확성·클래스 불균형이 핵심, 시계열 예측 모델은 적시성·결측값 연속성이 핵심, 추천 시스템은 유일성·정합성이 핵심, 이미지 인식 모델은 유효성(해상도·형식 기준)이 핵심이다.

3) 복수 소스 통합 데이터는 정합성 기준이 정의되지 않으면 오류 누적 구조: 여러 시스템·소스에서 데이터를 수집·통합하는 AI 사업에서 정합성 기준이 없으면 동일 고객·동일 제품·동일 이벤트에 대해 소스마다 다른 값이 존재하는 상태로 학습 데이터가 구성된다. 감리원은 복수 소스 통합 구조가 있는 경우 소스 우선순위 정의·불일치 허용 비율·불일치 처리 방법이 정의되었는지 확인해야 한다.

4) 레이블(Label) 품질 기준은 특성 품질 기준과 별도로 확인: 지도학습 AI 모델에서 학습 데이터의 '특성(Feature) 값' 품질과 '정답 레이블(Label)' 품질은 별개로 관리되어야 한다. 특성 품질 기준만 정의하고 레이블 정확성 기준(라벨 오류율 허용 비율·라벨러 간 일치율 기준)이 없으면 특성 데이터는 완벽해도 잘못된 레이블로 학습된 AI 모델이 산출될 수 있으므로 레이블 품질 기준을 별도로 점검해야 한다.

5) 임계값 미달 처리 절차 없는 품질 기준은 집행 불가능한 기준: 품질 기준 수치가 정의되어 있어도 '기준에 미달하면 어떻게 처리하는가'에 대한 절차가 없으면 실제 품질 관리 시 사업자가 미달 데이터를 그대로 학습에 투입하거나 임의로 처리하는 상황이 발생한다. 감리원은 {요구사항정의서}에 품질 기준 수치와 함께 임계값 미달 데이터의 처리 절차(제외·보정·재수집)가 함께 정의되었는지 확인해야 한다.

AG_SD41-3-10 | 벡터 DB 및 RAG 파이프라인 요구정의 적정성 점검

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 벡터 DB/임베딩

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 서비스의 핵심 인프라인 벡터 데이터베이스(Vector DB) 및 RAG(Retrieval-Augmented Generation) 파이프라인에 관한 요구사항이 {요구사항정의서}에 충분히 명시되었는지를 감리원이 점검하는 것을 목적으로 한다.

RAG 파이프라인은 LLM의 환각(Hallucination) 문제를 완화하고 도메인 특화 지식을 검색·활용하는 생성형 AI 서비스의 표준 아키텍처이며, 그 핵심 기반인 벡터 DB는 일단 도입 후 변경이 매우 어려운 인프라이므로 요구정의 단계에서 제품 선정 기준, RAG 전략, 데이터 규모 산정, 운영 SLA, 데이터 거버넌스 요건이 구체적으로 정의되어야 후속 설계·구현 단계에서 임의 결정으로 인한 리스크를 예방할 수 있다.

본 가이드라인은 ①벡터 DB 제품 선정 기준의 정량적 명시 여부, ②RAG 데이터 규모 산정의 적정성, ③운영 SLA 임계값 정의 여부, ④데이터 거버넌스 요건 반영 여부를 중심으로 점검한다.

2. 필요성

▶ 생성형 AI 서비스를 구성하는 기술 요소 가운데 벡터 DB와 RAG 파이프라인은 성능·비용·운영 부담의 3대 축에서 사업 성패를 좌우하는 핵심 결정 사항임에도, 현장에서는 요구정의 단계에서 이를 구체화하지 않고 설계 단계에서 사업자 임의로 결정하는 사례가 빈번하여 복합적인 리스크로 이어진다.

▶ 벡터 DB 제품별 성능·비용·운영 특성의 현저한 차이: Milvus, Weaviate, Qdrant, Pinecone, PostgreSQL+pgvector 등 주요 제품은 검색 정확도(Recall@10), P95 응답시간, 최대 벡터 수, 운영 복잡도, 라이선스 비용 측면에서 최대 수십 배 이상의 차이를 보이므로, 제품 선정 기준을 요구사항에 정량적으로 명시해야 한다.

▶ 사후 교체의 기술적·비용적 부담: 벡터 DB를 일단 도입한 후 변경하려면 전체 데이터 재임베딩, 인덱스 재구축, 연동 시스템 수정이 수반되며, 데이터 규모에 따라 수개월의 추가 작업이 발생할 수 있어 사업의 예산·일정·품질 목표를 훼손하는 직접적 요인이 된다.

▶ RAG 데이터 규모 미산정으로 인한 인프라 비용 초과: RAG 대상 문서 수, 청크 분할 전략, 임베딩 차원, QPS(초당 쿼리 수)는 벡터 저장소 용량과 컴퓨팅 비용에 직결되며, 요구정의 단계에서

이를 산정하지 않으면 구현 단계에서 인프라 비용이 예산을 초과하거나 성능 요건을 충족하지 못하는 상황이 발생한다.

▶ 데이터 거버넌스 부재로 인한 컴플라이언스 위험: RAG 인덱스에 민감 문서(개인정보, 비밀 등급 문서 등)가 무분별하게 포함되면 권한 없는 사용자가 검색 결과를 통해 해당 정보에 접근할 수 있으므로, 권한별 검색 범위 필터링, 인덱스 분리, 삭제 반영 절차를 요구사항으로 명시하지 않으면 운영 단계의 보안 사고로 이어질 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	벡터 DB 제품 선정 기준 (6차원)	{요구사항정의서}에서 검색 정확도(Recall@10), P95 응답 시간, 최대 벡터 수·확장성, 메타데이터 필터, 보안, 5년 TCO의 6개 차원이 선정 기준으로 정량 제시되었는지 확인
2	RAG 데이터 규모 산정 (5요소)	{요구사항정의서}에서 대상 문서 수, 청크 분할 전략, 임베딩 차원, 벡터 수, QPS의 5개 요소가 수치로 산정되어 저장 용량·컴퓨팅 요구량 산출 근거로 제시되었는지 확인
3	RAG 운영 SLA (6개 지표)	{요구사항정의서}에서 검색 정확도, 응답시간, 처리량, 가용률, 인덱스 갱신 지연, 오류율 등 운영 SLA 지표와 임계값이 명시되었는지 확인
4	데이터 출처·라이선스 관리	RAG 인덱스에 적재할 문서의 출처·저작권·라이선스 확인 절차가 {요구사항정의서}에 정의되었는지 확인하고, 외부 수집 문서의 이용 허락 범위를 면담으로 확인
5	민감 문서 인덱스 분리·권한 필터링	개인정보·비밀등급 문서의 인덱스 분리와 사용자 권한별 검색 범위 필터링이 {요구사항정의서}에 정의되었는지 확인하고, 권한 변경 시 재색인 여부를 면담으로 확인
6	인덱스 갱신·삭제 반영 절차	원본 문서 수정·삭제 시 벡터 인덱스에 반영되는 주기와 절차(재임베딩·소프트 삭제 등)가 {요구사항정의서}에 정의되었는지 확인
7	Agentic AI 도구 정의·승인 절차	에이전트가 호출 가능한 도구(Tool) 목록과 신규 도구 추가 시 승인 절차, 권한 범위가 {요구사항정의서}에 정의되었는지 확인
8	인프라 비용 추정 (TCO 5년)	벡터 저장소·임베딩·추론 비용을 포함한 5년 총소유비용(TCO) 추정 근거와 산정식이 제시되었는지 확인
9	{과업대비표} 요구사항 추적 완전성	RFP의 벡터 DB·RAG 관련 요구사항이 {과업대비표}를 통해 {요구사항정의서}에 누락 없이 연결되는지 대조 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	벡터 DB 제품 선정 기준(검색 정확도·P95 응답속도·확장성·메타데이터 필터·보안·TCO), RAG 데이터 규모 산정, 운영 SLA 임계값, 데이터 거버넌스 요건(출처·라이선스 관리, 민감 문서 인덱스 분리·권한 필터링), 원본 문서 수정·삭제 시 인덱스 갱신·삭제 반영 절차, Agentic AI가 호출 가능한 도구 목록과 신규 도구 추가 승인 절차·권한 범위가 {요구사항정의서}에 정량적으로 명시된 경우	현행 유지
보완권고 (△)	주요 요건은 정의되었으나 일부 선정 기준·SLA 임계값·데이터 거버넌스 요건의 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	벡터 DB·RAG 관련 요구사항이 미반영되었거나 선언적 문구 수준에 그쳐 설계 단계에서 사업자 임의 결정 위험이 있는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 기술 트렌드의 지속적 이해: 벡터 DB 및 RAG 파이프라인은 최근 2~3년 사이 급속히 발전한 기술 영역이므로, 감리원도 기술 트렌드에 대한 지속적인 이해를 바탕으로 오류 패턴과 실무적 판단 기준을 점검에 반영해야 한다.

2) 사후 교체 부담 고려: 벡터 DB는 도입 후 변경 시 전체 데이터 재임베딩·인덱스 재구축·연동 시스템 수정이 수반되므로, 제품 선정 기준이 정량적으로 명시되었는지를 우선 확인한다.

3) 인프라 비용 사전 산정: RAG 대상 문서 수·청크 분할 전략·임베딩 차원·QPS는 저장소 용량과 컴퓨팅 비용에 직결되므로, 5년 TCO 추정이 요구사항에 포함되었는지 점검한다.

4) 민감 문서 거버넌스 확인: RAG 인덱스에 민감 문서가 포함될 경우를 대비해 권한별 검색 범위 필터링·인덱스 분리·삭제 반영 절차가 요구사항으로 명시되었는지 면담을 통해 확인한다.

AG_SD41-3-11 | 데이터 분할 계획 및 검증용(Hold-out) 데이터 확보

데이터베이스 | 요구정의·분석 | 데이터 수집·처리 | 테스트시나리오

근거 | NIST AI RMF MEASURE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 성능을 객관적으로 평가하기 위한 학습·검증·테스트 데이터 분할(Data Split) 계획과 학습 과정에서 사용되지 않는 독립 검증용(Hold-out) 데이터 확보 계획이 {테스트계획서}에 구체적으로 수립되어 있는지를 제3자적 관점에서 점검하는 것이 목적이다.

AI 모델은 일반 소프트웨어와 달리 학습 데이터에 과도하게 최적화되는 과적합(Overfitting)이 발생할 수 있으며, 이를 사전에 방지하고 실제 운영 환경에서의 성능을 예측하기 위해 Hold-out 데이터를 학습 단계에서 철저히 격리하여 관리하여야 한다.

분석 단계부터 데이터 분할 계획이 구체적으로 수립되지 않으면 사업 완료 후 모델 성능 평가의 신뢰성 자체가 훼손되는 리스크가 발생하므로, 이를 사전에 예방하는 데 초점을 둔다.

2. 필요성

▶ AI 모델의 성능은 학습에 사용한 데이터로 측정하면 실제 운영 성능보다 과대평가되는 구조적 문제가 있다. 데이터 분할 계획이 분석 단계부터 수립되지 않으면 다음과 같은 리스크가 현실화된다.

▶ 학습 데이터와 테스트 데이터가 혼용되어 검사(Test) 단계의 성능 측정 결과가 실제 운영 성능과 크게 다른 "낙관적 편향(Optimistic Bias)"이 발생한다.

▶ Hold-out 데이터가 사전에 확보·격리되지 않으면 개발 과정에서 테스트 데이터가 암묵적으로 학습에 영향을 미치는 데이터 누수(Data Leakage)이 발생하여 성능 평가가 무의미해진다.

▶ 데이터 분할 비율·방법·이력이 문서화되지 않으면, 동일한 조건으로 모델을 재현하거나 성능을 재검증하는 것이 불가능해진다.

▶ 분할 데이터에 대한 접근 통제가 없으면 개발자가 테스트 데이터를 사전에 열람하여 의도적·비의도적으로 모델에 반영하는 "테스트 데이터 오염" 문제가 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 분할 방식 및 비율 명시 여부	{테스트계획서}에 전체 데이터를 학습(Train)·검증(Validation)·테스트(Test) 세트로 분할하는 방식과 비율이 명시되어 있는지 확인
2	Hold-out 검증 데이터 별도 확보·격리 계획 여부	Hold-out 테스트 데이터가 학습 시작 전에 별도로 분리되어 개발 과정에서 완전히 격리 보관되는 계획이 {테스트계획서}에 명시되어 있는지 확인한다. 확보 시점, 확보 규모(건수·전체 대비 비율), 최종 평가에 사용하는 시점이 함께 기술되었는지 확인
3	데이터 분포 대표성 확보 방안 포함 여부	분할된 각 데이터셋이 실제 운영 환경의 데이터 분포를 대표할 수 있도록 하는 방안이 계획에 포함되어 있는지 확인
4	데이터 누수(Data Leakage) 방지 계획 명시 여부	학습 데이터 정보가 검증·테스트 데이터에 의도치 않게 흘러들어 성능 평가를 왜곡하는 데이터 오염을 방지하는 절차가 {테스트계획서}에 명시되어 있는지 확인
5	분할 데이터 보안 및 접근통제 방안 명시 여부	Hold-out 테스트 데이터 및 각 분할 세트에 대한 접근 권한 통제와 보안 관리 방안이 {테스트계획서}에 명시되어 있는지 확인
6	데이터 분할 이력 관리 및 재현 가능성 확보 여부	데이터 분할 결과를 재현하고 감사할 수 있도록 분할 이력(분할 날짜, 방법, 랜덤 시드값, 담당자 등)을 관리하는 계획이 명시되어 있는지 확인
7	RAG 검색 평가용 Hold-out 셋 확보 여부	RAG 검색 품질 평가를 위한 독립 질의-정답 Hold-out 셋이 색인 대상 문서와 분리·격리되어 확보되는 계획이 {테스트계획서}에 수립되었는지 확인
8	RFP 요건과 {테스트계획서} 일치성	{제안요청서(RFP)}에 명시된 데이터 분할·Hold-out 관련 요건이 {테스트계획서}에 빠짐없이 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 서비스 특성에 맞는 데이터 분할 방식·비율이 수치로 명시되고, Hold-out 데이터의 확보 시점·규모·격리 방안·접근통제·최종 평가 사용 시점이 모두 기술되며, 데이터 오염 방지 절차, 분할 데이터의 분포 대표성 확보 방안, 분할 이력 관리 방안이 {테스트계획서}에 구체적으로 포함되고, RAG 활용 시 검색 품질 평가용 독립 Hold-out 셋이 색인 대상 문서와 분리·격리되어 확보 계획에 반영된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	분할 비율 미수치화, Hold-out 격리 방안 부재·불명확, 데이터 오염 방지 절차 누락, 클래스 불균형 대응 방안 미포함, 분할 이력 관리 계획 부재 등 핵심 요소 중 하나 이상이 결여된 경우	누락 항목 보완 권고 및 다음 감리 전까지 수정 확인
부적합 (X)	데이터 분할 계획 및 Hold-out 데이터 확보·격리 계획이 {테스트계획서}에 전혀 수립되지 않은 경우. 단, 외부 사전 학습 완료 모델(Foundation Model)을 추가 학습(Fine-tuning) 없이 API 형태로만 활용하여 자체 데이터 분할이 불필요한 구조인 경우에는 해당 모델의 성능 평가 기준 및 검증 방법이 대체 수단으로 계획되어 있는지 별도 확인한다	부적합 사항 시정조치 요구 및 이행 결과 확인 (API 활용 모델은 성능 검증 계획 대체 확인 후 기록)

5. 감리 시 유의사항 및 실무 Tip

1) 선언과 실질 계획의 구분: "데이터를 분할한다"는 선언과 "실질적으로 격리된 분할 계획"을 반드시 구분한다. {테스트계획서}에 "학습 데이터와 테스트 데이터를 분리하여 관리한다"는 문구가 있어도, 격리 시점·방법·접근통제 방안이 없으면 형식적 선언에 그친다.

2) Hold-out 격리 절차의 핵심 점검: 감리원은 Hold-out 데이터의 격리를 위한 구체적 절차(별도 저장 공간, 접근 권한 분리, 격리 시점)가 명시되어 있는지를 핵심 점검 기준으로 삼아야 한다. 사업PM 인터뷰 시 "Hold-out 데이터를 언제, 누가, 어디에 분리·보관하는지"를 직접 질의하여 계획의 실질성을 확인한다.

3) 분할 이력 문서화로 재현 가능성 확보: 분할 날짜·방법·랜덤 시드값·담당자 등 분할 이력이 문서화되지 않으면 동일 조건으로 모델을 재현하거나 성능을 재검증하는 것이 불가능하므로, 분할 이력 관리 계획의 명시 여부를 확인한다.

4) 테스트 데이터 접근 통제 확인: 분할 데이터에 대한 접근 통제가 없으면 개발자가 테스트 데이터를 사전에 열람하여 의도적·비의도적으로 모델에 반영하는 테스트 데이터 오염이 발생할 수 있으므로, 접근 권한 통제와 보안 관리 방안이 {테스트계획서}에 명시되었는지 확인한다.

AG_SD42-3-01 | 편향방지 피처엔지니어링 점검

데이터베이스 | 설계 | 데이터 수집·처리 | 편향/공정성

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 내용이 데이터베이스 설계 산출물에 빠짐없이 상세하고 구체적으로 반영되었는지 여부를 '감리원'이 제3자적 관점에서 검토하기 위한 것이다.

구체적으로, AI 모델 학습 전 단계에서 수행되는 피처 엔지니어링(Feature Engineering) — 특성 선택(Selection), 변환(Transformation), 생성(Generation) — 설계가 '전처리설계서'에 반영되었는지, 이 과정에서 편향(Bias)이 유입되거나 심화되지 않도록 하는 설계적 방안이 포함되었는지, 그리고 피처 엔지니어링 결과와 이력이 DB 테이블에 관리되는 구조가 {테이블정의서}에 반영되었는지를 점검하는 것을 목적으로 한다.

2. 필요성

▶ 피처 엔지니어링은 원본 데이터에서 AI 모델이 학습에 사용할 특성(Feature)을 선택·변환·생성하는 과정으로, 모델 성능과 공정성에 결정적인 영향을 미친다. 이 과정의 설계가 누락되거나 편향 방지 고려 없이 수행되면 다음과 같은 위험이 발생한다.

▶ 프록시 변수를 통한 간접 편향 유입: 보호변수(성별, 연령 등)를 직접 사용하지 않더라도, 보호변수와 높은 상관관계를 가진 프록시 변수(예: 우편번호→지역→인종, 이름→성별)가 특성으로 선택되면 간접적으로 편향이 유입된다. TTA 안내서 요구사항 07-2a는 보호변수 선정 시 충분한 분석을 요구하고 있다.

▶ 과도한 특성 제거로 인한 정보 손실: TTA 안내서 요구사항 07-2c는 전처리 시 특성이 과도하게 제거되었는지 검토하도록 요구하고 있다. 편향 방지를 이유로 유의미한 특성까지 무분별하게 제거하면 모델 성능이 저하될 수 있다.

▶ 특성 변환의 비재현성: 특성 변환 규칙(인코딩 방식, 스케일링 범위, 파생 변수 산출식 등)이 설계 문서에 명시되지 않으면 개발자마다 다른 방식으로 변환을 수행하여 학습 결과의 재현이 불가능해진다.

▶ 특성 이력 추적 불가: 어떤 원본 컬럼에서 어떤 변환을 거쳐 최종 학습 특성이 생성되었는지 이력이 DB에 기록되지 않으면, 모델 추론 결과의 원인 분석이 곤란하다.

3. 검토방법

No	점검 항목	확인 방법
1	최종 특성 목록·선택 근거 명시 여부	학습에 사용할 최종 특성(Feature) 목록과 각 특성의 선택 근거가 '전처리설계서'에 명시되었는지 확인
2	프록시 변수 식별·처리 방안 기술 여부	보호변수와 높은 상관관계를 가진 프록시 변수의 식별·처리 방안이 '전처리설계서' 편향 분석 섹션에 기술되었는지 확인
3	특성 변환 규칙 컬럼별 정의 여부	인코딩 방식·스케일링 범위·파생 변수 산출식 등 특성 변환 규칙이 '전처리설계서'에 컬럼별로 구체적으로 정의되었는지 확인
4	피처 스토어·피처 테이블 설계 여부	피처 엔지니어링 결과(최종 특성값)를 저장하는 피처 스토어 또는 피처 테이블이 {테이블정의서}·{ERD}에 설계되었는지 교차 검토
5	피처 엔지니어링 이력 테이블 설계 여부	변환 일시·적용 규칙·입출력 특성 수·수행자 등 피처 엔지니어링 수행 이력을 기록하는 이력 테이블이 {테이블정의서}에 설계되었는지 확인
6	편향 방지·특성 추출 요구사항 추적 가능 여부	{요구사항추적표}의 편향 방지·특성 추출 요구사항이 '전처리설계서'·{테이블정의서}의 구체적 설계로 추적 가능한지 매핑 검토
7	임베딩 공간 보호변수 편향 측정 설계 여부	임베딩 공간에서 보호변수 그룹별 코사인 유사도·거리 분포 편향을 측정·저장하는 설계가 '편향 측정 설계' 또는 '품질 관리 설계서'에 포함되었는지 확인
8	프롬프트 편향 유발 표현 검토 절차 설계 여부	프롬프트 템플릿·시스템 메시지에 특정 성별·직업 고정 표현 등 편향 유발 표현이 포함되지 않도록 검토하는 절차가 '프롬프트 거버넌스 설계'에 설계되었는지 확인
9	보호변수 프록시 검토 절차 설계 여부	우편번호·이름 등 간접 식별 가능 프록시 피처에 대한 검토 절차가 '피처 정의서' 또는 '편향 설계 검토 회의록'에 설계되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	최종 특성 목록과 선택 근거, 프록시 변수 식별·처리 방안, 컬럼별 변환 규칙이 '전처리설계서'에 구체적으로 정의되고, 피처 테이블·변환 이력 테이블이 {테이블정의서}에 설계되어 {요구사항추적표}로 추적 가능하며, 임베딩 공간에서 보호변수 그룹별 코사인 유사도·거리 분포의 편향을 측정·저장하는 설계와 프롬프트 템플릿·시스템 메시지의 편향 유발 표현 검토 절차가 포함되는 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	특성 목록과 변환 규칙은 정의되었으나 프록시 변수 분석 누락, 피처 이력 테이블 미설계, 임베딩 편향 측정·프롬프트 편향 검토 절차 미비 등 일부 설계 요소가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	피처 엔지니어링 설계가 '전처리설계서'에 전혀 반영되지 않았거나, 편향 방지 고려 없이 특성 선택·변환이 수행되도록 설계되어 편향 유입·재현 불가 위험이 존재하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '프록시 변수 분석 결과서' 확인: '전처리설계서' 내에 프록시 변수 분석이 별도 섹션 또는 첨부로 포함되어 있는지 확인한다. 분석 방법은 상관계수 분석, 카이제곱 독립성 검정 등 다양하며, 감리원은 방법의 적합성보다 '분석이 수행되었고 결과가 설계에 반영되었는지'를 확인하는 것이 핵심이다.

2) 피처 스토어(Feature Store) 설계 수준 확인: 최신 MLOps 체계에서는 피처 스토어를 별도 시스템으로 구축하기도 하지만, DB 감리 관점에서는 {테이블정의서}에 피처 저장 테이블이 설계되어 있는지를 확인하면 충분하다. 별도 피처 스토어 시스템의 기술적 구현은 시스템구조 영역의 점검 대상이다.

3) 생성형 AI 특화 피처: LLM 기반 서비스의 경우, 텍스트 임베딩(Embedding) 벡터가 핵심 특성이다. 임베딩 모델명, 벡터 차원 수, 임베딩 생성 일시 등의 메타정보가 피처 테이블 또는 관련 메타 테이블에 관리되는 설계인지 확인한다. 또한 프롬프트 특성(프롬프트 길이, 토큰 수, 언어 등)이 학습·분석 목적으로 추출·저장되는 설계 여부도 점검 대상이다.

4) 변환 규칙의 '기준값 저장' 확인: Min-Max 스케일링의 최솟값·최댓값, Z-Score의 평균·표준편차 등 변환 기준값은 추론 시에도 동일하게 적용되어야 한다. 이 기준값을 DB에 저장하는 테이블(예: TB_TRANSFORM_PARAM)이 설계되었는지 확인하면, 학습-추론 간 변환 일관성을 확보하는 설계인지 판단할 수 있다.

5) 운영 산출물 연계: 피처 테이블을 기반으로 한 '피처 모니터링 대시보드 설계서', '특성 드리프트(Feature Drift) 탐지 기준 정의서' 등 운영 산출물 작성 계획이 수립되었는지 확인한다. 운영 과정 자체의 점검사항은 본 가이드라인 범위 밖이나, 운영을 고려한 산출물 계획의 적정성은 점검 대상이다.

AG_SD42-3-02 | 학습데이터 버전-모델 매핑점검

데이터베이스 | 설계 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 내용이 데이터베이스 설계 산출물에 빠짐없이 상세하고 구체적으로 반영되었는지 여부를 '감리원'이 제3자적 관점에서 검토하기 위한 것이다.

구체적으로, AI 학습에 사용된 데이터셋의 버전을 관리하는 테이블이 {테이블정의서}에 설계되었는지, 그리고 특정 AI 모델이 어떤 버전의 데이터셋으로 학습되었는지를 역추적할 수 있도록 모델 버전과 데이터셋 버전을 매핑(Mapping)하는 컬럼이 설계에 포함되었는지를 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템에서는 학습 데이터셋이 지속적으로 갱신·보완되며, 동일 모델이라도 학습에 사용된 데이터셋 버전에 따라 성능이 달라진다. 데이터셋 버전과 모델 버전 간의 매핑이 DB 수준에서 관리되지 않으면 다음과 같은 위험이 발생한다.

▶ 모델 재현성(Reproducibility) 확보 실패: 특정 모델의 학습 결과를 재현하려면 해당 모델이 어떤 버전의 데이터셋으로 학습되었는지 정확히 알아야 한다. 매핑이 없으면 동일 조건의 재학습이 불가능하다.

▶ 모델 성능 저하 원인 규명 곤란: 운영 중 모델 성능이 저하되었을 때, 데이터셋 버전 변경이 원인인지 모델 변경이 원인인지 구분할 수 없다. TTA 생성 AI 안내서 요구사항 04-2e는 신규 데이터 확보 시 모델 성능평가 재수행을 요구하고 있다.

▶ 데이터 변경 영향 추적 불가: TTA 안내서 요구사항 04-2c(데이터 변경 시 버전 관리), 04-2d(이해관계자를 위한 정보 제공)는 데이터 변경의 체계적 관리를 요구하고 있으며, 버전 매핑 없이는 변경 영향을 추적할 수 없다.

▶ 규제 대응 및 감사 추적 곤란: AI 시스템에 대한 사후 감사 시 '현재 운영 중인 모델이 어떤 데이터로 학습되었는가?'라는 기본적인 질문에 답할 수 없으면, 모델의 투명성 및 책임성을 확보할 수 없다.

3. 검토방법

No	점검 항목	확인 방법
1	학습 데이터셋 버전 관리 전용 테이블 정의 여부	학습 데이터셋 버전 관리 전용 테이블(예: TB_DATASET_VERSION)이 {테이블정의서}에 독립적으로 정의되었는지 확인
2	버전 테이블 필수 컬럼 포함 여부	버전 ID·버전 번호·생성일시·변경 사유·데이터 건수·스냅샷 참조 등 필수 컬럼이 데이터셋 버전 테이블에 포함되었는지 {테이블정의서} 컬럼 목록에서 확인
3	모델-데이터셋 버전 매핑 설계 여부	모델 버전과 데이터셋 버전을 매핑하는 테이블 또는 컬럼(TB_MODEL_DATASET_MAP, 모델 이력 테이블 내 DATASET_VER_ID FK 등)이 설계되었는지 {테이블정의서}·{ERD}에서 교차 검토
4	데이터셋 버전 테이블 참조 관계 연결 여부	데이터셋 버전 테이블이 원천 데이터·모델 학습 이력·추론 결과 테이블 등과 참조 관계로 연결되었는지 {ERD} 관계선에서 확인
5	버전 간 차이(Diff) 정보 저장 설계 여부	추가·삭제·변경 건수 등 데이터셋 버전 간 차이(Diff) 정보를 저장하는 컬럼 또는 별도 테이블이 설계되었는지 {테이블정의서}에서 확인
6	데이터 버전 관리 요구사항 추적 가능 여부	{요구사항추적표}의 데이터 버전 관리·모델 추적 요구사항이 {테이블정의서}의 구체적 테이블·컬럼으로 추적 가능한지 매핑 검토
7	임베딩 모델-컬렉션-색인-응답 시점 매핑 설계 여부	임베딩 모델 버전·임베딩 저장 벡터 DB 컬렉션·색인(Index) 파라미터·응답 시점 매핑이 설계되었는지 '벡터 DB 설계서'·'모델 운영 설계서'에서 확인
8	RAG 응답 생성 정보 기록 설계 여부	RAG 기반 서비스에서 응답 생성에 사용된 LLM 버전·임베딩 모델 버전·청크 정책·검색 컬렉션 스냅샷 ID가 추론 결과에 함께 기록되도록 설계되었는지 '추론 결과 저장 설계'에서 확인
9	파인튜닝·LoRA 어댑터 버전 매핑 설계 여부	파인튜닝·LoRA 어댑터 버전 관리 시 기반 모델 ID·어댑터 ID·학습 데이터셋 버전·평가셋 버전이 1:1 매핑 가능하도록 설계되었는지 '모델 학습 이력 설계'에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	데이터셋 버전 관리 테이블과 필수 컬럼이 {테이블정의서}에 정의되고, 모델 버전-데이터셋 버전 매핑 구조가 {ERD}상 참조 관계로 연결되어 있으며, RAG·파인튜닝 등 생성형 AI 특화 버전 매핑까지 설계되어 양방향 추적이 가능한 경우	현행 유지
보완권고 (△)	버전 관리 테이블은 설계되었으나 매핑 컬럼(FK) 누락, 버전 간 차이(Diff) 정보 미설계, 임베딩 모델·컬렉션 매핑 또는 추론 로그 버전 기록 등 일부 설계 요소가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	데이터셋 버전 관리 테이블과 모델-데이터셋 매핑 설계가 {테이블정의서}에 전혀 반영되지 않아 모델 재현성과 감사 추적성을 확보할 수 없는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 버전 네이밍 규칙 확인: 데이터셋 버전 번호가 체계적인 네이밍 규칙(예: 시맨틱 버전닝 MAJOR.MINOR.PATCH)으로 관리되는 설계인지 확인한다. 'v1', 'v2', '최종', '최종_수정' 등 비체계적 명명은 혼란을 유발하므로, {데이터베이스설계서}에 네이밍 규칙이 정의되었는지 점검한다.

2) '양방향 추적' 시나리오 검증: 감리원은 사업자에게 다음 두 가지 시나리오를 제시하여 설계의 실효성을 확인하는 것이 효과적이다. (1) '현재 운영 모델 v3.0은 어떤 데이터셋 버전으로 학습되었는가?' (2) '데이터셋 v2.1에서 편향이 발견되었는데, 이 데이터셋으로 학습된 모든 모델은 무엇인가?'

3) 생성형 AI 특화 버전 관리: LLM 기반 서비스의 경우, 사전학습(Pre-training) 데이터셋, 파인튜닝(Fine-tuning) 데이터셋, RAG 문서 컬렉션 등 여러 종류의 데이터셋이 하나의 모델에 연결된다. 이 경우 매핑 테이블이 데이터셋 유형(DATASET_TYPE_CD: PRE_TRAIN, FINE_TUNE, RAG_DOC 등)을 구분하는 컬럼을 포함하도록 설계되었는지 확인한다.

4) 데이터셋 스냅샷 전략 확인: 데이터셋 버전 테이블에 스냅샷 저장 경로(SNAPSHOT_PATH) 컬럼이 있더라도, 실제 스냅샷이 전체 복제(Full Copy)인지 변경분만 저장(Delta)인지에 따라 DB 설계가 달라진다. 설계서에 스냅샷 전략이 명시되었는지 확인하되, 스토리지 수준의 기술적 구현은 시스템구조 영역의 점검 대상이므로 DB 감리원은 테이블 설계 반영 여부까지만 점검한다.

5) 운영 산출물 연계: 버전 테이블을 기반으로 한 '데이터셋 변경 이력 리포트 템플릿', '모델-데이터셋 매핑 현황 대시보드 설계서' 등 운영 산출물 작성 계획이 수립되었는지 확인한다. 운영 과정 자체의 점검사항은 본 가이드라인 범위 밖이나, 운영을 고려한 산출물 계획의 적정성은 점검 대상이다.

AG_SD42-3-03 | 비식별화-가명화 뷰 설계점검

데이터베이스 | 설계 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습·운영 데이터에 포함된 개인정보 및 민감정보에 대해, 데이터베이스 설계 수준에서 비식별화 보호 설계가 적정하게 반영되었는지를 감리원이 제3자적 관점에서 검토하기 위한 것이다.

EDB(RDB) 영역에서는 실 테이블(베이스 테이블)의 원본 데이터에 직접 접근하지 않고 비식별화(masking)된 데이터만 조회할 수 있는 가명화 뷰(View)가 {테이블정의서}에 설계되었는지, 뷰 로직 내 마스킹 함수·규칙이 구체적으로 정의되었는지, 뷰에 대한 접근 권한이 역할별로 분리 설계되었는지를 점검한다.

벡터 DB 영역에서는 RAG 파이프라인에서 사용되는 청크(Chunk) 원문 텍스트에 개인정보가 포함되는 경우, 임베딩 생성 전 단계에서의 비식별화 처리 설계 여부 및 벡터 DB API 레이어에서의 접근 제어 설계 여부를 함께 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템의 학습·운영 데이터에는 개인식별정보(주민번호, 전화번호 등)와 보호변수(성별, 연령 등)가 포함될 수 있다. EDB(RDB)와 벡터 DB를 함께 운용하는 AI 사업에서 비식별화 보호 설계가 부재하면 다음과 같은 복합적 위험이 발생한다.

▶ 개인정보 과다 노출: 개발자·데이터 분석가·라벨러 등이 학습 데이터 조회 시 원본 개인정보가 그대로 노출되면, 개인정보 보호법상 필요 최소한의 개인정보 처리 원칙에 위배될 수 있다.

▶ AI 모델을 통한 개인정보 복원 위험(Memorization): 생성형 AI의 경우, 학습 데이터에 포함된 개인정보가 모델에 기억(Memorization)되어 추론 결과에 노출될 수 있다. 학습 전 단계에서 비식별화 접근이 설계되어야 이 위험을 사전에 차단할 수 있다.

▶ 벡터 DB 원문 텍스트 노출: RAG용 벡터 DB에는 문서를 청크 단위로 분할하여 저장하는데, 이 청크 원문 텍스트(대화 내역, 문서 본문 등)에 개인정보가 포함될 수 있다. 벡터 DB는 SQL 기반 뷰(View) 개념이 없으므로, 임베딩 생성 전 전처리 단계에서의 비식별화가 설계되지 않으면 원문 개인정보가 직접 조회·노출될 수 있다.

▶ 임베딩 역추론(Embedding Inversion Attack) 위험: 벡터는 수치 데이터이지만 고정밀 역변환 연구를 통해 원문 텍스트를 상당 부분 복원할 수 있다는 연구 결과가 있다. 원문 청크 저장 방식과 벡터 DB 접근 권한 설계를 점검할 필요가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	원본 테이블 비식별화 뷰 정의 여부	개인정보·민감정보 포함 원본 테이블에 대한 비식별화(masking) 뷰가 {테이블정의서}에 정의되었는지 확인
2	컬럼별 마스킹 함수·규칙 정의 여부	SUBSTR·REPLACE·HASH 등 마스킹 함수·규칙이 뷰 로직 내 컬럼별로 구체적으로 정의되었는지 {테이블정의서}에서 확인
3	뷰 접근 권한 역할별 분리 설계 여부	개발자→마스킹 뷰·DBA→원본 테이블 등 뷰 접근 권한이 역할별로 분리 설계되었는지 {데이터베이스설계서} 접근 제어 섹션에서 확인
4	원본 테이블 우회 접근 차단 설계 여부	뷰를 우회한 원본 테이블 직접 접근 경로를 차단하는 GRANT/REVOKE 정책 등이 {데이터베이스설계서}에 포함되었는지 확인
5	역할별 차등 마스킹 다중 뷰 설계 여부	분석가용 부분마스킹 뷰·개발자용 완전마스킹 뷰 등 역할별 차등 마스킹 다중 뷰 설계가 고려되었는지 {테이블정의서}에서 확인
6	비식별화·접근 제어 요구사항 추적 가능 여부	{요구사항추적표}의 비식별화·접근 제어 요구사항이 {테이블정의서}의 구체적 뷰명·마스킹 규칙으로 추적 가능한지 매핑 검토
7	벡터 DB 청크의 임베딩 전 비식별화 설계 여부	RAG용 벡터 DB에 저장되는 청크(원문 텍스트)에 개인정보 포함 시 임베딩 생성 전 비식별화 처리 절차가 'AI 파이프라인 설계서'·'데이터 비식별화 계획서'에 설계되었는지 확인
8	벡터 DB 컬렉션 접근 권한 분리 설계 여부	벡터 검색 연산 역할과 원문 텍스트 직접 조회 역할 등 벡터 DB 컬렉션 접근 권한이 역할별로 분리 설계되었는지 '벡터 DB 설계서'·'AI 데이터 아키텍처 설계서'에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	개인정보 포함 테이블에 대한 마스킹 뷰와 컬럼별 마스킹 규칙이 {테이블정의서}에 구체적으로 정의되고, 역할별 접근 권한 분리·원본 직접 접근 차단 설계 및 벡터 DB 청크의 임베딩 전 비식별화·접근 제어 설계가 모두 확인되는 경우	현행 유지
보완권고 (△)	마스킹 뷰는 설계되었으나 일부 컬럼의 마스킹 규칙 미정의, 역할별 차등 마스킹 미고려, 원본 테이블 우회 접근 차단 정책 또는 벡터 DB 접근 제어 설계 미흡 등 부분적 보완이 필요한 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	개인정보 포함 테이블에 대한 비식별화 뷰·마스킹 규칙이 전혀 설계되지 않았거나, 벡터 DB 청크 원문의 비식별화 처리가 누락되어 개인정보 보호법 위반 위험이 존재하는 경우	시정조치 요구 및 이행 결과 확인 (개인정보 보호법 준수 촉구)

5. 감리 시 유의사항 및 실무 Tip

1) 벡터 DB에 뷰가 없다는 점 명시: 벡터 DB는 SQL 기반 뷰(View) 개념이 없으므로, EDB와 동일한 가명화 뷰 설계를 요구하는 오류를 범하지 않도록 주의한다. 단, PostgreSQL의 pgvector 확장 사용 시 Row-Level Security 및 뷰 설계 적용이 가능하므로 이 경우에는 EDB 기준으로 점검할 수 있다.

2) DBMS 내장 마스킹 기능 활용 가능성 확인: Oracle Dynamic Data Masking, PostgreSQL Row-Level Security, SQL Server Dynamic Data Masking 등 DBMS 자체의 마스킹 기능을 활용하는 설계인 경우, 뷰 대신 DBMS 정책 기반 마스킹이 적용될 수 있다. 이 경우에도 마스킹 대상 컬럼·규칙이 설계서에 구체적으로 명시되어 있는지 확인한다.

3) 학습 파이프라인의 접근 경로 확인: AI 모델 학습 배치가 EDB에 접근할 때 마스킹 뷰를 통하는지, 원본 테이블을 직접 조회하는지를 설계서에서 확인한다. 학습 목적상 원본이 필요한 경우(예: 정확한 수치 기반 학습)에는 별도의 비식별화 테이블(물리적 복제+마스킹)을 통한 접근 설계가 대안이 될 수 있다.

4) 마스킹 함수 성능 영향 고려: 복잡한 정규표현식(Regex) 함수나 해싱(Hashing) 로직이 포함된 뷰는 조회 성능에 영향을 미칠 수 있다. {데이터베이스설계서}에 뷰 성능 영향 분석 또는 물리적 마스킹 테이블(Materialized View 등) 대안이 검토되었는지 확인하되, DB 감리원은 '성능 고려 여부가 설계서에 기술되었는지'만 점검한다.

5) 모호한 설계 표현의 구체화 유도: '적절히 비식별화한다'와 같은 모호한 표현은 향후 분쟁의 원인이 된다. 예를 들어 '전화번호 컬럼에 SUBSTR(PHONE, 1, 3) || '-****-' || SUBSTR(PHONE, 8) 함수를 적용한 VW_MEMBER_MASKED 뷰를 정의한다'와 같이 컬럼 단위로 마스킹 규칙을 명시하도록 유도하고, 벡터 DB의 경우 임베딩 전 비식별화 처리 규칙을 동일한 수준으로 구체화하도록 권고한다.

AG_SD42-3-04 | 데이터 정합성 검증 배치 설계점검

데이터베이스 | 설계 | 데이터 수집·처리 | 데이터 품질

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습·서비스 데이터의 정합성을 주기적으로 검증하는 배치(Batch) 프로세스의 설계가 '전처리설계서'에 적정하게 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

EDB(RDB) 영역에서는 Null 값, 형식(Format) 오류, 도메인 범위 위반, 참조 무결성 위반, 중복 데이터 등 정합성 문제가 빈번히 발생하므로 검증 대상·검증 규칙·실행 스케줄·검증 결과 저장·오류 데이터 처리 절차 등을 종합적으로 검토하며, RAG·Agent 서비스에서 벡터 DB를 병행 운용하는 경우 임베딩 벡터의 차원 일치성, EDB 원문 데이터와 벡터 DB 임베딩 간의 동기화 정합성(고아 임베딩·누락 임베딩), 임베딩 모델 버전 일치성 등 벡터 DB 고유의 검증 배치 설계 여부를 함께 점검한다.

2. 필요성

▶ AI 서비스에서 데이터 정합성 검증 배치는 학습 데이터 품질의 마지막 방어선으로, EDB와 벡터 DB를 함께 운용하는 AI 사업에서 검증 배치 설계가 부재하면 복합적 위험이 발생한다.

▶ AI 학습 데이터 품질 보증: Null 값, 형식 불일치, 범위 초과 등의 데이터가 학습셋에 포함되면 모델이 잘못된 패턴을 학습하므로, DB 제약조건(NOT NULL, CHECK)만으로 포착하기 어려운 복합적 정합성 문제를 주기적 배치 검증으로 탐지하여 오류 데이터의 학습·서비스 활용을 사전에 차단한다.

▶ 복합 검증 규칙 실행: DB 제약조건으로 검증 가능한 규칙 외에 '연령 0~150 범위인가', '이메일 형식이 올바른가', '코드값이 코드 테이블에 존재하는가' 등 업무·도메인 수준의 복합 규칙은 배치 검증으로만 확인할 수 있다.

▶ 데이터 드리프트 조기 감지: 입력 데이터의 분포가 학습 시점과 달라지는 데이터 드리프트를 검증 배치에서 통계적으로 감지하여 모델 재학습 시점을 판단하는 근거로 활용할 수 있다.

▶ 오류 데이터의 격리 및 추적: 검증 실패 데이터를 즉시 격리하고 오류 사유를 기록해야 데이터 엔지니어가 원인을 파악하고 정제할 수 있으며, TTA 개발안내서에서도 데이터 품질 검증·관리 체계를 요구하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	검증 대상 및 검증 규칙 설계 완전성	검증 대상 테이블·컬럼이 식별되고, Null·형식·범위·참조·중복 등 검증 규칙이 구체적으로 설계되었는지 확인
2	검증 스크립트 명세 설계	검증 규칙별 SQL/스크립트 로직이 명세되었는지, 실행 순서·의존관계가 정의되었는지 확인
3	검증 배치 스케줄 설계	검증 실행 주기·시간대가 데이터 적재 주기 및 AI 학습 파이프라인과 연계 설계되었는지 확인
4	검증 결과 저장 테이블 설계	검증 실행 결과(통과/실패 건수, 오류 상세)를 저장하는 테이블이 설계되었는지 확인
5	오류 데이터 격리·처리 설계	검증 실패 데이터의 격리(Quarantine) 테이블 및 재처리·폐기 절차가 설계되었는지 확인
6	AI 학습 데이터 연계 및 KPI 임계값	검증 통과 데이터만 학습에 사용하는 필터링, 품질 KPI 임계값, 알림 기준이 설계되었는지 확인
7	EDB-벡터 DB 동기화 정합성 검증 배치 (AI 특화)	EDB 원문 수정·삭제 후 벡터 DB 임베딩 갱신 여부를 검증하는 배치(고아 임베딩·누락 임베딩·스탈(Stale) 임베딩 탐지 포함)가 설계되었는지 확인
8	벡터 DB 벡터 차원·모델 버전 일관성 검증 (AI 특화)	벡터 DB 컬렉션 내 임베딩 벡터의 차원수 일치성 및 생성 모델 버전 혼재 여부를 검증하는 배치가 설계되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	EDB 검증 규칙이 Null·형식·범위·참조·중복 등 5개 유형 이상 정의되고 스크립트 명세·스케줄·결과 테이블·격리 테이블·KPI 임계값이 모두 설계된 경우. 벡터 DB 사용 시 EDB-벡터 DB 교차 정합성 검증 배치(고아·누락·스탈 임베딩) 및 벡터 차원·모델 버전 일관성 검증까지 설계된 경우	현행 유지
보완권고 (△)	EDB 검증 규칙이 일부 정의되었으나 스크립트 명세 미흡, 결과 테이블 미설계, 격리 절차 누락, KPI 미정의 등 일부 미흡한 경우. 벡터 DB 사용 시 EDB-벡터 DB 교차 정합성 검증 배치가 일부 누락된 경우	누락된 검증 규칙 추가 및 스크립트 명세 구체화 등 해당 항목 보완 요청 후 재확인
부적합 (X)	EDB 데이터 정합성 검증 배치가 '전처리설계서'에 전혀 설계되지 않은 경우. 벡터 DB를 사용하면서 EDB-벡터 DB 교차 정합성 검증 배치가 어떤 산출물에도 정의되지 않은 경우	부적합 사항 시정조치(설계 보완) 요청 후 재점검

5. 감리 시 유의사항 및 실무 Tip

1) DB 제약조건만으로는 AI 데이터 품질을 보장할 수 없다: NOT NULL, CHECK, FK 등 DB 제약조건은 적재 시점의 단순 규칙만 검증하며, '코드 테이블 변경 후 고아 코드 발생', '정상 범위이나 AI 학습에 부적합한 이상치' 등은 배치 검증으로만 탐지 가능하므로 DB 제약조건과 별도로 배치 검증 설계가 존재하는지 확인한다.

2) EDB 검증과 벡터 DB 검증의 구조적 차이 인식: EDB의 배치 검증은 SQL로 실행하고 결과를 동일 DB에 저장할 수 있으나, 벡터 DB는 SQL이 없으므로 검증 스크립트는 Python(SDK 활용) 등으로 작성되며 검증 결과는 EDB에 저장해야 한다. DB 감리원은 '어떤 컬렉션의 어떤 속성을 검증하는가'라는 설계 명세 여부를 점검하며, 벡터 DB 검증 스크립트 코드 구현은 응용시스템 영역 점검 대상임에 유념한다.

3) 고아 임베딩이 RAG 품질에 미치는 영향 강조: EDB에서 원문이 삭제되었는데 벡터 DB의 임베딩이 함께 갱신되지 않으면 RAG 시스템이 삭제된 문서를 검색 결과로 반환하여 잘못된 정보를 제공하는 직접적 서비스 오류로 이어지므로, 단순 데이터 품질 문제를 넘어 서비스 신뢰성 위협으로 보고 EDB-벡터 DB 동기화 정합성 검증 배치 설계의 존재 여부를 반드시 확인한다.

4) 검증 결과 저장 없이 '실행만 하는 배치'는 무의미: 결과를 저장하지 않으면 '어제 검증 통과율이 몇 %였는가', '지난주 대비 오류가 증가했는가'를 확인할 수 없으므로, EDB와 벡터 DB 검증 결과 모두 시계열 분석이 가능하도록 EDB에 결과 테이블이 설계되었는지 반드시 확인한다.

5) 모호한 설계 표현의 구체화 유도: '벡터 DB 데이터를 주기적으로 검증한다'와 같은 모호한 표현은 분쟁의 원인이 되므로, '매일 02:00 EDB TB_DOCUMENT 테이블과 벡터 DB doc_collection 간 source_id 교차 검증을 실행하여, 고아 임베딩 탐지 시 1시간 내 삭제 처리하고 검증 결과를 VDB_VALIDATION_RESULT에 기록한다'와 같이 컬렉션·테이블·스케줄·처리 절차가 명시되도록 유도한다.

AG_SD42-3-05 | 개인정보 암호화 컬럼 설계점검

데이터베이스 | 설계 | 데이터 수집·처리 | 암호화

근거 | 「개인정보 보호법」 · KCMVP · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 개인정보가 포함된 컬럼에 대한 암호화 적용 대상 식별 및 암호화 알고리즘 지정이 {데이터베이스설계서}에 적정하게 설계되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 시스템에서는 학습 데이터, 추론 입력·출력, 사용자 피드백 등에 개인정보가 포함될 수 있으며, 일반 업무 시스템과 달리 AI 학습 과정에서의 암호화·복호화 처리, 비식별화와 암호화의 구분 적용, 벡터 DB·피처 스토어 등 AI 전용 저장소의 암호화 설계까지 고려해야 한다.

따라서 법적 암호화 의무 항목의 설계 반영뿐만 아니라 AI 데이터 파이프라인 전반에 걸친 암호화 설계의 적정성을 종합적으로 검토한다.

2. 필요성

▶ AI 서비스에서 개인정보 암호화는 법적 의무이자 AI 데이터 보호의 기본 요건이다. 개인정보 보호법에서는 고유식별정보, 비밀번호, 바이오정보 등에 대한 암호화를 의무화하고 있으며, AI 시스템은 대량의 개인 데이터를 수집·학습·추론에 활용하므로 암호화 설계의 중요성이 더욱 높다.

▶ 법적 암호화 의무 이행: 개인정보 보호법 제24조(고유식별정보), 제24조의2(주민등록번호), 제29조(안전성 확보조치)에 따라 고유식별정보·비밀번호·바이오정보 등은 DB 저장 시 암호화가 의무이다. 미이행 시 과태료 및 행정처분 대상이다.

▶ AI 학습 데이터 내 개인정보 보호: AI 학습용으로 수집된 데이터에 성명, 연락처, 주소 등 개인정보가 포함될 수 있다. 암호화 없이 저장하면 데이터 유출 시 대량 개인정보 침해가 발생한다.

▶ AI 전용 저장소의 암호화 사각지대 방지: 벡터 DB, 피처 스토어, 추론 로그 테이블 등 AI 전용 저장소에도 개인정보가 유입될 수 있으나, 일반 RDBMS 중심의 암호화 설계에서 이러한 저장소가 누락되는 경우가 빈번하다.

▶ 암호화와 AI 성능의 균형: 과도한 컬럼별 암호화는 AI 학습·추론 시 복호화 오버헤드를 유발한다. 법적 의무 항목과 추가 보호 항목을 구분하고, 암호화 방식(양방향/단방향)과 적용 시점을 적정하게 설계해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	암호화 대상 컬럼 식별	법적 의무 항목(고유식별정보, 비밀번호, 바이오정보) 및 AI 서비스 특성에 따른 추가 항목이 빠짐없이 식별되었는지 확인
2	암호화 알고리즘 지정 적정성	대상 컬럼별 암호화 알고리즘(AES, SHA, ARIA 등)과 방식(양방향/단방향)이 법적 기준에 부합하게 지정되었는지 확인
3	AI 전용 저장소 암호화 설계	벡터 DB, 피처 스토어, 추론 로그 등 AI 전용 저장소의 개인정보 암호화가 설계되었는지 확인
4	암호화 방식 (컬럼/TDE/API) 설계	암호화 적용 방식(컬럼 암호화, TDE, API 암호화)이 테이블·컬럼 특성에 맞게 선정되었는지 확인
5	암호화 키 관리 설계	암호화 키의 생성·저장·갱신·폐기 절차가 설계되었는지, 키 저장소가 DB와 분리 설계되었는지 확인
6	AI 학습 과정에서의 암호화·복호화 처리 설계	학습 데이터 추출 시 복호화 시점, 암호화 상태의 학습 불가 항목 처리, 비식별화와의 구분 적용이 설계되었는지 확인
7	임베딩 자산 암호화·가명화 설계 (AI 특화)	임베딩 자산에 대해 (1) 저장 시 암호화, (2) 전송 시 암호화, (3) 페이로드 내 식별자에 대한 토큰화/가명화가 함께 설계되었는지 확인. 3개 통제 중 2개 이상 설계 시 적합으로 판단
8	프롬프트·응답 로그 개인정보 보호 설계 (AI 특화)	프롬프트·응답 로그에 포함되는 개인정보의 암호화·접근통제·보존기간이 설계되었는지 확인. 3개 통제 모두 명시 시 적합으로 판단
9	다중 저장소 KMS·키 회전 정책 일관성 설계 (AI 특화)	KMS·키 회전 정책이 RDB 컬럼 암호화·Object Storage SSE·벡터 DB 암호화에 일관되게 적용되도록 설계되었는지 확인. 다중 저장소에 걸친 키 관리 일관성이 명시된 경우 적합으로 판단

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	법적 의무 항목이 모두 식별되고, 알고리즘·방식이 적정하게 지정되었으며, AI 전용 저장소·키 관리·학습 과정 처리까지 설계되고, 임베딩 자산의 저장·전송 암호화와 식별자 토큰화·가명화, 프롬프트·응답 로그의 개인정보 암호화·접근 통제·보존기간, 다중 저장소(RDB 컬럼 암호화·Object Storage SSE·벡터 DB)에 걸친 KMS·키 회전 정책의 일관성이 설계된 경우	암호화 대상 완전 식별, AES-256/SHA-256 지정, 벡터 DB 포함, KMS 분리 설계 확인 → 현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	RDBMS 암호화는 설계되었으나, AI 전용 저장소 누락, 키 관리 미설계, 학습 과정 복호화 절차 미명시 등 일부 미흡한 경우	벡터 DB·추론로그 누락, 키 갱신 미설계 등 → 해당 항목 보완 요청 후 재확인
부적합 (X)	법적 의무 암호화 항목이 식별되지 않았거나, 암호화 설계가 {데이터베이스설계서}에 전혀 반영되지 않은 경우	암호화 대상 미식별, 알고리즘 미지정, 암호화 설계 부재 → 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) RDBMS만 점검하고 AI 전용 저장소를 누락하지 않도록 주의: 가장 흔한 미흡 사례는 사용자 테이블(RDBMS)의 암호화만 설계하고, 벡터 DB·추론 로그·피처 스토어 등 AI 전용 저장소의 개인정보 암호화를 누락하는 것이다. 감리원은 {데이터베이스설계서}의 전체 저장소 목록에서 AI 전용 저장소의 암호화 포함 여부를 반드시 확인한다.

2) 암호화와 비식별화를 혼동하지 않도록 구분 확인: 사업자가 '비식별화 처리'를 '암호화'로 기재하거나 그 반대인 경우가 있다. 암호화는 키로 복원 가능, 비식별화는 복원 불가. 감리원은 법적 의무 항목에는 반드시 '암호화'가, AI 학습 활용 항목에는 필요에 따라 '비식별화'가 적용되었는지 구분하여 확인한다.

3) 암호화 대상의 테이블 간 일관성 확인: 동일한 주민등록번호가 사용자 테이블, 학습 데이터 테이블, 이력 테이블 등 복수 테이블에 존재할 수 있다. 한 테이블에서만 암호화하고 다른 테이블에서 평문으로 저장하면 의미가 없으므로, 동일 항목의 전 테이블 일관 적용 여부를 확인한다.

4) AI 학습 시 복호화 범위를 최소화하도록 권고: 학습 데이터 추출 시 전체 컬럼을 복호화하면 보안 위험이 증가한다. 암호화된 컬럼 중 학습에 불필요한 항목(주민번호 등)은 복호화 없이 제외하고, 학습에 필요한 항목만 선별적으로 처리하도록 설계되었는지 확인한다.

AG_SD42-3-06 | 테스트용 데이터관리 점검

데이터베이스 | 설계 | 데이터 수집·처리 | 테스트시나리오

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「개인정보 보호법」 · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 작성되는 {테스트계획서} 및 'AI 테스트 데이터 관리 설계서'를 대상으로, 기획/분석 단계에서 정의된 요구사항이 테스트 데이터 관리 설계에 빠짐없이 구체적으로 반영되었는지를 사업관리/품질보증 관점에서 검토하는 것을 목적으로 한다.

특히, AI 시스템의 성능과 신뢰성에 직접 영향을 주는 테스트 데이터셋의 편향(Bias) 방지 및 오염(Contamination) 방지 설계가 설계 산출물에 적절히 반영되었는지를 문서 완결성 검토 관점에서 확인한다.

2. 필요성

▶ AI 시스템은 일반 소프트웨어와 달리 테스트 데이터의 품질과 구성이 시스템 성능 및 공정성에 결정적 영향을 미치므로, 설계 단계에서 테스트 데이터 관리 체계가 충분히 설계되지 않으면 이후 구현·테스트 단계에서 다음과 같은 문제가 발생할 수 있다.

▶ 편향된 테스트 데이터로 인한 AI 성능 평가의 신뢰성 저하: 특정 집단이나 상황에 편중된 데이터로 구성된 테스트 셋은 AI 모델의 실제 오류를 검출하지 못할 수 있다.

▶ 테스트 데이터 오염(Contamination)으로 인한 검증 무효화: 학습 데이터와 테스트 데이터가 중복되거나 교차 오염될 경우, 모델 성능 측정값이 과도하게 부풀려져 실제 운영 성능과 괴리가 발생한다.

▶ 요구사항-설계 간 추적성 단절: 분석 단계에서 정의된 AI 품질 요구사항이 설계서에 반영되지 않으면, 이후 검사 단계에서 수용 기준 모호로 인한 분쟁이 발생할 수 있다.

▶ 개인정보 및 민감정보 포함 위험: 테스트 데이터 내 실제 개인정보가 사용될 경우, 법적 위반이 발생할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	테스트 데이터 수집·확보 계획 명시 여부	산출물 내 테스트 데이터 출처, 수량, 수집 절차 기술 내용 확인
2	편향 방지 설계 반영 여부	데이터 분포 다양성 설계 및 편향 점검 기준 포함 여부 확인

No	점검 항목	확인 방법
3	테스트 데이터 오염 방지 절차 명시 여부	학습-검증-테스트 데이터 분리 기준 및 절차 확인
4	테스트 데이터 내 민감정보 처리 방안	개인정보·민감정보 처리 방안(비식별화·가명처리 등) 기술 내용 확인
5	요구사항-테스트 추적성 반영 여부	요구사항 추적 매트릭스에 AI 품질 요구사항이 반영되었는지 확인
6	데이터 버전 관리 및 변경 이력 설계	테스트 데이터셋의 버전 관리 체계 기술 내용 확인
7	합성/가상 데이터 활용 시 품질 기준	합성 데이터 사용 여부 및 품질 기준 명시 여부 확인
8	데이터 라벨링 기준 및 품질 관리	라벨링 지침 및 일관성 검증 기준 확인
9	AI 평가셋 유형별 분리 정의 여부	'AI 테스트 데이터 관리 계획'에서 Golden Set·회귀셋·적대적 입력셋·환각 검출셋·Jailbreak 시도셋이 분리 정의되었는지 검토
10	평가셋의 학습 데이터 분리·격리 및 누출 방지 설계 여부	'데이터 격리 정책'에서 평가셋이 학습 데이터와 분리·격리되어 누출(Test Leakage) 방지를 위한 저장·접근 통제가 설계되었는지 검토
11	평가셋 버전 관리 및 재현 가능성 설계 여부	'평가 운영 설계'에서 평가 시점·모델 버전·인덱스 스냅샷 기록 등 평가셋 버전 관리와 재현 가능성이 설계되었는지 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	테스트 데이터 수집·확보 계획, 편향 방지 설계, 학습-테스트 데이터 분리·오염 방지 절차, 민감정보 처리 방안, 합성·가상 데이터 활용 시 품질 기준, 라벨링 지침과 일관성 검증 기준, 요구사항 추적성, 버전 관리 체계, 평가셋 분리 정의·격리·재현 가능성이 설계 산출물에 구체적으로 반영된 경우	현행 유지
보완권고 (△)	일부 항목이 반영되었으나 편향 허용 기준 미정의, 데이터 격리·접근 통제 설계 미흡, 버전 관리·변경 이력 체계 불명확, 합성 데이터 품질 기준 미명시 등 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	테스트 데이터 관리 설계가 산출물에 전혀 반영되지 않았거나, 테스트 데이터 내 개인정보 처리 요건 미준수 등 법적 위반 리스크가 확인된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (개인정보 보호법 준수 촉구)

5. 감리 시 유의사항 및 실무 Tip

1) 테스트 데이터 최신성 유지 기준 확인: 편향 방지를 위한 정적 비율(예: 성별 비율 5:5)도 중요하지만, AI 모델은 시간이 지남에 따라 현실 데이터의 변화(데이터 드리프트)를 겪는다. 테스트 데이터의 최신성 유지를 위한 주기적 업데이트 기준(예: 분기별 최신 데이터 10% 교체 테스트 등)이 설계되었는지 확인한다.

2) 테스트 데이터 누출(Leakage) 방지 설계 확인: 테스트 데이터가 학습 데이터로 흘러들어가 누출이 발생하면 평가 신뢰성이 무너진다. 격리·접근 통제·해시 기반 누출 탐지 정책이 설계되었는지 확인한다.

3) AI 테스트 데이터의 다층 구성 확인: AI 사업의 테스트 데이터는 단일 'QA 셋'이 아니라 (1) 회귀(Regression) 평가셋, (2) 적대적 입력(Adversarial) 셋, (3) Jailbreak/안전성 셋, (4) 환각 검출 셋, (5) 도메인 일반화 셋 등으로 다층 구성된다. 각 셋은 목적과 갱신 주기가 다르므로 단일 폴더/테이블이 아니라 별도 관리 기준이 설계되어 있는지 확인한다.

AG_SD42-3-07 | 공공데이터 연계·활용 요건 점검

데이터베이스 | 설계 | 데이터 수집·처리 | 학습데이터

근거 | 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」·「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)·「저작권법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 설계 단계에서, 공공데이터 포털·열린데이터·공공 API 등 외부 공공데이터를 연계·활용할 때 적법성·데이터 품질·API 안정성·정합성 요건이 설계에 적절히 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 '데이터관리계획서', '전처리설계서', {데이터베이스설계서}와 이용 약관·협약서를 근거로 ① 공공누리(KOGL) 유형별 활용 적합성과 출처 표시, ② 데이터 품질 기준, ③ 공공 API 호출 한도·가용성 대응, ④ 다기관 데이터의 정합성 확보 방안을 확인한다.

2. 필요성

▶ 공공데이터 활용은「공공데이터의 제공 및 이용 활성화에 관한 법률」과 공공누리(KOGL) 이용조건의 준수 의무를 수반한다. AI 학습에 활용할 수 있는 유형인지 확인하지 않으면 라이선스 위반이 된다.

▶ AI 학습·추론에 활용되는 공공데이터의 품질 저하는 모델 성능과 서비스 신뢰성에 직접 영향을 미친다. 결측·이상값·갱신 지연에 대한 품질 기준이 정의되어야 한다.

▶ 공공 API의 호출 한도(Rate Limit)와 가용성은 AI 서비스의 SLA에 직결된다. 한도 초과나 장애 시의 대체 경로가 설계되지 않으면 서비스 중단으로 이어진다.

▶ 여러 기관의 데이터를 통합할 때 코드 체계·기준 시점·항목 정의의 불일치로 정합성 문제가 발생한다. 통합 매핑 규칙과 검증 절차가 설계에 포함되어야 한다.

▶ 출처 표시 의무를 사람이 수동으로 이행하는 구조는 AI 응답에 적용할 수 없다. 응답 생성 시 출처가 자동으로 부기되도록 설계되어야 이용조건을 지속적으로 준수할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	공공누리(KOGL) 유형별 활용 적합성	'데이터관리계획서'에서 활용 대상 공공데이터의 공공누리 유형을 확인하고, AI 학습·상업적 이용이 허용되는 유형인지 이용 약관 원문과 대조
2	출처 표시 요건의 자동 이행 설계	AI 응답 생성 시 데이터 출처가 자동으로 부기되도록 설계되었는지 '전처리설계서' 또는 {응용시스템설계서}에서 확인

No	점검 항목	확인 방법
3	재배포·2차 가공 제한 준수	이용 약관의 재배포·2차 가공 제한 조건과 실제 활용 방식(벡터화·재배포 여부)이 일치하는지 대조하고, 해석이 모호한 경우 제공기관 질의 이력을 확인
4	유료·비공개 데이터의 별도 계약	유료 또는 비공개 데이터 활용 시 별도 협약서·계약서가 체결되어 '데이터관리계획서'에 첨부되었는지 확인
5	공공데이터 품질 기준 정의	결측률·이상값 허용 범위·갱신 주기 등 품질 기준이 수치로 정의되고, 기준 미달 시 처리 방안이 '전처리설계서'에 기술되었는지 확인
6	공공 API 호출 한도·가용성 대응 설계	API 호출 한도와 장애 시 캐시·대체 데이터원 활용 등 대응 설계가 반영되고 SLA 목표와 정합하는지 {데이터베이스설계서}에서 확인
7	다기관 데이터 통합 시 정합성 확보	코드 체계·기준 시점·항목 정의의 불일치를 해소하는 매핑 규칙과 정합성 검증 절차가 '전처리설계서'와 {데이터베이스설계서}에 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	라이선스·품질·API 안정성·정합성이 정량 기준으로 설계되고 출처 표시 자동화가 포함되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	일부 영역 검토는 수행했으나 데이터 정합성 또는 API 변경 대응이 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	라이선스·품질 검토가 부재하고 '공공데이터 활용' 선언 수준의 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) KOGL 단서 조항까지 검토: KOGL 유형 표시만 보고 AI 학습 활용 가능성을 단정하지 말고, 기관별 별도 약관과 데이터셋 상세 페이지의 단서 조항까지 검토해야 한다.

2) 데이터 활용 목록 표준화: 데이터 활용 목록은 '데이터셋명·기관·라이선스·갱신·용도·법무검토일'의 6열 표준으로 관리하도록 권고한다.

3) API 한도 초과 대비: 공공 API는 일일 호출 한도가 무료/유료 등급에 따라 달라지는 점을 반영하고, 자체 캐싱(TTL 1시간 등)으로 한도 초과 위험을 완화한다.

4) 코드 마스터 통합 관리: 법정동·산업분류 등 코드 마스터는 통합 마스터 테이블로 관리하고, 시점이 다른 데이터를 결합할 때 기준일 매핑으로 모순을 방지한다.

5) 출처 표시 자동화: 출처 표시는 응답 메타데이터에 포함해 UI에서 자동 렌더링되도록 하고, '공공데이터'로만 일괄 표기하지 않도록 구체적 데이터셋(원본 URL·발행기관·라이선스) 표기가 보존되는지 점검한다.

AG_SD42-3-08 | DB 성능 최적화 실행계획 설계점검

데이터베이스 | 설계 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 서비스 운영에 필요한 복잡한 쿼리(대량 데이터 조인, 집계, 학습 데이터 추출, 피처 생성 등)에 대한 실행계획(Execution Plan) 최적화 설계가 {데이터베이스설계서}에 적정하게 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 시스템은 수억 건 학습 데이터 추출, 실시간 피처 조합 쿼리, 다차원 성능 지표 집계, 대량 벡터 유사도 검색 등 일반 업무 시스템보다 쿼리 복잡도와 데이터 규모가 크므로, 단순 인덱스 설계를 넘어 쿼리 유형별 실행계획 최적화 전략, 비정규화 설계, 사전 집계(Pre-Aggregation), 뷰·머티리얼라이즈드 뷰 활용, 쿼리 힌트 전략 등 성능 최적화 설계의 적정성을 종합적으로 검토한다.

2. 필요성

▶ AI 서비스에서 DB 성능 최적화 설계는 학습 파이프라인 처리시간, 실시간 추론 응답속도, 모니터링 대시보드 응답성 등 서비스 전체의 성능에 직접적인 영향을 미친다. 설계 단계에서 최적화를 고려하지 않으면, 구현·운영 단계에서 대규모 재설계가 불가피해진다.

▶ 학습 데이터 추출 성능: 수억 건 테이블에서 조건별(기간, 카테고리, 라벨, 데이터셋 버전) 학습 데이터를 추출하는 복잡 쿼리가 수 시간 소요되면, 학습 파이프라인 전체 일정이 지연된다. 적절한 인덱스·파티셔닝·비정규화 설계가 필수적이다.

▶ 실시간 피처 조합 쿼리 성능: 추론 시 피처 스토어에서 여러 테이블을 조인하여 피처 벡터를 생성하는 쿼리가 수 초 이상 소요되면, AI 서비스의 응답시간 SLA를 위반한다.

▶ 모니터링 집계 쿼리 성능: 모델별·일별·사용자그룹별 성능 지표를 집계하는 다차원 쿼리가 대시보드에서 수 분 이상 소요되면, 운영 모니터링이 사실상 불가능해진다.

▶ 대량 배치 처리 성능: ETL, 정합성 검증, 데이터 이행 등 대량 데이터를 처리하는 배치 쿼리가 배치 윈도우를 초과하면 운영에 심각한 영향을 미친다.

3. 검토방법

No	점검 항목	확인 방법
1	주요 쿼리 식별 및 성능 목표 정의	AI 서비스의 핵심 쿼리 유형이 식별되고, 각 쿼리의 성능 목표(응답시간, 처리량)가 정의되었는지 확인

No	점검 항목	확인 방법
2	인덱스 최적화 설계	쿼리 접근 패턴에 맞는 인덱스(복합, 커버링, 부분, 함수 기반)가 설계되었는지 확인
3	비정규화·사전 집계 설계	조인 부하 감소를 위한 비정규화, 집계 쿼리 가속을 위한 사전 집계 테이블/MV가 설계되었는지 확인
4	파티셔닝·테이블스페이스 설계	대용량 테이블의 파티셔닝과 I/O 분산을 위한 테이블스페이스 배치가 설계되었는지 확인
5	쿼리 실행계획 분석 및 힌트 전략 설계	주요 쿼리의 실행계획(EXPLAIN) 분석 결과가 포함되고, 필요 시 쿼리 힌트 전략이 설계되었는지 확인
6	AI 특화 성능 설계	벡터 검색, 대량 배치 INSERT, 피처 조합 쿼리 등 AI 고유 쿼리의 성능 최적화가 설계되었는지 확인
7	벡터 검색 실행계획·튜닝 기록 보존 설계 (AI 특화)	벡터 검색의 실행 계획(쿼리 플랜)·튜닝 기록(인덱스 파라미터 변경·캐시 적중률) 보존 구조가 설계되었는지 확인. 벡터 DB 측 실행 계획·튜닝 이력 저장 구조가 명시된 경우 적합으로 판단
8	하이브리드 검색 단계별 후보 수·지연 측정 설계 (AI 특화)	하이브리드 검색(벡터+키워드+필터)에서 단계별 후보 수(top-k 단계, rerank top-n 단계)와 지연 측정 지점이 설계되었는지 확인. 단계별 측정 포인트가 정의된 경우 적합으로 판단
9	LLM 추론 지연 최적화 설계 (AI 특화)	LLM 추론 지연 최적화(스트리밍 응답, KV 캐시, 프롬프트 캐시) 적용 여부와 측정 방법이 설계되었는지 확인. 최적화 기법 선택 근거와 측정 방법이 명시된 경우 적합으로 판단

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	주요 쿼리가 식별되고, 인덱스·비정규화·사전집계·파티셔닝·실행계획 분석·AI 특화 최적화가 모두 설계된 경우	핵심 쿼리 식별·실행계획 분석·비정규화·MV·벡터 인덱스 설계 확인 → 현행 유지
보완권고 (△)	인덱스 설계는 있으나, 비정규화·사전 집계 미설계, 실행계획 분석 미포함, AI 특화 쿼리 최적화 누락 등 일부 미흡	PK 인덱스만 존재, 실행계획 미분석, MV 미설계 등 → 해당 항목 보완 요청 후 재확인
부적합 (X)	성능 최적화 설계가 {데이터베이스설계서}에 전혀 반영되지 않고, 성능 설계서도 미작성된 경우	최적화 설계 부재, 성능 목표 미정의 → 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '인덱스가 있다'와 '최적화되었다'는 다르다: PK 인덱스만 존재하고 실제 쿼리의 WHERE 조건에 맞는 복합 인덱스가 없으면, 인덱스가 무의미하다. 감리원은 인덱스 존재 여부뿐 아니라 '주요 쿼리의 접근 패턴과 인덱스가 일치하는가'를 확인해야 한다.

2) 실행계획(EXPLAIN) 분석 결과가 핵심 증빙: 최적화 설계의 적정성을 판단하는 가장 객관적인 근거는 실행계획 분석 결과이다. 감리원은 '성능 설계서'에 핵심 쿼리의 EXPLAIN 결과가 포함되어 있는지 확인하고, 미포함 시 추가를 권고한다. 단, 실행계획의 세부 해석은 구현 단계 감리에서 수행한다.

3) 사전 집계/MV 미설계가 가장 흔한 성능 설계 누락: 모니터링 대시보드에서 매번 원본 테이블을 집계하면 수 분 이상 소요되는 경우가 빈번하다. 일별·주별 사전 집계 테이블이나 MV가 설계되었는지 확인하고, 미설계 시 보완을 권고한다.

4) AI 특화 쿼리 최적화를 일반 쿼리와 구분 확인: 대량 범위 스캔(학습 추출), 대량 INSERT(로그 적재), 벡터 ANN 검색은 일반 업무 쿼리와 최적화 전략이 다르다. 감리원은 이러한 AI 고유 쿼리에 대한 별도 최적화 설계가 존재하는지 확인한다.

AG_SD42-3-09 | EDB-벡터 DB-하이브리드 데이터모델 정합성

데이터베이스 | 설계 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, EDB(Enterprise/관계형 DB, 이하 RDB)와 벡터 DB를 함께 운용하는 하이브리드 데이터 아키텍처가 AI 서비스의 요구사항을 충실히 반영하여 설계되었는지를 '감리원'이 제3자적 관점에서 검토하기 위한 것이다.

특히, RDB에 저장되는 정형 데이터(학습 이력, 운영 메타데이터, 추론 결과 등)와 벡터 DB에 저장되는 비정형 임베딩 데이터(문서 임베딩, 이미지 특징 벡터 등) 간의 설계 일관성, 동기화 정합성, 그리고 AI 생명주기 전반을 지원하는 데이터 구조의 완결성을 확인하는 것을 목적으로 한다.

2. 필요성

▶ AI 서비스(특히 RAG 기반 생성형 AI, Agent 시스템)는 정형 데이터를 관리하는 RDB와 비정형 임베딩 벡터를 관리하는 벡터 DB를 병행 운용하는 구조가 일반화되어 있다. 두 DB 시스템이 적절한 설계 없이 독립적으로 운영될 경우 다음과 같은 위험이 발생한다.

▶ 데이터 불일치(Inconsistency): RDB의 원문 데이터가 수정·삭제되어도 벡터 DB의 임베딩이 갱신되지 않으면, RAG 검색 결과가 실제와 다른 오래된 정보를 반환하게 된다. 이는 사용자에게 잘못된 정보를 제공하는 직접적 원인이 된다.

▶ 추적 불가(Traceability 결여): 벡터 DB의 임베딩 벡터와 RDB의 원문 데이터 간 연계 키(ID)가 설계 단계에서 정의되지 않으면, AI 추론 결과의 근거 문서를 역추적할 수 없어 설명가능성(Explainability)이 훼손된다. TTA 생성 AI 안내서 요구사항 04-2a에서도 데이터 흐름 및 계보의 추적 가능성 확보를 요구하고 있다.

▶ 편향·품질 관리 부재: 임베딩 데이터의 출처, 버전, 갱신 이력이 RDB에 메타데이터로 관리되지 않으면, 편향된 학습 데이터가 파이프라인에 유입되어도 탐지할 수 없다. NIST AI RMF는 데이터 품질 문제가 AI 신뢰성의 주요 위험 요인임을 명시하고 있다.

▶ 운영 산출물 누락: 하이브리드 DB 구조에 대한 설계 산출물이 불완전하면, 운영 단계에서 데이터 드리프트(Drift) 탐지 및 모델 재학습 파이프라인 구축이 어려워진다.

3. 검토방법

No	점검 항목	확인 방법
1	RDB AI 데이터 엔티티의 ERD 정의 여부	학습 데이터셋·추론 결과·메타데이터 등 RDB(EDB) 영역의 AI 데이터 엔티티가 {ERD}에 명확히 정의되었는지 확인
2	벡터 DB 컬렉션 구조·차원·스키마 명시 여부	벡터 DB의 컬렉션 구조·임베딩 벡터 차원수·메타데이터 스키마가 '벡터 DB 설계서'에 명시되었는지 확인
3	임베딩-RDB 연계 키 정의 여부	벡터 DB 임베딩 레코드와 RDB 원문 데이터 간 연계 키(source_id 등)가 'AI 데이터 아키텍처 설계서'·{ERD}에 정의되었는지 교차 검토
4	EDB-벡터 DB 동기화 정책 설계 여부	INSERT·UPDATE·DELETE 시 임베딩 갱신 절차 등 EDB-벡터 DB 간 데이터 동기화 정책이 '데이터 동기화 설계서'에 설계되었는지 확인
5	임베딩 모델 버전 메타데이터 관리 설계 여부	임베딩 생성에 사용된 AI 모델(모델명·버전)이 메타데이터로 RDB에 관리되는 설계가 {데이터베이스설계서}·'벡터 DB 설계서'에 반영되었는지 확인
6	청크-추론 결과 연계 구조 설계 여부	RAG 파이프라인에서 검색된 청크(Chunk) 데이터와 최종 추론 결과 간 연계 구조가 'AI 파이프라인 설계서'에 설계되었는지 확인
7	보호변수 속성의 RDB 메타데이터 정의 여부	성별·연령 등 데이터 편향 관리용 보호변수 속성이 RDB 메타데이터 엔티티에 정의되었는지 {ERD}·'AI 데이터 아키텍처 설계서'에서 확인
8	AI 데이터 요구사항 추적 가능 여부	{요구사항추적표}의 AI 데이터 요구사항이 RDB 엔티티·벡터 DB 컬렉션으로 추적 가능한지 설계 산출물과 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	RDB의 AI 데이터 엔티티와 벡터 DB 컬렉션 구조가 설계서에 명확히 정의되고, 양측 간 연계 키·동기화 정책·임베딩 모델 버전 메타데이터·청크-추론 결과 연계 구조와 편향 관리를 위한 보호변수 속성의 RDB 메타데이터 엔티티 정의가 모두 설계되어 {요구사항추적표}로 추적 가능한 경우	현행 유지
보완권고 (△)	RDB와 벡터 DB 설계는 각각 존재하나 연계 키 정의 불일치, 동기화 정책(UPDATE·DELETE 시 임베딩 갱신) 미정의, 임베딩 모델 버전 관리 또는 보호변수 속성 설계 누락 등 일부 미흡 사항이 존재하는 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	벡터 DB 관련 설계 산출물이 어디에도 존재하지 않거나, EDB-벡터 DB 간 연계·동기화 설계가 전혀 반영되지 않아 데이터 정합성과 추적 가능성을 보장할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 벡터 DB 산출물 부재 시 대응: AI 사업에서 벡터 DB를 사용하면서도 별도 설계서를 작성하지 않는 경우가 빈번하다. 이 경우 'AI 데이터 아키텍처 설계서' 또는 'AI 파이프라인 설계서'에 해당 내용이 포함되어 있는지 확인하고, 아무 산출물에도 없다면 설계 누락으로 판정한다. 사전에 사업자에게 벡터 DB 관련 설계 내용이 어느 산출물에 기술되어 있는지 확인해 두면 감리가 효율적이다.

2) 물리적 FK 부재에 대한 판단 기준: 대규모 AI 데이터 환경(NoSQL, 데이터 레이크, 벡터 DB 등)에서는 성능 이유로 물리적 FK를 설정하지 않는 것이 일반적이다. 따라서 물리적 FK 미설정 자체를 결함으로 판정하지 않도록 주의한다. 대신 '논리적 ERD상의 관계 명확성'과 'ETL·파이프라인 레벨에서의 정합성 유지 설계'를 집중 확인한다.

3) EDB-벡터 DB 연계 키의 명명 일관성 확인: 두 DB 간 연계 키(예: source_id, doc_id)의 명칭과 데이터 타입이 RDB 설계서와 벡터 DB 설계서에서 일치하는지를 반드시 교차 확인한다. 명칭 불일치는 연계 오류의 주요 원인이 된다.

4) 임베딩 모델 변경 리스크 인식: 임베딩 모델이 업그레이드되면 기존 벡터와 새 벡터가 호환되지 않아 전체 데이터를 재임베딩해야 한다. 이에 대한 관리 방안(모델 버전 관리 테이블, 재임베딩 파이프라인 설계)이 설계서에 포함되어 있는지 점검하고, 없는 경우 운영 리스크로 보완 권고한다.

5) RAG 아키텍처의 청크(Chunk) 설계 확인: RAG 패턴에서 문서를 청크 단위로 분할하여 임베딩하는 경우, 청크 크기(Chunk Size), 청크 겹침(Overlap) 정책, 청크 인덱스 관리 방법이 설계서에 정의되어 있는지 확인한다. 이 설계가 불명확하면 검색 품질이 저하되고 추론 결과의 정확성이 낮아진다.

AG_SD42-3-10 | 성능지표 데이터마트 설계점검

데이터베이스 | 설계 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 성능 지표(정확도, 재현율, F1-Score, 응답시간 등)를 산출·분석·모니터링하기 위한 데이터 마트(Data Mart) 스키마가 {데이터베이스설계서}에 적정하게 설계되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 배포 후에도 모델 성능이 시간에 따라 변동(드리프트)할 수 있으므로 성능 지표를 시계열로 축적하고 다차원 분석(모델별, 기간별, 데이터셋별, 사용자 그룹별)이 가능한 OLAP 구조의 데이터 마트가 반영되어야 하며, 단순 로그 테이블이 아닌 분석·집계에 최적화된 스타 스키마 또는 스노우플레이크 스키마 기반의 마트 설계 적정성을 종합적으로 검토한다.

2. 필요성

▶ AI 서비스에서 모델 성능 지표의 체계적 관리는 서비스 품질 유지와 모델 재학습 의사결정의 기반이며, 성능 지표를 단순 로그로만 저장하면 다차원 분석이 불가능하여 성능 저하 원인 파악, 모델 간 비교, 시계열 추이 분석 등 핵심 모니터링 활동을 수행할 수 없다.

▶ AI 모델 성능 드리프트 감지: 학습 시점의 성능과 운영 시점의 성능을 시계열로 비교하여 성능 저하(모델 드리프트)를 조기에 감지하려면 성능 지표가 시간 축을 포함한 마트 구조로 축적되어야 한다.

▶ 다차원 성능 분석 지원: 모델별, 버전별, 데이터셋별, 사용자 그룹별 성능을 비교 분석하려면 팩트 테이블과 차원 테이블로 구성된 OLAP 구조가 필요하며, 단일 로그 테이블로는 다차원 슬라이싱·다이싱이 불가능하다.

▶ 모델 재학습 의사결정 근거: 재학습 시점을 판단하려면 성능 지표의 추이(하락 경향, 임계값 이탈)를 분석해야 하며, 이를 위한 집계 테이블(일별, 주별, 월별 평균 성능)이 마트에 포함되어야 한다.

▶ AI 거버넌스·감사 대응: TTA 개발안내서와 NIST AI RMF MEASURE 기능에서는 AI 시스템의 성능을 지속적으로 측정·문서화할 것을 요구하고 있으며, 성능 지표 마트는 이러한 요구사항의 데이터베이스 설계적 구현이다.

3. 검토방법

No	점검 항목	확인 방법
1	마트 스키마 구조 설계	스타/스노우플레이크 스키마 기반으로 팩트·차원 테이블이 분리 설계되었는지 확인
2	성능 지표 컬럼 설계	AI 모델 성능 지표(정확도, 재현율, F1 등)가 팩트 테이블의 측정값 컬럼으로 설계되었는지 확인
3	차원 테이블 설계	모델, 버전, 데이터셋, 기간, 사용자그룹 등 분석 축이 차원 테이블로 설계되었는지 확인
4	시계열 축적 및 집계 테이블 설계	성능 지표가 시간 축으로 축적되고, 일별/주별/월별 집계 테이블이 설계되었는지 확인
5	원천 데이터 적재 설계	추론 로그 등 원천에서 마트로의 적재 주기·방식·변환 규칙이 설계되었는지 확인
6	편향·공정성 분석 구조 설계	보호 변수별(성별, 연령 등) 성능 분석이 가능하도록 마트에 편향 지표 관련 설계가 포함되었는지 확인
7	AI 특화 운영 지표 마트 구조 (AI 특화)	AI 특화 운영 지표(Recall@k, MRR, nDCG, 응답지연 p50/p95, 토큰 사용량, 환각/거부율, 도구 호출 성공률)를 저장·집계할 마트 구조가 설계되었는지 확인
8	RAG·LLM 품질 지표 구분 저장 구조 (AI 특화)	RAG 품질 지표(검색 적중률, 인용 정확도, 그라운딩 점수)와 LLM 품질 지표(유해도, 일관성, 안전성 위반율)를 구분 저장하는 구조인지 확인. 검색 지표·생성 지표가 구분 설계된 경우 적합
9	에이전트 운영 지표 차원 설계 (AI 특화)	에이전트 운영 지표(작업 성공률, 평균 단계 수, 도구 실패 분포)가 추가 마트 차원으로 설계되었는지 확인. 에이전트 운영 지표 3종 이상이 정의된 경우 적합

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	스타/스노우플레이크 스키마 기반 마트가 설계되고, 팩트·차원·집계 테이블이 구체적으로 정의되며, 편향 지표 반영과 적재 설계가 포함되고, AI 특화 운영 지표(Recall@k·MRR·nDCG, 응답지연 p50/p95, 토큰 사용량, 환각·거부율, 도구 호출 성공률) 저장·집계 구조와 RAG 품질 지표(검색 적중률·인용 정확도·그라운딩 점수) 및 LLM 품질 지표(유해도·일관성·안전성 위반율)의 구분 저장 구조, 에이전트 운영 지표(작업 성공률·평균 단계 수·도구 실패 분포) 차원이 설계된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	마트 테이블은 설계되었으나 차원 분리가 미흡하거나, 집계 테이블 누락, 편향 지표 미반영 등 OLAP 구조가 불완전한 경우(단일 테이블에 지표+차원 혼재, 집계 테이블 미설계, 보호 변수 차원 누락)	해당 항목 보완 요청 후 재확인
부적합 (X)	성능 지표 마트가 {데이터베이스설계서}에 전혀 설계되지 않았거나, 단순 로그 테이블만 존재하는 경우(마트 미설계, 성능 지표 저장 구조 부재, 로그 테이블만 존재)	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '로그 테이블'과 '마트'를 구분해야 한다: 추론 로그 테이블(원시 추론 결과 저장)은 OLTP 성격이고 성능 지표 마트(집계·분석용)는 OLAP 성격이므로, 사업자가 '로그 테이블에 성능 지표를 저장한다'고 기술한 경우 다차원 분석이 가능한 마트 구조인지 반드시 확인한다. 로그 테이블에서 직접 집계 쿼리를 실행하면 운영 DB 성능에 영향을 줄 수 있다.

2) AI 모델 유형에 따라 성능 지표가 다르다: 분류 모델(정확도, F1), 생성 모델(BLEU, 환각률), 추천 모델(nDCG, MAP) 등 모델 유형마다 핵심 지표가 다르므로, 해당 사업의 AI 모델 유형에 맞는 지표 컬럼이 팩트 테이블에 설계되었는지 확인하되 지표 산출 알고리즘 자체의 적합성은 응용시스템 영역에서 점검한다.

3) 집계 테이블 미설계는 가장 흔한 미흡 사례: 팩트 테이블만 설계하고 집계 테이블(일별/주별 평균)을 누락하는 경우가 빈번하며, 집계 없이 매번 원시 데이터에서 평균을 계산하면 대량 데이터에서 쿼리 성능이 급격히 저하되므로, 실시간 대시보드를 위한 일/주/월 단위 집계(Aggregation) 테이블 생성이 마트 설계에 포함되어 있는지 반드시 확인한다.

4) 편향 지표 분석 가능 구조는 AI 감리의 핵심 확인 사항: 전체 평균 정확도가 95%여도 특정 성별·연령대에서 70%일 수 있으므로, 보호 변수별 성능을 분석할 수 있는 차원 구조(사용자그룹 차원)가 마트에 설계되어 있는지 확인하는 것은 AI 감리 고유의 점검 사항이다.

5) 평가 기준선(베이스라인) 저장 구조 확인: AI 운영 지표의 평가 기준선 자체도 마트에 저장되어, 신규 모델 배포 시 회귀 여부 판단이 가능하도록 설계되었는지 확인한다.

AG_SD42-3-11 | 정제 전후 적재공간 분리점검

데이터베이스 | 설계 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「개인정보 보호법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 내용이 데이터베이스 설계 산출물에 빠짐없이 상세하고 구체적으로 반영되었는지 여부를 감리원이 제3자적 관점에서 검토하기 위한 것이다.

구체적으로, AI 시스템에서 사용하는 데이터를 정제 전 원천 데이터(Raw Data)와 정제·전처리 후 학습용 데이터(Training Data)로 구분하여, 이들의 적재 공간이 {데이터베이스설계서}에서 물리적(별도 스키마, 별도 데이터베이스, 별도 저장소) 또는 논리적(구분 컬럼, 뷰, 네이밍 규칙)으로 분리 설계되었는지를 점검하는 것을 목적으로 한다.

2. 필요성

▶ AI 시스템의 데이터는 원천 수집 상태(Raw), 전처리 완료 상태(Cleansed), 학습 투입 상태(Training-Ready) 등 여러 단계를 거치며, 각 단계의 데이터가 동일 공간에 혼재되면 다음과 같은 위험이 발생한다.

▶ 비정제 데이터의 학습 투입 위험: Raw Data와 Training Data가 동일 테이블이나 스키마에 혼재되면, 정제되지 않은 원천 데이터가 실수로 학습에 투입되어 모델 품질이 저하될 수 있다.

▶ 원본 데이터 훼손: 전처리 과정에서 Raw Data가 직접 변경·삭제되면 원본 복구가 불가능해진다. TTA 안내서 요구사항 05-1a는 정제 전과 후의 데이터 특성 설명을 요구하고 있으며, 이를 위해서는 정제 전 원본이 보존되어야 한다.

▶ 데이터 계보 추적 곤란: 분리된 적재 공간이 없으면, 학습에 사용된 데이터가 원천의 어떤 상태에서 어떤 정제를 거쳤는지 추적하기 어렵다. TTA 안내서 요구사항 04-2a(데이터 흐름 및 계보 추적)의 실효성이 떨어진다.

▶ 접근 권한 관리 곤란: Raw Data에는 개인정보·민감정보가 포함될 수 있으나, Training Data는 비식별화된 상태여야 한다. 동일 공간에 혼재 시 접근 권한을 차등 적용하기 어렵다.

3. 검토방법

No	점검 항목	확인 방법
1	Raw·Training 적재 공간 분리 전략 명시 여부	Raw Data(원천)와 Training Data(정제 후 학습용)의 적재 공간 분리 전략이 {데이터베이스설계서} 아키텍처 섹션에 명시되었는지 확인

No	점검 항목	확인 방법
2	Raw·Training 영역 스키마·네이밍 구분 여부	Raw 영역과 Training 영역 테이블이 별도 스키마 또는 네이밍 규칙으로 구분되었는지 {테이블정의서} 테이블 목록·스키마에서 확인
3	ERD 상 영역 시각 구분 여부	Raw 영역과 Training 영역이 시각적으로 구분된 서브젝트 영역(Subject Area)으로 {ERD}에 표현되었는지 확인
4	Raw→Training 데이터 이동 경로 정의 여부	Raw Data에서 Training Data로의 데이터 이동(ETL/ELT) 경로가 {데이터베이스설계서} 또는 'AI 파이프라인 설계서'에 정의되었는지 교차 검토
5	Raw·Training 접근 권한 차등 설계 여부	Raw Data와 Training Data에 대한 접근 권한이 차등 설계되었는지 {데이터베이스설계서} 접근 제어 섹션에서 확인
6	데이터 분리·보존 요구사항 추적 가능 여부	{요구사항추적표}의 데이터 분리·보존 요구사항이 {데이터베이스설계서}의 구체적 설계로 추적 가능한지 매핑 검토
7	4계층(원본·정제·임베딩·색인) 영역 분리 명시 여부	원본·정제·임베딩·색인 4계층 영역 분리가 RDB 스키마 또는 Object Storage 버킷·경로 정책으로 'AI 데이터 아키텍처 설계서'·스토리지 명명 규칙에 명시되었는지 확인
8	격리(Quarantine) 영역 설계 여부	이상·유해·중독 의심 데이터가 학습·임베딩 흐름에서 배제되도록 격리(Quarantine) 영역이 별도로 정의·설계되었는지 '데이터 분리 보관 설계'에서 확인
9	임시·실험·운영 영역 분리 설계 여부	임시(Temporary)·실험(Sandbox) 영역과 운영(Production) 영역이 분리되어 실험성 임베딩·모델이 운영 RAG에 유입되지 않도록 설계되었는지 스토리지·DB 영역 분리 정책에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	Raw Data와 Training Data의 적재 공간 분리 전략(물리적 또는 논리적)이 {데이터베이스설계서}에 명시되고, {테이블정의서}·{ERD}에 영역 구분이 일관되게 반영되며, 원본(Raw)·정제(Cleansed)·임베딩(Embedded)·색인(Indexed) 4계층 영역 분리가 스키마 또는 스토리지 경로 정책으로 명시되고, 데이터 이동(ETL/ELT) 경로와 접근 권한 차등 설계, 격리·운영/비운영 영역 분리까지 확인되는 경우	현행 유지
보완권고 (△)	분리 전략은 정의되어 있으나 일부 산출물({ERD}의 서브젝트 영역 구분, 격리 영역 정의, 접근 권한 차등, 요구사항 추적 등)에 반영이 미흡하거나 분리 수준이 불명확한 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	Raw Data와 Training Data의 적재 공간 분리가 물리적·논리적으로 전혀 설계되지 않았거나, 관련 설계 산출물이 부재하여 비정제 데이터의 학습 투입·원본 훼손 위험이 방치된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) Data Lake vs. Data Warehouse 구분: AI 사업에서 Raw Data는 Data Lake(비정형 포함, 스키마리스)에, Training Data는 Data Warehouse 또는 Feature Store(정형, 스키마 기반)에 적재하는 것이 일반적이다. {데이터베이스설계서}에 이 구분이 명시되어 있는지 확인하되, 저장소 인프라의 기술적 세부 구현 방식은 사업자의 재량 범위이므로 분리 구조의 명시 여부를 중심으로 점검한다.

2) 소규모 프로젝트의 최소 분리 수준: 예산·인프라 제약으로 별도 DB/스키마 구성이 어려운 경우, 최소한 테이블 네이밍 접두어(RAW_, TRAIN_)와 구분 컬럼(DATA_STAGE_CD)을 통한 논리적 분리가 설계되어야 한다. 분리가 전혀 없는 설계는 프로젝트 규모와 무관하게 부적합으로 판정한다.

3) 생성형 AI 특화 적재 구조: LLM 기반 서비스의 경우, 원천 텍스트(Raw), 청킹·전처리된 텍스트(Cleansed), 벡터 임베딩(Training-Ready)이 서로 다른 저장소 유형(RDB, 문서DB, 벡터 DB)에 적재될 수 있다. {데이터베이스설계서}에 이러한 이기종 저장소 간 데이터 흐름이 정의되었는지 확인한다.

4) Raw Data 보관 기간(Retention) 확인: Raw Data를 무기한 보관하면 스토리지 비용이 증가하고, 너무 짧게 보관하면 원본 복원이 불가능해진다. {데이터베이스설계서}에 Raw Data 보관 기간과 폐기 정책이 명시되었는지 확인한다.

5) 운영 산출물 연계: 적재 공간 분리 구조를 기반으로 한 '데이터 적재 현황 모니터링 대시보드 설계서', '스토리지 용량 관리 계획서' 등 운영 산출물 작성 계획이 수립되었는지 확인한다. 운영 과정 자체의 점검사항은 본 가이드라인 범위 밖이나, 운영을 고려한 산출물 계획의 적정성은 점검 대상이다.

AG_SD43-3-01 | 편견금칙어 사전 적용점검

데이터베이스 | 구현 | 데이터 수집·처리 | 금칙어/가드레일

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 편견 표현 사전 및 금칙어(Blacklist) 사전이 데이터베이스에 구축·관리되고, 사전 DB 업데이트 시 시스템의 필터링 기능에 즉각 반영되어 AI 서비스의 입출력에서 해당 표현이 차단되는지를 {시스템화면}과 관련 시험 결과를 중심으로 확인하고, 사전 관리 체계의 구현·동작 적정성을 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

편견/금칙어 사전은 AI 서비스가 유해·차별적·부적절한 표현을 생성하거나 처리하는 것을 방지하는 핵심 안전장치이다.

사전의 DB 관리 체계, 실시간 반영 메커니즘, 필터링 동작의 정확성을 종합적으로 검증한다.

2. 필요성

▶ 생성형 AI를 포함한 AI 서비스는 편견(Bias)이 포함된 표현이나 금칙어(욕설, 혐오 표현, 차별 표현, 불법 정보 등)를 생성·출력할 위험이 있다. 편견/금칙어 사전은 이러한 위험을 기술적으로 통제하는 1차 방어선이며, 사전이 DB에 체계적으로 관리되고 업데이트가 즉각 반영되어야 실질적 보호가 가능하다.

▶ AI 안전성(Safety) 확보: 금칙어 사전이 적용되지 않으면, AI 서비스가 욕설·혐오 표현·차별적 발언·불법 정보 등을 출력하여 사용자 피해와 사회적 물의를 초래할 수 있다. TTA 개발안내서는 금칙어 필터링을 AI 안전성의 핵심 요건으로 정의한다.

▶ 편견(Bias) 방지: 성별·인종·연령·장애 등에 대한 편견 표현이 AI 응답에 포함되면 공정성이 훼손된다. 편견 표현 사전을 통해 학습 데이터 및 AI 출력에서 편견 표현을 탐지·차단해야 한다.

▶ 실시간 반영의 중요성: 새로운 유해 표현은 지속적으로 출현하므로, 사전에 신규 단어를 추가하면 시스템 재시작 없이 즉각 필터링에 반영되어야 한다. DB 업데이트와 필터링 반영 간 시간 차이가 클수록 유해 표현 노출 위험이 증가한다.

▶ NIST AI RMF 공정성 요건: NIST AI RMF는 유해 편향 관리(Fair – with Harmful Bias Managed)를 AI 신뢰성의 핵심 특성으로 정의하며, 편견 탐지·완화를 위한 기술적 조치를 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	편견/금칙어 사전 DB 구축 여부	테이블 구조·분류 체계·단어 수 등 편견/금칙어 사전 DB가 '사전 설계서'대로 구축되었는지 실제 DB 구조와 대조
2	등록 금칙어 입력·출력 차단 여부	사전에 등록된 금칙어가 AI 서비스 입력·출력에서 실제로 차단되는지 {시스템화면}에서 직접 확인
3	사전 업데이트 필터링 즉각 반영 여부	사전 DB 업데이트(신규 단어 추가) 시 필터링에 즉각 또는 설계된 주기 내 반영되는지 {테스트결과서} 업데이트 반영 시험 결과에서 확인
4	사전 관리 화면 구현 여부	추가·수정·삭제·검색 등 사전 관리 화면이 구현되었는지 {시스템화면}에서 직접 확인
5	사전 변경 이력 DB 기록·관리 여부	추가·수정·삭제 일시·변경자·변경 사유 등 사전 변경 이력이 '데이터베이스 설계서' 이력 테이블에 기록·관리되는지 확인
6	금칙어 필터링 오차단·미탐지율 목표 이내 여부	금칙어 필터링의 오차단(정상 표현 차단)·미탐지(금칙어 미차단)율이 목표 이내인지 'AI 안전성 시험 보고서' 차단율·오차단율에서 확인
7	금칙어/편견 요구사항 추적 가능 여부	{요구사항추적표}의 금칙어/편견 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	편견/금칙어 사전 DB가 설계서대로 구축되고, 입력·출력 필터링 및 사전 업데이트의 즉각 반영이 시험으로 확인되며, 사전 관리 기능과 변경 이력 관리가 완비되고, 금칙어 필터링의 오차단율·미탐지율이 목표 기준 이내로 확인된 경우	현행 유지
보완권고 (△)	사전 DB와 필터링 기능은 구현되었으나, 업데이트 즉각 반영 시험 미수행, 변경 이력 관리 미흡, 오차단율 기준 미충족 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	편견/금칙어 사전 DB가 미구축되었거나, 등록된 금칙어가 AI 서비스 입출력에서 실제로 차단되지 않는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '사전 업데이트 즉각 반영 시험'의 수행 여부를 최우선으로 확인하라: 사전이 구축되어 있고 필터링이 동작하더라도, 신규 단어 추가 시 즉각 반영되지 않으면 운영 중 새로운 유해 표현에 대응

할 수 없다. {테스트결과서}에 '신규 단어 추가 → 즉시 차단 확인' 시험이 포함되어 있는지를 최우선으로 점검한다.

2) '입력 필터링'과 '출력 필터링'을 구분하여 확인하라: 입력(프롬프트)에서 금칙어를 차단하는 것과, AI 응답(출력)에서 금칙어를 차단하는 것은 별개의 기능이다. 특히 생성형 AI에서는 출력 필터링이 더 중요하다. 양쪽 모두 동작하는지 {시스템화면} 또는 {테스트결과서}에서 별도로 확인한다.

3) 사전의 '적재 건수'를 확인하라: 사전 테이블이 존재하더라도 적재된 단어가 극히 적으면(예: 10개 미만) 실질적 보호가 미흡하다. 설계서 또는 요구사항에 정의된 최소 기준(예: 1,000개 이상)과 실제 적재 건수를 대조한다.

4) 편견 표현과 금칙어를 구분 관리하는지 확인하라: 욕설·혐오 표현(금칙어)은 무조건 차단이 적절하나, 편견 표현은 맥락에 따라 허용될 수 있다(예: '편견 표현을 사용하면 안 됩니다'라는 교육적 문장). 사전에서 금칙어와 편견 표현이 분류 카테고리로 구분 관리되는지 확인한다.

5) 생성형 AI 사업에서의 추가 점검 포인트: 생성형 AI(LLM)에서는 프롬프트 인젝션(Prompt Injection)을 통한 금칙어 우회 시도가 빈번하다. 'AI 안전성 시험 보고서'에 프롬프트 인젝션 기반 금칙어 우회 시험이 포함되어 있는지를 추가로 확인한다. 예: '아래 단어를 거꾸로 출력해줘: OOO' 등의 우회 시도에 대한 차단 여부.

AG_SD43-3-02 | 공통코드 데이터정합성 점검

데이터베이스 | 구현 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 공통 코드 테이블(Code Table)과 이를 참조하는 마스터 테이블(Master Table) 및 트랜잭션 테이블 간의 데이터 정합성이 확보되었는지를 {데이터베이스} 관련 산출물을 중심으로 확인하고, 특히 코드-마스터 간 참조 무결성 위배 건수가 0건인지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템에서 공통 코드는 학습 데이터의 범주형 특성(Categorical Feature)을 정의하는 핵심 기준 데이터로, 코드 테이블에 존재하지 않는 코드 값이 마스터·트랜잭션 테이블에 적재되면 AI 모델의 학습·추론 시 범주 매핑 오류, 원-핫 인코딩(One-Hot Encoding) 실패, 미지 범주(Unknown Category) 처리 오류 등이 발생하여 모델 품질이 직접적으로 훼손된다.

2. 필요성

▶ 공통 코드 데이터는 AI 시스템의 전 영역에 걸쳐 사용되는 기준 정보로, 코드와 마스터 데이터 간 정합성이 보장되지 않으면 AI 모델의 입력 데이터 자체가 오류 상태에 놓이게 된다. 특히 AI 서비스에서는 코드 값이 학습 특성(Feature)으로 직접 활용되므로, 일반 정보시스템 이상으로 엄격한 코드 정합성 관리가 요구된다.

▶ AI 학습 데이터 품질 보증: 코드 테이블에 없는 코드 값(고아 코드, Orphan Code)이 학습 데이터에 포함되면, 전처리 시 인코딩 실패 또는 미지 범주 처리로 인해 학습 품질이 저하된다. 코드-마스터 정합성은 학습 데이터 품질의 기본 전제이다.

▶ 추론 결과 신뢰성 확보: 운영 중 신규 코드가 코드 테이블에 추가되지 않은 채 트랜잭션에 사용되면, 모델이 해당 코드를 해석하지 못하여 추론 오류나 서비스 장애가 발생할 수 있다.

▶ 데이터 무결성 기본 원칙 준수: 참조 무결성(Referential Integrity)은 데이터베이스 설계의 가장 기본적인 원칙이며, 정보시스템 감리 수행 가이드에서도 핵심 상세 검토 항목으로 정의하고 있다.

▶ 코드 체계 표준화 검증: AI 시스템이 레거시 시스템과 연계되는 경우, 양 시스템 간 코드 체계의 일관성이 보장되어야 데이터 교환·이행 시 코드 매핑 오류가 방지된다.

3. 검토방법

No	점검 항목	확인 방법
1	코드-마스터 참조 무결성 위배 0건 여부	코드 테이블과 마스터 테이블 간 참조 무결성 위배 건수가 0건인지 {테스트결과서} 또는 '데이터 품질 검증 보고서'의 무결성 검증 결과에서 확인
2	공통 코드 체계 DB 구현 여부	'공통 코드 정의서'에 정의된 코드 체계(그룹 코드·상세 코드)가 {데이터베이스} 코드 테이블에 실제 구현되었는지 대조
3	고아 코드 부재 여부	마스터 테이블 코드 컬럼에 코드 테이블에 없는 값(고아 코드)이 존재하지 않는지 '데이터 품질 검증 보고서'의 고아 코드 검출 결과에서 확인
4	미사용 코드 관리 방안 정의 여부	미사용 코드(Dead Code)의 관리 방안이 '공통 코드 정의서'의 코드 상태 관리(활성/비활성) 항목에 정의되었는지 확인
5	코드-마스터 참조 관계 명시 여부	코드-마스터 간 참조 관계(FK 등)가 '데이터베이스 설계서' ERD에 명시되었는지 확인
6	AI 학습 범주형 특성 코드 값 일치 여부	AI 학습에 사용되는 범주형 특성(Feature) 컬럼의 코드 값이 코드 테이블과 100% 일치하는지 {테스트결과서}의 학습 데이터 코드 검증 결과에서 확인
7	코드 정합성 요구사항 추적 가능 여부	{요구사항추적표}의 코드 정합성 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	전체 코드 참조 관계에서 무결성 위배 및 고아 코드가 0건으로 검증 범위와 함께 확인되고, '공통 코드 정의서'의 코드 체계가 DB에 일치 구현되었으며, AI 학습 데이터의 범주형 특성 코드 값 정합성과 미사용 코드 관리 방안까지 확인되는 경우	현행 유지
보완권고 (△)	무결성 검증은 수행되었으나 검증 범위가 일부 테이블에 한정되거나, AI 학습 전용 테이블 검증·미사용 코드 관리 방안·코드 참조 관계의 설계서 명시 등 일부 항목이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	코드-마스터 간 참조 무결성 위배 또는 고아 코드가 존재하거나, 코드 정합성 검증 자체가 수행되지 않았고 코드 체계 정의도 부재한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '위배 0건'의 검증 범위를 반드시 확인하라: {테스트결과서}에 '코드 무결성 위배 0건'으로 기재되어 있더라도, 검증 대상 범위(어떤 테이블의 어떤 컬럼을 검증했는지)가 명시되어 있는지 확인한다. 일부 테이블만 검증하고 전체 0건으로 보고하는 사례가 있으므로, 검증 범위의 완전성을 점검하는 것이 중요하다.

2) AI '학습 데이터 테이블'의 코드 정합성을 별도로 확인하라: 일반 업무 테이블의 코드 정합성은 확보되었으나, AI 학습 전용 테이블(학습_데이터셋, Feature Store 등)은 별도 적재 파이프라인을 거치므로 코드 정합성이 누락되는 사례가 빈번하다. AI 학습 테이블에 대한 별도 검증이 수행되었는지 반드시 확인한다.

3) FK 미설정 사유가 합리적인지 면담으로 확인하라: AI 사업에서는 대량 데이터 성능 또는 NoSQL 사용을 이유로 FK를 생략하는 경우가 많다. 이 경우 FK 미설정 자체를 부적합으로 판정하지 말고, 애플리케이션 레벨의 대안(적재 전 코드 검증, 주기적 배치 검증 등)이 구현되었는지를 확인한다.

4) 코드 변경이 AI 모델에 미치는 영향 관리 체계를 확인하라: 운영 중 신규 코드가 추가되면 AI 모델의 범주 인코딩에 '미지 범주(Unknown Category)'가 발생한다. 코드 변경 시 모델 재학습 필요 여부를 판단하는 절차가 수립되었는지 면담으로 확인한다.

5) '공통 코드 정의서'가 없는 경우의 대응: 별도의 '공통 코드 정의서'가 작성되지 않은 사업의 경우, '데이터베이스 설계서'의 코드 테이블 설계 항목이나 '데이터 표준 정의서'에서 코드 체계를 확인할 수 있다. 코드 체계 정의 자체가 부재한 경우, 보완 권고 대상이다.

AG_SD43-3-03 | 데이터계보 이력저장 확인점검

데이터베이스 | 구현 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습 데이터를 수집·변환·적재하는 ETL/전처리 파이프라인의 작업 이력이 {시스템로그}(로그 테이블, 이력 테이블 등)에 설계대로 정상 적재되었는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

특히, 각 파이프라인 단계별 수행 일시, 입력/출력 건수, 적용된 변환 규칙 버전, 성공/실패 상태, 오류 내역 등의 이력 정보가 체계적으로 기록되어, AI 모델 학습에 사용된 데이터가 '어떤 원본에서, 어떤 처리를 거쳐, 최종 훈련 데이터가 되었는지'를 역추적할 수 있는 데이터 계보(Data Lineage) 체계가 물리적으로 구현되었는지를 확인한다.

2. 필요성

▶ AI 시스템에서 데이터 계보(Data Lineage) 이력의 저장은 단순한 로그 관리가 아니라, AI 모델의 신뢰성·재현성·감사 가능성을 보장하는 핵심 인프라이다. ETL/전처리 파이프라인 이력이 적재되지 않으면 다음과 같은 문제가 발생한다.

▶ AI 모델 재현성(Reproducibility) 상실: 학습 데이터가 어떤 전처리를 거쳤는지 이력이 없으면, 동일한 조건으로 모델을 재학습하거나 결과를 재현하는 것이 불가능하다. 이는 AI 시스템의 신뢰성 검증을 근본적으로 저해한다.

▶ 데이터 품질 문제 원인 역추적 불가: 모델 성능 저하 시, 학습 데이터의 어떤 전처리 단계에서 문제가 발생했는지 파악하려면 파이프라인 이력이 필수적이다. 이력 없이는 '데이터가 어디서 오염되었는지' 특정할 수 없다.

▶ 감사(Audit) 및 규제 대응 불가: TTA 신뢰할 수 있는 인공지능 개발안내서와 NIST AI RMF는 데이터 처리 이력의 기록과 추적을 요구한다. 파이프라인 이력이 미적재되면 AI 감사 시 데이터 처리 과정에 대한 증빙이 불가능하다.

▶ 운영 안정성 저해: ETL 작업의 성공/실패 이력, 오류 건수, 처리 소요 시간 등이 기록되지 않으면, 파이프라인 장애 발생 시 신속한 원인 분석과 복구가 어려워진다.

3. 검토방법

No	점검 항목	확인 방법
1	이력 테이블(로그 테이블) 생성 확인	설계서에 정의된 파이프라인 이력 테이블이 {시스템로그}로서 물리 DB에 생성되었는지, 테이블 구조가 설계와 일치하는지 확인
2	파이프라인 단계별 이력 적재 완전성	설계된 모든 ETL/전처리 단계(수집, 정제, 변환, 적재 등)에 대한 이력이 빠짐없이 적재되었는지 확인
3	필수 이력 항목 적재 확인	각 파이프라인 이력 레코드에 필수 항목(작업ID, 수행일시, 단계명, 입력건수, 출력건수, 성공/실패, 변환규칙 버전 등)이 모두 기록되었는지 확인
4	입출력 건수 정합성	이력에 기록된 입력 건수와 출력 건수가 실제 데이터 건수와 일치하는지, 단계 간 건수 체인이 정합한지 확인
5	오류/실패 이력 기록 확인	파이프라인 수행 중 발생한 오류 및 실패 건에 대한 이력(오류 코드, 오류 메시지, 실패 건수)이 기록되었는지 확인
6	이력 데이터 보존 정책 확인	파이프라인 이력의 보존 기간, 아카이빙 정책이 설계되었는지 확인
7	시험 결과 기반 이력 적재 검증	파이프라인 이력 적재 관련 시험이 수행되었는지, 시험 결과가 적합한지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	이력 테이블이 설계대로 생성되고, 모든 파이프라인 단계의 이력이 빠짐없이 적재되며, 필수 항목 NULL 0건, 입출력 건수 체인이 정합하고, 오류·실패 건의 이력(오류 코드·오류 메시지·실패 건수)이 기록되며, 이력 데이터의 보존 기간·아카이빙 정책이 설계되고 이력 적재 관련 시험이 수행되어 적합 판정된 경우	예: 이력 테이블 전수 생성, 파이프라인 전 단계 이력 적재, NULL 0건, 건수 체인 정합 — 현행 유지
보완권고 (△)	이력 테이블은 생성되었으나, 일부 단계의 이력 누락, 일부 필수 항목 NULL, 또는 건수 체인 불일치가 소수 발견된 경우	예: 특성추출 단계 이력 미적재, 변환규칙 버전 NULL, 건수 체인 불일치 등 — 해당 항목 보완 요청 후 재확인
부적합 (X)	이력 테이블 자체가 미생성이거나 파이프라인 이력이 전혀 적재되지 않아 데이터 계보 추적이 불가능한 경우	예: 이력 테이블 미생성, 파이프라인 로그 미존재 — 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 파이프라인 도구의 내장 이력 저장 기능을 확인하라: Apache Airflow, Luigi, AWS Glue, Azure Data Factory 등의 파이프라인 도구는 작업 이력을 자체 메타데이터 DB에 자동 저장한다. 이 경우 별도 이력 테이블을 설계하지 않더라도 도구 내장 이력이 존재할 수 있으므로, 이력의 소재(자체 설계 테이블 vs 도구 메타데이터 DB)를 먼저 확인한다.

2) 입출력 건수 체인의 '끊김'을 핵심 지표로 삼으라: 데이터 계보의 신뢰성은 단계 간 건수 체인의 정합성으로 판단할 수 있다. 이전 단계 출력 건수와 다음 단계 입력 건수가 일치하지 않는 곳이 있다면 데이터 유실 또는 비정상 처리가 발생한 것이므로, 건수 체인의 끊김을 중점적으로 점검한다.

3) 변환 규칙 버전 기록의 존재 여부를 반드시 확인하라: AI 전처리에서는 결측값 처리, 정규화, 인코딩 등의 변환 규칙이 버전별로 달라질 수 있다. 이력에 '어떤 버전의 변환 규칙을 적용했는지'가 기록되어 있어야 모델 재현이 가능하다. 변환 규칙 버전 항목이 이력에 누락되었거나 NULL인 경우, 데이터 계보의 완전성이 훼손된 것이므로 보완을 요구한다.

4) 실패 이력의 '상세도'를 평가하라: 파이프라인 이력에 '실패'라고만 기록되고 오류 코드, 오류 메시지, 실패 건수가 없다면 장애 원인 분석이 불가능하다. 실패 이력은 오류 분류 코드, 구체적 오류 메시지, 실패 데이터 건수, 재수행 여부까지 포함해야 실용적이다. 실패 이력의 상세도를 직접 확인하고 불충분한 경우 보완을 권고한다.

5) 이력 보존 기간이 모델 생명주기 이상인지 확인하라: AI 모델은 배포 후 수개월~수년간 운영되며, 이 기간 동안 학습 데이터의 처리 이력을 역추적할 수 있어야 한다. 이력 보존 기간이 모델 운영 예상 기간보다 짧으면 추후 감사나 재학습 시 이력이 소실된다. 보존 기간이 최소 모델 생명주기 이상으로 설계되었는지 확인한다.

AG_SD43-3-04 | 레거시데이터 이행 ETL 검증 점검

데이터베이스	구현	데이터 수집·처리	정제(클렌징)/변환

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 기존 레거시 시스템의 데이터를 신규 AI 시스템 데이터베이스로 이행(Migration)하는 ETL(Extract-Transform-Load) 작업이 완료되었는지를 {데이터이행결과}를 중심으로 확인하고, 특히 원본(Source)과 타겟(Target) 간 건수 일치 및 오류 데이터 0건 달성 여부를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 레거시 데이터를 학습·추론의 기초 자산으로 활용하므로, 데이터 이행 과정에서 건수 불일치, 값 왜곡, 코드 변환 오류 등이 발생하면 AI 모델의 학습 품질과 서비스 신뢰성에 직접적인 영향을 미친다.

이에 따라 이행 완전성(Completeness)과 정합성(Consistency)을 정량적으로 검증한다.

2. 필요성

▶ 레거시 데이터 이행은 AI 서비스 구축에서 가장 기본적이면서도 오류 발생 빈도가 높은 작업이다. 이행 과정에서의 데이터 손실·왜곡은 AI 학습 데이터의 근본적 품질을 훼손하며, 이행 후 발견되지 않은 오류는 모델 성능 저하의 원인으로 장기간 잠복할 수 있다.

▶ 데이터 완전성(Completeness) 보장: 원본 시스템의 전체 데이터가 타겟 시스템으로 누락 없이 이행되었는지 확인해야 한다. 건수 불일치가 발생하면 학습 데이터의 대표성이 훼손되어 모델 편향이 발생할 수 있다.

▶ 데이터 정합성(Consistency) 확보: 이행 과정의 데이터 타입 변환, 코드 매핑, 인코딩 처리에서 값 왜곡이 발생하지 않았는지 검증해야 한다. 특히 AI 학습에 직접 사용되는 수치형 데이터의 정밀도 손실은 모델 정확도에 치명적 영향을 미친다.

▶ 오류 데이터 0건 달성: 이행 시 발생한 변환 오류, 제약조건 위반, NULL 처리 오류 등이 모두 해소되어 오류 건수가 0건임을 확인해야 한다. 미해소 오류가 잔존하면 AI 서비스 운영 중 장애의 원인이 된다.

▶ AI 학습 데이터 신뢰성: 이행된 레거시 데이터가 AI 모델의 학습 기초 데이터로 활용되는 경우, 이행 정확성이 곧 모델 신뢰성의 기반이 된다. TTA 개발안내서는 학습 데이터의 품질 확보를 핵심 요구사항으로 정의하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	이행 테이블 원본-타겟 건수 일치 여부	이행 대상 전체 테이블에 대해 원본-타겟 간 건수가 일치하는지 {데이터이행결과} 테이블별 건수 비교 항목에서 확인
2	이행 오류 데이터 0건 여부	이행 완료 후 변환 실패·제약조건 위반 등 오류 데이터 건수가 0건인지 {데이터이행결과} 오류 건수 항목과 '이행 오류 처리 대장'에서 확인
3	변환 규칙 정확 적용 여부	코드 매핑·타입 변환 등 '데이터 매핑 정의서' 정의 변환 규칙이 정확히 적용되었는지 {데이터이행결과} 변환 검증 항목과 대조
4	이행 대상 범위 일치 여부	이행 대상 범위가 '이행 계획서' 정의 범위와 누락 없이 일치하는지 {데이터이행결과}와 대조
5	이행 데이터 무결성 검증 여부	PK 중복·FK 참조·NOT NULL 위반 등 이행 데이터 무결성 검증이 수행되었는지 {테스트결과서} 무결성 시험 결과에서 확인
6	AI 학습 수치형 데이터 정밀도 손실 검증 여부	AI 학습에 직접 사용되는 수치형 데이터의 정밀도 손실 여부가 검증되었는지 {데이터이행결과} 샘플 데이터 비교 항목에서 확인
7	데이터 이행 요구사항 추적 가능 여부	{요구사항추적표}의 데이터 이행 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	이행 대상 전체 테이블의 원본-타겟 건수가 일치하고 오류 데이터 0건이 확인되며, 변환 규칙 적용·무결성 검증·정밀도 검증이 계획된 범위대로 완료된 경우	현행 유지
보완권고 (△)	이행은 완료되었으나 일부 테이블 건수 불일치, 오류 처리 대장에 미해소 건 잔존, 수치형 데이터 정밀도 검증 미수행 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	건수 검증 자체가 수행되지 않았거나 대량의 건수 불일치·오류 데이터가 미해소 상태로 잔존하여 이행 완전성을 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '건수 비교표의 산술 검증'을 반드시 수행하라: {데이터이행결과}의 테이블별 건수 비교표에서 원본 건수와 타겟 건수의 차이를 직접 계산하고, 합계 건수의 산술적 정합성도 확인한다. 사업자가 차이 건수를 0으로 기재했더라도, 실제 원본·타겟 건수의 뺄셈 결과와 일치하는지 교차 검증하는 것이 중요하다.

2) '이행 오류 처리 대장'의 전건 해소를 확인하라: 이행 과정에서 오류가 발생하는 것 자체는 자연스러운 현상이지만, 모든 오류가 원인 분석·조치 완료·재이행 확인까지 마무리되었는지가 핵심이다. '조치 중' 또는 '분석 중' 상태로 남아있는 건이 있으면 보완 권고 대상이다.

3) AI 학습용 수치 컬럼의 '정밀도 손실'에 특별히 주의하라: 일반 이행에서는 데이터 타입 변환 시 소수점 이하 정밀도 손실이 크게 문제되지 않을 수 있으나, AI 학습에 사용되는 수치형 특성(Feature) 컬럼에서는 미세한 정밀도 차이도 모델 성능에 영향을 줄 수 있다. 주요 수치 컬럼에 대한 원본-타겟 샘플 값 비교가 {데이터이행결과}에 포함되었는지 확인한다.

4) '건수 일치'만으로는 충분하지 않음을 인지하라: 건수가 일치하더라도 데이터 내용이 왜곡될 수 있다(예: 코드 매핑 오류로 '남성'이 모두 '여성'으로 변환). 건수 검증과 함께 코드 변환·인코딩·정밀도 등 '내용 정합성' 검증이 수행되었는지를 반드시 병행 확인한다.

5) 이행 계획서와 이행 결과의 '범위 대조'를 수행하라: '데이터 이행 계획서'에 이행 대상으로 정의된 테이블 목록과 {데이터이행결과}에 기재된 이행 완료 테이블 목록을 대조하여, 계획 대비 누락된 테이블이 없는지 확인한다. 특히 AI 사업에서 추가된 학습 데이터 전용 테이블의 이행 누락이 빈번하므로 주의가 필요하다.

AG_SD43-3-05 | 이상데이터 실제 제거율점검

데이터베이스 | 구현 | 데이터 수집·처리 | 정제(클렌징)/변환

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에 정의된 데이터 전처리 규칙에 따라 이상(Abnormal) 데이터가 실제로 식별·제거되었는지 여부를 '전처리결과서'를 중심으로 확인하고, 특히 IQR(사분위 범위) 기반 노이즈 데이터의 제거 여부 및 제거율의 적정성을 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

이상 데이터 제거는 AI 모델의 학습 품질에 직접적 영향을 미치는 핵심 전처리 활동으로, 설계 시 수립한 이상치 탐지 기준(IQR 1.5배 규칙 등)이 충실히 적용되었는지, 제거 전후의 데이터 건수·비율이 산출물에 정량적으로 기록되었는지를 검증한다.

2. 필요성

▶ AI 시스템의 학습 데이터에 이상치(Outlier)나 노이즈가 포함된 채 모델 학습이 수행되면, 모델 정확도 저하, 예측 편향, 과적합(Overfitting) 등 심각한 품질 문제를 초래할 수 있다. 특히 생성형 AI의 경우, 오염된 학습 데이터가 환각(Hallucination) 발생 빈도를 높이고 출력 신뢰성을 크게 훼손한다.

▶ 학습 데이터 품질 보증: IQR 등 통계적 기법으로 정의된 이상치 탐지 기준이 실제 구현에서 누락 없이 적용되었는지 확인하지 않으면, 설계와 구현 간 괴리로 인해 데이터 품질이 보증되지 않는다.

▶ 모델 신뢰성 확보: 노이즈 데이터가 학습에 포함되면 추론 결과의 신뢰성이 저하되며, 특히 공공 AI 서비스에서는 잘못된 의사결정 지원으로 이어질 수 있어 전처리 완전성 검증이 필수적이다.

▶ 정량적 추적 가능성: 전처리 전후의 데이터 건수, 제거율, 제거 사유를 문서화하지 않으면 향후 모델 성능 저하 시 원인 추적이 불가능하며, 감리 시점에서도 적정성 판단이 어렵다.

▶ 재현성 및 감사 추적성 보장: 이상 데이터 제거 과정이 기록되지 않으면 동일 데이터에 대한 재처리 시 결과 일관성을 보장할 수 없으며, 제3자 감사 시 검증 근거가 부재하게 된다.

3. 검토방법

No	점검 항목	확인 방법
1	이상치 탐지 기준 구현·적용 여부	IQR 1.5배 규칙 등 이상치 탐지 기준이 '데이터 전처리 설계서' 정의대로 구현·적용되었는지 '전처리결과서' 적용 기준과 대조
2	전처리 전후 데이터 건수 정량 기록 여부	전체 건수·제거 건수·잔여 건수 등 전처리 전후 데이터 건수가 '전처리결과서'에 정량 기록되었는지 확인
3	이상 데이터 제거율 허용 범위 이내 여부	이상 데이터 제거율이 사전 정의된 허용 범위 내에 있는지 '전처리결과서' 제거율과 설계서 허용 범위를 대조
4	노이즈 데이터 제거 샘플 확인 가능 여부	IQR 범위 밖 노이즈 데이터가 실제로 제거·대체되었는지 '전처리결과서' 샘플 데이터 전후 비교 첨부로 확인
5	제거 데이터 사유 기록 여부	이상치 유형·해당 컬럼·IQR 초과 값 등 제거된 데이터 사유가 '전처리결과서'에 기록되었는지 확인
6	전처리 전후 분포 변화 문서화 여부	히스토그램·기술통계량 등 전처리 전후 데이터 분포 변화가 '데이터 품질 검증 보고서'에 비교·문서화되었는지 확인
7	이상 데이터 제거 요구사항 추적 가능 여부	{요구사항추적표}의 이상 데이터 제거 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 이상치 탐지 기준(IQR 등)이 그대로 적용되고, 전처리 전후 건수·제거율·제거 사유가 정량 기록되며, 샘플 기반 전후 비교와 전처리 전후 데이터 분포 변화(히스토그램·기술통계량) 비교가 문서화되고 제거율이 허용 범위 이내인 경우	현행 유지
보완권고 (△)	이상 데이터 제거는 수행되었으나 전후 건수·제거 사유 기록 미비, 분포 변화 비교 미문서화, 제거율 허용 범위 미정의 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	이상치 탐지·제거가 수행되지 않았거나 설계 근거 없이 임의로 수행되어 적정성을 검증할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '수작업 문서검토'에만 의존하지 말고, DB 쿼리를 통한 교차 검증을 수행하라: 가이드라인 본문에는 "수작업 문서검토 및 면담을 통해 수행하며, 별도 쿼리는 필요하지 않다"고 되어 있으나, DBA 관점에서 종료단계 감리 시 '전처리결과서'의 내용이 실제 물리 DB와 일치하는지 확인하는 것

은 필수이다. 정제 전 Raw 테이블과 정제 후 Mart 테이블의 건수를 COUNT 쿼리로 직접 대조하여 '전처리결과서'의 수치와 일치하는지 확인한다.

2) IQR 기준값의 '출처'를 확인하라: '전처리결과서'에 기재된 Q1, Q3, IQR 값이 어떤 데이터셋(전체 원본 vs. 샘플)에서 산출된 것인지 확인한다. 샘플 기반으로 IQR을 산출한 경우 전체 데이터 적용 시 제거율이 크게 달라질 수 있으므로, 산출 기반의 적정성을 사업자에게 면담으로 확인할 필요가 있다.

3) '제거 0건'도 의심하라: 특정 컬럼에서 이상 데이터가 0건으로 보고된 경우, 실제로 이상치가 없는 것인지, 아니면 전처리 자체가 누락된 것인지 구분해야 한다. 해당 컬럼의 최솟값·최댓값이 IQR 범위 내에 있는지 '전처리결과서' 또는 '데이터 품질 검증 보고서'에서 확인한다.

4) 도메인에 따른 제거율 기대치를 사전에 파악하라: 금융 거래 데이터는 극단적 이상치가 빈번하여 제거율이 높을 수 있고, 센서 데이터는 장비 오류로 인한 노이즈가 집중적으로 발생할 수 있다. 감리 착수 전 해당 도메인의 데이터 특성에 대한 기본적인 이해를 갖추면 제거율의 합리성을 판단하는 데 도움이 된다.

5) 설계서가 없는 경우의 대응 전략: '데이터 전처리 설계서'가 별도 산출물로 작성되지 않은 사업의 경우, {데이터베이스 설계서}나 '데이터 구축 계획서'의 전처리 관련 항목을 대체 근거로 활용할 수 있다. 어떠한 설계 근거도 없이 전처리가 수행된 경우에는 부적합으로 판정할 수 있다.

AG_SD43-3-06 | 데이터정합성 Null형식 배치검증 점검

데이터베이스 | 구현 | 데이터 수집·처리 | 데이터 품질

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습 데이터에 대한 Null 값 검증, 형식(Format) 검증, 쓰레기 데이터(Garbage Data) 차단 등의 정합성 검증이 배치(Batch) 파이프라인으로 구현되어 정상 동작하는지를 '전처리결과서'를 중심으로 확인하고, 형식 위반 및 품질 미달 데이터가 학습 파이프라인에 유입되기 전에 사전 차단되는 체계를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

데이터 정합성 검증은 AI 학습 데이터 품질의 가장 기본적인 방어선이다.

Null, 형식 위반, 쓰레기 데이터가 학습에 포함되면 모델 학습 실패, 예측 정확도 저하, 전처리 파이프라인 장애 등이 발생하므로, 배치 검증 파이프라인의 설계·구현·동작 적정성을 검증한다.

2. 필요성

▶ AI 학습 데이터는 다양한 출처(DB, 파일, API, 크롤링 등)에서 수집되므로, Null 값, 형식 불일치, 의미 없는 쓰레기 데이터가 혼입될 가능성이 높다. 이러한 품질 미달 데이터가 전처리 단계에서 사전 차단되지 않으면, 학습 과정 전체에 누적적 품질 저하를 초래한다.

▶ 학습 파이프라인 안정성: 필수 컬럼에 Null 값이 포함되면 전처리 스크립트가 오류로 중단되거나, 인코딩·정규화 과정에서 예외가 발생하여 파이프라인 전체가 실패할 수 있다. 사전 검증으로 Null을 차단해야 파이프라인이 안정적으로 동작한다.

▶ 모델 정확도 보장: 날짜 컬럼에 텍스트가 적재되거나, 수치 컬럼에 문자열이 혼입되는 등 형식 위반 데이터는 특성(Feature) 변환 시 왜곡을 유발하여 모델 정확도를 저하시킨다.

▶ 쓰레기 데이터 제거: 의미 없는 반복 문자열(예: 'aaaa...'), 테스트 데이터, 빈 문자열, 특수문자만으로 구성된 데이터 등은 학습에 유해하므로 사전 차단이 필요하다.

▶ 자동화된 품질 게이트(Quality Gate): 데이터 정합성 검증이 수작업이 아닌 배치 파이프라인으로 자동화되어야, 모든 데이터 적재 시 일관된 품질 기준이 적용되고 인적 오류가 방지된다.

3. 검토방법

No	점검 항목	확인 방법
1	정합성 검증 규칙의 배치 파이프라인 구현 여부	Null·형식·도메인·쓰레기 등 설계서 정의 정합성 검증 규칙이 배치 파이프라인으로 구현되었는지 '전처리 설계서' 규칙과 '전처리결과서' 실행 결과를 대조
2	형식·Null 위반 건수 정량 기록 여부	검증 실행 후 형식 위반·Null 위반 건수가 '전처리결과서'에 컬럼별로 정량 기록되었는지 확인
3	위반 데이터 학습 유입 차단 여부	위반 데이터가 학습 파이프라인에 유입되지 않도록 차단(격리·삭제·대체)되었는지 '전처리결과서' 차단 처리 항목 또는 '품질 검증 보고서'에서 확인
4	검증 배치 자동 통합·정기 실행 여부	검증 배치가 데이터 적재 파이프라인에 자동 통합되어 정기 실행되는지 '파이프라인 설계서' 배치 실행 구조에서 확인
5	검증 결과 리포트 자동 생성 여부	검증 결과 리포트(대시보드·보고서)가 자동 생성되는지 '전처리결과서' 또는 '품질 검증 보고서' 리포트 형식에서 확인
6	위반 건수의 허용 기준 이내 여부	위반 건수가 설계서 정의 허용 기준(예: Null률 0%, 형식 위반율 0.1% 미만) 이내인지 '전처리결과서' 위반율과 대조
7	데이터 정합성 요구사항 추적 가능 여부	{요구사항추적표}의 데이터 정합성 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 정합성 검증 규칙(Null·형식·도메인·쓰레기)이 배치 파이프라인으로 자동 구현되고, 위반 건수가 정량 기록되며 위반 데이터 전건이 차단되어 허용 기준 이내이고, 검증 결과 리포트(대시보드·보고서)가 자동 생성되는 경우	현행 유지
보완권고 (△)	검증 배치는 구현되었으나 수동 실행 방식이거나, 탐지 건수와 차단 건수 불일치, 결과 리포트 미생성 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	정합성 검증 배치가 미구현되어 Null·형식 위반·쓰레기 데이터가 차단 없이 학습 파이프라인에 유입되는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '위반 건수 = 차단 건수' 일치를 산술 검증하라: '전처리결과서'에서 가장 핵심적인 점검은 탐지된 위반 건수와 실제 차단(격리·삭제) 건수가 일치하는지 여부이다. 위반 85건이 탐지되었는데 차단이 80건이면 5건은 학습에 유입된 것이므로, 이 산술적 일치를 반드시 확인한다.

2) '필수 컬럼 Null 0건'을 최우선으로 확인하라: AI 학습의 핵심 특성(Feature) 컬럼에 Null이 존재하면 학습 파이프라인이 중단되거나 모델 정확도가 저하된다. '전처리결과서'에서 필수(NOT NULL) 컬럼의 Null 건수가 0건인지를 최우선으로 확인한다.

3) '수동 실행' 방식의 검증 배치는 보완 권고하라: 정합성 검증이 담당자의 수동 실행에 의존하면, 실행 누락 시 품질 미달 데이터가 학습에 유입된다. 파이프라인에 자동 통합되어 매 적재 시 자동 실행되는 구조인지 확인하고, 수동 방식이면 보완 권고한다.

4) '품질 게이트(Quality Gate)' 개념의 구현 여부를 확인하라: 정합성 검증 결과가 허용 기준을 초과하면 학습 파이프라인이 자동 중단되는 '품질 게이트' 구조가 이상적이다. '파이프라인 설계서'에 게이트 단계가 설계되어 있는지, 위반 시 중단 메커니즘이 구현되었는지 확인한다.

5) 쓰레기 데이터 검증의 '규칙 정의'를 확인하라: 쓰레기 데이터의 판별 기준이 명확히 정의되지 않으면, 정상 데이터가 오차단되거나 쓰레기 데이터가 미탐지될 수 있다. '데이터 전처리 설계서'에 쓰레기 데이터의 구체적 판별 규칙(빈 문자열, 반복 패턴, 테스트 키워드 목록, 최소 길이 등)이 정의되어 있는지 확인한다.

AG_SD43-3-07 | 외부크롤링 적재 무결성점검

데이터베이스 | 구현 | 데이터 수집·처리 | 학습데이터

근거 | 「저작권법」 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 외부 웹사이트·API 등에서 크롤링(Crawling)으로 수집한 데이터가 {데이터베이스}에 누락 없이 적재되었는지(누락률 1% 미만), 각 수집 데이터에 저작권(라이선스) 정보가 명시적으로 기록되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 외부 크롤링 데이터를 학습·추론의 핵심 자원으로 활용하므로, 수집 과정에서의 데이터 누락은 학습 데이터 대표성을 훼손하고, 저작권 미관리는 법적 리스크를 초래한다.

이에 따라 수집 무결성(누락률)과 저작권 관리의 적정성을 정량적으로 검증한다.

2. 필요성

▶ 생성형 AI를 포함한 AI 시스템은 대규모 외부 데이터(웹 문서, 이미지, 뉴스, SNS 데이터 등)를 크롤링하여 학습에 활용하는 경우가 빈번하다. 크롤링 데이터의 무결성과 저작권 관리가 미흡하면, AI 모델의 학습 품질 저하와 함께 저작권 침해라는 법적 위험에 직면하게 된다.

▶ 수집 완전성 보장: 크롤링 대상 중 네트워크 오류, 차단(Blocking), 타임아웃 등으로 수집에 실패한 건이 일정 비율을 초과하면 학습 데이터의 대표성이 훼손된다. 누락률 1% 미만 달성은 수집 완전성의 최소 기준이다.

▶ 저작권(라이선스) 리스크 관리: 크롤링 데이터는 원저작자의 저작권이 적용되므로, 수집 데이터별 라이선스 유형(CC BY, CC BY-SA, 공공 도메인, API 이용약관 등)이 DB에 명시적으로 기록되어야 향후 법적 분쟁 시 근거를 제시할 수 있다.

▶ AI 학습 데이터 계보(Lineage) 확보: 수집원(Source URL), 수집 일시, 라이선스 상태가 데이터와 함께 관리되어야, 특정 학습 데이터의 출처와 이용 권한을 사후에 추적할 수 있다.

▶ TTA·NIST 요구사항 준수: TTA 개발안내서는 학습 데이터의 수집 출처 명시와 이용 권한 확인을 요구하며, NIST AI RMF는 데이터 수집·선택의 과학적 무결성(Scientific Integrity)을 강조하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	수집 누락률 1% 미만 여부	수집 대상 건수 대비 실제 적재 건수의 누락률이 1% 미만인지 '데이터 수집 결과 보고서' 수집원별 건수 비교에서 확인
2	수집 실패 원인 건별 기록 여부	수집 실패 건에 대한 실패 원인이 건별로 '데이터 수집 결과 보고서'에 기록되었는지 확인
3	수집 데이터 저작권 정보 기록 여부	수집 데이터별 저작권(라이선스) 정보가 {데이터베이스} 라이선스 메타 컬럼 또는 '저작권 관리 대장'에 명시적으로 기록되었는지 확인
4	수집 출처·일시 메타 관리 여부	수집 데이터별 출처(Source URL)·수집 일시가 {데이터베이스} 수집 테이블의 메타 컬럼으로 관리되는지 확인
5	중복 수집 데이터 탐지·처리 여부	중복 수집 데이터의 탐지·처리(Deduplication)가 수행되었는지 {테스트결과서} 중복 검출 시험 결과에서 확인
6	robots.txt 준수 확인·문서화 여부	크롤링 대상 사이트의 robots.txt 준수 여부가 '데이터 수집 설계서' 크롤링 정책 항목에 확인·문서화되었는지 확인
7	외부 데이터 수집·저작권 요구사항 추적 가능 여부	{요구사항추적표}의 외부 데이터 수집·저작권 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	전체 및 수집원별 누락률이 1% 미만으로 확인되고, 수집 실패 건별 원인이 기록되며, 라이선스·출처(Source URL)·수집 일시가 메타 정보로 관리되고, 중복 처리와 robots.txt 준수 여부까지 문서화된 경우	현행 유지
보완권고 (△)	누락률 기준은 충족하나 실패 원인 기록이 일부 누락되었거나, 라이선스 컬럼의 NULL·미확인 건 비율이 높거나, 중복 처리·robots.txt 준수 확인 등 일부 항목의 문서화가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	누락률이 1% 이상이거나 산출 근거가 부재한 경우, 또는 저작권(라이선스) 정보 관리 체계가 없거나 robots.txt 준수 확인이 전혀 수행되지 않아 법적 리스크가 방치된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (관련 법령·지침 미준수 시 즉시 보완 요구 및 법적 리스크 보고)

5. 감리 시 유의사항 및 실무 Tip

1) 누락률 산술 검증을 반드시 수행하라: '데이터 수집 결과 보고서'에 기재된 대상 건수, 성공 건수, 실패 건수의 합이 정확한지, 누락률 계산이 맞는지 직접 검증한다. 사업자가 성공 건수만 기재하고 실패 건수를 누락하는 사례가 있으므로, 대상 - 성공 = 실패의 산술 검증이 중요하다.

2) 라이선스 컬럼의 'NULL 건수'를 반드시 확인하라: DB에 라이선스 컬럼이 존재하더라도, 실제 값이 NULL이거나 '미확인'으로 기재된 건이 대량 존재할 수 있다. 사업자에게 라이선스 NULL 건수를 확인하고, NULL 비율이 높으면(5% 이상) 보완을 요청한다.

3) '개별 수집원'의 누락률도 확인하라: 전체 합산 누락률이 1% 미만이더라도, 특정 수집원의 누락률이 극단적으로 높으면(예: 10% 이상) 해당 수집원의 데이터 대표성이 훼손된다. 수집원별 개별 누락률도 함께 확인하고, 이상치가 있으면 사유를 면담으로 확인한다.

4) 생성형 AI 사업에서 저작권 리스크에 특별히 주의하라: 생성형 AI(LLM 등)의 학습 데이터는 대규모 웹 크롤링으로 수집되는 경우가 많으며, 최근 저작권 분쟁이 급증하고 있다. 학습 데이터에 저작권 침해 가능성이 있는 콘텐츠(뉴스 기사, 저작물 본문 등)가 포함되지 않았는지, 라이선스 검토 절차가 수행되었는지를 특별히 확인한다.

5) robots.txt 확인은 '절차 수행 여부'를 점검하라: 감리원이 robots.txt의 기술적 내용을 직접 분석할 필요는 없다. '데이터 수집 설계서'에 크롤링 대상 사이트별로 robots.txt 확인 절차가 수행되었다는 기록이 있는지, 크롤링 금지(Disallow) 사이트를 수집 대상에서 제외하였는지를 확인하면 충분하다.

AG_SD43-3-08 | 모델 추론결과 프롬프트 응답 로깅점검

데이터베이스 | 구현 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 추론(Inference) 과정에서 발생하는 사용자 질의(프롬프트)와 모델 응답이 {데이터베이스}에 100% 누락 없이 적재(로깅)되는지를 확인하고, 로깅 테이블의 설계·구현·동작 적정성을 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

추론 로그는 AI 모델의 성능 모니터링, 품질 개선, 이상 탐지, 감사 추적, 법적 분쟁 대응의 핵심 근거 데이터이다.

로그가 누락되면 서비스 장애 원인 분석 불가, 유해 응답 추적 불가, 모델 드리프트(Drift) 탐지 불가 등 심각한 운영 리스크가 발생한다.

2. 필요성

▶ AI 서비스에서 모델 추론 로그는 일반 정보시스템의 트랜잭션 로그에 해당하는 핵심 운영 데이터이다. 특히 생성형 AI 서비스에서는 프롬프트와 응답 로그가 유해 콘텐츠 감사, 환각(Hallucination) 분석, 사용자 민원 대응, 모델 재학습 피드백의 유일한 근거이므로 100% 적재가 필수적이다.

▶ 감사 추적성(Auditability) 확보: 특정 사용자에게 어떤 질의에 대해 어떤 응답이 제공되었는지 사후에 추적할 수 있어야 한다. 이는 유해 응답 발생 시 원인 분석, 법적 분쟁 시 증거 확보, 규제기관 감사 대응의 전제 조건이다.

▶ 모델 성능 모니터링: 추론 로그를 분석하여 모델의 응답 품질 변화(드리프트), 응답 지연, 오류율 등을 실시간으로 모니터링할 수 있어야 한다. 로그 누락 시 모니터링 지표의 정확성이 훼손된다.

▶ 모델 개선 피드백: 사용자의 실제 질의 패턴과 모델 응답을 분석하여 모델 재학습(Fine-tuning) 또는 프롬프트 최적화에 활용한다. 추론 로그는 AI 서비스 지속 개선의 핵심 데이터 자산이다.

▶ TTA·NIST 요구사항 준수: TTA 개발안내서는 AI 시스템의 설명 가능성(Explainability)과 투명성(Transparency)을 요구하며, 추론 이력의 기록·관리는 이를 뒷받침하는 기술적 기반이다.

3. 검토방법

No	점검 항목	확인 방법
1	추론 건수-로그 적재 건수 100% 일치 여부	응용서버 추론 건수와 DB 로그 적재 건수가 100% 일치(누락 0건)하는지 {테스트결과서} 추론 건수-로그 건수 비교 결과에서 확인
2	로그 테이블 필수 항목 기록 여부	요청ID·사용자ID·프롬프트·응답·모델버전·일시 등 필수 항목이 로그 테이블에 기록되는지 '로깅 설계서' 필수 항목과 실제 테이블 구조를 대조
3	로그 테이블 구조 설계 일치 여부	컬럼·타입·인덱스·파티셔닝 등 로그 테이블 구조가 '데이터베이스 설계서'와 일치하는지 실제 DB 구조와 대조
4	비동기 로깅 메시지 유실 방지 여부	비동기 로깅 시 메시지 유실 없이 DB에 적재되는지 {테스트결과서} 비동기 적재 안정성 시험 결과에서 확인
5	로그 개인정보 마스킹·비식별화 적용 여부	로그에 개인정보 포함 시 마스킹 또는 비식별화 처리가 적용되었는지 {테스트결과서} 개인정보 처리 시험 또는 로그 테이블에서 확인
6	로그 보존 기간·삭제 정책 정의·구현 여부	로그 보존 기간 및 삭제(아카이빙) 정책이 '로깅 설계서'에 정의되고 구현되었는지 확인
7	추론 로깅 요구사항 추적 가능 여부	{요구사항추적표}의 추론 로깅 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	응용서버 추론 건수와 DB 로그 적재 건수가 100% 일치하고, 필수 항목·테이블 구조가 설계서와 일치하며, 비동기 로깅 시 메시지 유실 없이 적재됨이 시험으로 확인되고, 개인정보 마스킹과 보존·삭제 정책이 구현된 경우	현행 유지
보완권고 (△)	추론 로깅은 구현되었으나 일부 필수 항목(모델 버전 등) 누락, 비동기 유실·피크 부하 시험 미수행, 보존 정책 미정의 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	추론 로그 적재가 대량 누락되거나 개인정보가 평문으로 적재되는 등 로깅 체계가 실질적으로 미구현된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '추론 건수 = 로그 건수' 산술 검증을 최우선으로 수행하라: {테스트결과서}에서 응용서버의 추론 처리 건수와 DB 로그 적재 건수의 차이를 직접 계산하고, 차이가 0건인지 확인하는 것이 가장

핵심적인 점검이다. 99.9%도 대량 서비스에서는 수천 건 누락을 의미하므로, 100%(0건 누락)를 기준으로 판정한다.

2) '모델 버전' 컬럼의 존재를 반드시 확인하라: 모델 버전이 기록되지 않으면, 모델 업데이트 후 성능 변화 분석이 불가능하고 장애 시 원인 모델을 특정할 수 없다. 로그 테이블에 model_version 컬럼이 존재하고 실제로 값이 기록되는지 확인한다.

3) 프롬프트 내 '개인정보 마스킹'을 별도로 확인하라: 사용자가 프롬프트에 개인정보(이름, 전화번호, 주소 등)를 입력할 수 있으므로, 로그 적재 시 개인정보 패턴을 탐지·마스킹하는 처리가 구현되었는지 확인한다. 개인정보가 평문으로 로그에 적재되면 개인정보 보호법 위반 소지가 있다.

4) '피크 부하' 시험 포함 여부를 확인하라: 정상 부하에서 100% 적재가 확인되더라도, 피크 부하(대량 동시 요청) 시에는 로그 적재 지연·누락이 발생할 수 있다. {테스트결과서}에 피크 부하(정상의 3~5배) 시험이 포함되어 있는지 확인한다.

5) 로그 보존 기간이 '무제한'이 아닌지 확인하라: AI 추론 로그는 일 수만~수십만 건 발생하므로 보존 기간이 미정의되면 스토리지가 폭증한다. '로깅 설계서'에 보존 기간(예: 90일)과 만료 시 삭제·아카이빙 정책이 정의되어 있는지 확인한다. 특히 개인정보가 포함된 로그는 '개인정보 처리 방침서'의 보유 기간과 일치해야 한다.

AG_SD43-3-09 | 정제 전후DB분리 환경점검

데이터베이스 | 구현 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「개인정보 보호법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 학습 데이터의 정제 전 원본 데이터(Raw Zone)와 정제 후 가공 데이터(Marts Zone)가 논리적 또는 물리적으로 분리된 환경에서 관리되고 있는지를 {데이터베이스} 관련 산출물을 중심으로 확인하고, 분리 설계가 구현 단계에서 충실히 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템의 데이터 파이프라인은 원본 수집 → 정제·전처리 → 학습용 가공의 단계를 거치며, 각 단계의 데이터가 혼재되면 데이터 계보(Lineage) 추적 불가, 전처리 오류 시 원본 복구 불가, 학습 데이터 오염 등의 심각한 문제가 발생한다.

이에 따라 데이터 존(Zone) 분리의 설계·구현 적합성을 검증한다.

2. 필요성

▶ AI 시스템은 일반 정보시스템과 달리 원본 데이터에 대한 다양한 전처리(이상치 제거, 정규화, 특성 추출, 샘플링 등)를 거쳐 학습 데이터를 생성하며, 이 과정에서 원본과 가공본의 명확한 분리가 이루어지지 않으면 데이터 품질 관리, 재현성 확보, 장애 대응에 심각한 차질이 발생한다.

▶ 원본 데이터 보전: Raw Zone에 원본 데이터가 변경 없이 보존되어야 전처리 오류 발생 시 원본으로부터 재가공이 가능하다. 원본과 가공본이 동일 스키마·테이블에 혼재되면 원본 복구가 불가능해진다.

▶ 데이터 계보(Lineage) 추적: Raw → Staging → Marts의 단계별 존 분리가 이루어져야 특정 학습 데이터가 어떤 원본에서 어떤 전처리를 거쳐 생성되었는지 추적할 수 있다. 이는 AI 모델의 재현성과 감사 추적성의 전제 조건이다.

▶ 접근 권한 분리: 원본 데이터(개인정보 포함 가능)와 비식별화된 학습 데이터를 분리하여, 역할별로 접근 가능한 Zone을 차등 통제할 수 있어야 한다. 이는 개인정보보호와 데이터 보안의 기본 원칙이다.

▶ 전처리 파이프라인 독립성: 정제 전/후 데이터가 분리되어야 전처리 파이프라인의 수정·재실행이 기존 가공 데이터에 영향을 주지 않으며, 파이프라인 간 의존성을 최소화할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	Zone 분리 구조 DB 구현 여부	Raw/Staging/Marts Zone 분리 구조가 '데이터베이스 설계서' 정의대로 {데이터베이스}에 실제 구현되었는지 스키마·테이블 구조와 대조
2	Raw Zone 원본 데이터 보존 여부	Raw Zone 원본 데이터가 전처리 과정에서 변경·삭제되지 않고 보존되는지 {테스트결과서} 원본 보전 시험 결과 또는 사업자 면담으로 확인
3	Zone 간 데이터 이동 경로 구현 여부	Raw→Staging→Marts Zone 간 데이터 이동 경로가 'ETL/ELT 설계서'대로 구현되었는지 실제 파이프라인과 대조
4	Zone별 접근 권한 분리·통제 여부	Zone별 접근 권한이 분리되어 비인가 사용자의 Raw Zone 직접 접근이 통제되는지 {테스트결과서} 접근 권한 시험 또는 '보안 설계서'에서 확인
5	Marts Zone 학습 데이터 구분 가능 여부	Marts Zone 학습 데이터가 스키마명·접두사·메타 태깅 등으로 Raw Zone 원본과 명확히 구분 가능한지 '데이터베이스 설계서' 네이밍 규칙과 실제 테이블명에서 확인
6	Zone 분리 수준 명시·구현 여부	Zone 분리가 논리적/물리적 어느 수준으로 이루어졌는지 '데이터 아키텍처 설계서'에 명시되고 구현되었는지 확인
7	Zone 분리 요구사항 추적 가능 여부	{요구사항추적표}의 Zone 분리 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 Zone 분리 구조(Raw/Staging/Marts)가 {데이터베이스}에 실제 구현되고, 원본 보전·이동 경로·Zone별 접근 권한 분리가 모두 확인되며, Marts Zone의 학습 데이터가 스키마명·접두사·메타 태깅으로 Raw Zone 원본과 구분되고 Zone 분리 수준(논리적·물리적)이 설계서에 명시·구현된 경우	현행 유지
보완권고 (△)	Zone 분리는 구현되었으나 네이밍 규칙 미적용으로 Zone 식별 곤란, 접근 권한 미분리, 분리 수준(논리적/물리적) 미명시 등 일부 요건이 미충족된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	원본 데이터와 가공 데이터가 동일 스키마·테이블에 혼재되어 데이터 계보 추적과 원본 복구가 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 스키마명·테이블명의 '접두사'로 Zone 분리를 빠르게 확인하라: 사업자가 제출한 DB 카탈로그 또는 ERD에서 스키마명(raw_, stg_, marts_)이나 테이블명 접두사를 확인하는 것만으로도 Zone 분리 여부를 신속히 파악할 수 있다. 접두사 규칙이 없으면 분리가 형식적일 가능성이 높다.

2) 'Raw Zone에 쓰기 권한'이 부여된 계정을 확인하라: Raw Zone은 원본 보전이 핵심이므로, 전처리 파이프라인 계정에도 SELECT(읽기)만 부여되어야 한다. INSERT/UPDATE/DELETE 권한이 있으면 원본 변경 위험이 있으므로, 사업자에게 해당 권한의 필요성과 보호 조치를 면담으로 확인한다.

3) 'ETL 이동 방식'의 Copy vs. Move를 반드시 확인하라: Raw → Marts로 데이터를 이동할 때 'Move(원본 삭제 후 이동)' 방식이면 Raw Zone의 원본이 사라진다. 반드시 'Copy(복사)' 방식인지 'ETL/ELT 설계서'에서 확인하고, 불명확한 경우 사업자 면담으로 확인한다.

4) 클라우드 환경에서의 Zone 분리 패턴을 이해하라: 클라우드 기반 AI 사업에서는 Raw Zone을 오브젝트 스토리지(S3, GCS 등)에, Marts Zone을 데이터 웨어하우스(BigQuery, Redshift 등)에 배치하는 것이 일반적이다. 이 경우 '별도 인스턴스'가 아닌 '별도 서비스'로 물리적 분리가 이루어지므로, 분리 여부 판단 시 이를 고려한다.

5) 2단계(Raw-Marts) 구조도 적합할 수 있음을 인지하라: 모든 사업에서 3단계(Raw-Staging-Marts) 분리가 필요한 것은 아니다. 소규모 사업이나 전처리 단계가 단순한 경우 2단계(Raw-Marts) 구조도 적합하다. 감리원은 '몇 단계로 분리했는가'보다 '설계서에 정의된 분리 구조가 실제로 구현되었는가'를 중심으로 점검한다.

AG_SD43-3-10 | 데이터 소스변경 모니터링 알림구현 적정성 점검

데이터베이스 | 구현 | 데이터 수집·처리 | 데이터드리프트

근거 | NIST AI RMF MANAGE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 학습 및 추론에 사용되는 데이터 소스(수집 URL, 외부 연계 API, 원천 DB 스키마 등)에 변경이 발생할 경우, 이를 자동으로 감지하여 관리자에게 알림(에러 알림)을 발송하는 모니터링 메커니즘이 설계대로 구현되었는지, 그리고 해당 기능에 대한 시험이 적정하게 수행되었는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

AI 시스템은 데이터 소스의 안정성에 전적으로 의존하므로, 수집 URL의 접속 불가, 외부 API 응답 스키마 변경, 원천 테이블의 컬럼 추가·삭제·타입 변경 등이 발생하면 데이터 파이프라인이 중단되거나 오류 데이터가 학습에 투입되어 모델 성능이 급격히 저하될 수 있다.

이러한 변경을 조기에 감지하고 관리자에게 즉시 알리는 체계가 구축되어야만 AI 서비스의 안정성과 데이터 품질을 확보할 수 있다.

2. 필요성

▶ 데이터 파이프라인 중단 방지: AI 시스템의 수집 URL이 변경되거나 외부 API가 폐기·변경되면 데이터 수집이 중단되어 모델 재학습 및 추론 서비스 전체가 영향을 받는다. 변경 감지 알림이 없으면 장애 인지까지 수 시간~수일이 소요되어 서비스 가용성이 심각하게 저하될 수 있다.

▶ 오류 데이터 학습 차단: 원천 DB의 스키마(컬럼명, 데이터 타입, 제약 조건 등)가 사전 통보 없이 변경되면, ETL 파이프라인이 정상 동작하는 것처럼 보이면서도 실제로는 잘못된 데이터를 적재할 수 있다. 이를 감지하지 못하면 오염된 데이터가 모델 학습에 투입되어 성능 저하와 편향 심화를 초래한다.

▶ 데이터 드리프트(Data Drift) 조기 경보: 외부 데이터 소스의 분포 특성이 변경되는 데이터 드리프트는 AI 모델 성능 저하의 주요 원인이다. 소스 변경 알림 체계는 드리프트 발생의 전조 신호를 조기에 포착하여 모델 재학습 시점을 판단하는 데 핵심적인 역할을 한다.

▶ 규제 준수 및 감사 추적성 확보: 공공 AI 사업에서는 데이터 수집 경로와 소스의 변경 이력을 투명하게 관리해야 한다. NIST AI RMF의 MANAGE 기능에서도 AI 시스템 운영 중 발생하는 변경 사항의 문서화와 대응 체계 구축을 요구하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	변경 감지 대상 정의 완전성	AI 학습·추론에 사용되는 모든 데이터 소스(수집 URL, API, 원천 DB 테이블 등)가 모니터링 대상으로 빠짐없이 등록되어 있는지 확인
2	스키마 변경 감지 로직 구현	원천 테이블의 컬럼 추가·삭제·타입 변경, API 응답 필드 변경 등 스키마 수준의 변경을 자동 감지하는 로직이 구현되었는지 확인
3	수집 URL·API 접속 상태 감지	수집 URL의 접속 불가(HTTP 에러), API 엔드포인트 변경·폐기, 인증 토큰 만료 등 접속 장애를 자동 감지하는 기능 구현 여부 확인
4	알림 발송 채널·대상 적정성	변경 감지 시 알림이 적절한 채널(이메일, SMS, 메신저 등)을 통해 지정된 관리자에게 발송되도록 구현되었는지 확인
5	알림 등급 분류 및 에스컬레이션	변경 유형별 심각도 등급(Critical, Warning, Info)이 정의되고, 등급에 따른 에스컬레이션 절차가 구현되었는지 확인
6	변경 이력 로그 기록 적정성	데이터 소스 변경 감지 이력(감지 시각, 변경 유형, 이전/이후 상태, 알림 발송 결과 등)이 로그 테이블에 기록되는지 확인
7	시험 시나리오 완전성	데이터 소스 변경 감지·알림에 대한 시험 시나리오가 다양한 변경 유형(스키마, URL, 인증, 데이터 포맷 등)을 포괄하는지 확인
8	시험 결과의 합격 기준 충족	시험 결과가 사전 정의된 합격 기준(감지 시간, 알림 도달 시간, 오탐률 등)을 충족하는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	데이터 소스 변경 감지·알림 기능이 설계대로 구현되고, 다양한 변경 유형에 대한 시험이 수행되어 합격 기준을 충족하며, 변경 이력 로그가 체계적으로 기록되는 경우	현행 유지
보완권고 (△)	변경 감지·알림 기능은 구현되었으나, 일부 변경 유형에 대한 시험이 누락되었거나, 알림 등급·에스컬레이션 절차가 불완전한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	데이터 소스 변경 감지·알림 기능이 구현되지 않았거나, 핵심 변경 유형에 대한 시험이 전혀 수행되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모니터링 대상의 '완전성'을 최우선으로 점검하라: 변경 감지 로직이 아무리 정교해도, 모니터링 대상에서 누락된 소스는 감지할 수 없다. '데이터 수집 설계서'의 전체 소스 목록과 메타 관리 테이블의 등록 건수를 1:1로 대조하는 것이 가장 효과적인 점검 방법이다.

2) '알림 피로(Alert Fatigue)' 방지 설계를 확인하라: 모든 변경을 동일 등급으로 알림하면 관리자가 알림을 무시하게 되어, 정작 심각한 변경(소스 접속 불가 등)에 대한 대응이 지연될 수 있다. 심각도 등급 분류(Critical/Warning/Info)와 등급별 차별화된 알림 채널이 설계되어 있는지를 반드시 확인한다.

3) 시험 시나리오에 '비정상 경로'가 포함되었는지 확인하라: 정상적인 스키마 변경 감지뿐 아니라, 네트워크 단절 중 변경 발생, 동시 다발적 변경, 알림 서버 장애 시 재발송 등 비정상 경로(Edge Case) 시험이 {테스트결과서}에 포함되어 있는지 점검하면 시험의 충분성을 보다 정밀하게 평가할 수 있다.

4) 변경 이력 로그의 '감사 추적성'에 주목하라: 변경 이력 로그는 단순한 기술 운영 로그가 아니라, AI 모델의 데이터 품질 추적과 규제 대응을 위한 감사 증적이다. 로그 테이블에 변경 감지 시각, 변경 전후 상태, 조치 결과가 모두 기록되는지 확인하고, 로그 보관 기간이 사업 요구사항에 부합하는지도 함께 점검한다.

5) 설계-구현-시험의 '일관성 체인'을 추적하라: 감리의 핵심은 '알림·통보 정책 정의서'(설계) → 실제 구현(알림 발송 채널·대상) → {테스트결과서}(시험)의 3단계가 일관되게 연결되어 있는지를 확인하는 것이다. 설계서에는 이메일·SMS 알림으로 정의되어 있는데 시험은 이메일만 수행한 경우, 또는 시험 시나리오에 설계서에 없는 항목이 포함된 경우에는 설계-구현-시험 간 불일치로 판단하여 보완을 요구한다.

AG_SD44-3-01 | 보호변수 배제 분리 보관 확인점검

데이터베이스 | 시험 | 데이터 수집·처리 | 편향/공정성

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」제34조 · 「개인정보 보호법」 · EU AI Act · 「장애인차별금지 및 권리구제 등에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델 훈련에 사용되는 데이터 스키마(훈련 테이블/뷰)에 성별, 인종, 민족, 연령, 장애 여부, 종교, 출신 지역 등 보호변수(Protected Attributes)가 직접 컬럼으로 포함되지 않았는지, 그리고 해당 변수들이 설계서에 정의된 대로 별도 스키마 또는 별도 테이블에 분리 보관되었는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

이를 통해 AI 모델이 보호변수를 직접 학습 특성(Feature)으로 사용하여 차별적 의사결정을 내리는 위험을 데이터베이스 구조 수준에서 사전 차단하고, 공정성(Fairness) 확보를 위한 설계 의도가 물리 DB에 올바르게 구현되었는지를 확인한다.

2. 필요성

▶ AI 시스템의 공정성은 알고리즘 수준뿐만 아니라 데이터베이스 구조 수준에서부터 확보되어야 한다. 보호변수가 훈련 스키마에 직접 포함되면, 모델이 이를 학습 특성으로 사용하여 특정 집단에 대한 차별적 예측을 생성할 위험이 있다. 이는 다음과 같은 문제로 이어진다.

▶ 직접 차별(Direct Discrimination) 위험: 성별, 인종 등 보호변수가 훈련 데이터의 컬럼으로 직접 포함되면, AI 모델이 이를 예측의 핵심 변수로 학습하여 '남성이므로 대출 승인', '특정 인종이므로 채용 탈락'과 같은 직접 차별적 의사결정을 내릴 수 있다.

▶ 법적 컴플라이언스 위반: 개인정보 보호법, AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 2026.1.22 시행), EU AI Act 등 국내외 법규는 AI 시스템의 차별 금지를 명시하고 있다. 보호변수가 훈련 스키마에 포함된 상태에서 모델이 배포되면 법적 제재 및 분쟁의 대상이 될 수 있다.

▶ 공정성 검증 체계 무력화: 보호변수가 훈련 스키마와 분리되지 않으면, 사후에 '모델이 보호변수를 사용했는지 여부'를 검증하기 어려워진다. 분리 보관은 공정성 감사(Fairness Audit)의 물리적 전제 조건이다.

▶ 대리변수(Proxy Variable) 위험 관리의 출발점: 보호변수를 훈련 스키마에서 배제하더라도, 우편번호·학교명 등 대리변수를 통해 간접적으로 보호변수가 추론될 수 있다. 그러나 직접 변수의 배제는 대리변수 관리의 필수적인 첫 단계이다.

3. 검토방법

No	점검 항목	확인 방법
1	보호변수 목록 정의 확인	설계서에 보호변수 목록(성별, 인종, 연령, 장애 등)이 명시적으로 정의되었는지 확인
2	훈련 스키마 내 보호변수 직접 컬럼 부재 확인	훈련용 테이블/뷰의 컬럼 목록에 보호변수에 해당하는 직접 컬럼이 존재하지 않는지 확인
3	보호변수 분리 보관 테이블 구현 확인	보호변수가 별도 테이블/스키마에 분리 보관되었는지, 설계서대로 구현되었는지 확인
4	훈련 스키마-분리 스키마 간 접근통제 확인	훈련 프로세스가 분리 보관된 보호변수 테이블에 접근할 수 없도록 DB 권한이 설정되었는지 확인
5	훈련용 뷰(View) 정의 적정성	훈련 데이터 추출용 뷰가 보호변수를 배제하고 정의되었는지, 뷰 정의문(DDL)을 점검
6	대리변수(Proxy) 식별 및 문서화 확인	보호변수와 높은 상관관계를 가진 대리변수가 식별되고 관리 방안이 문서화되었는지 확인
7	시험 결과 기반 보호변수 배제 확인	보호변수 배제 관련 시험(컬럼 존재 여부, Feature 목록 검증 등)이 수행되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	보호변수 목록이 설계서에 명시되고, 훈련 스키마 내 보호변수 직접 컬럼이 0건이며, 훈련 데이터 추출용 뷰가 보호변수를 배제하고 정의되고, 분리 보관 테이블이 설계대로 구현되어 접근통제가 적용되며, 보호변수와 상관관계가 높은 대리변수(Proxy)가 식별·문서화되고 보호변수 배제 시험 결과가 확인된 경우	현행 유지 (사례: 보호변수 8개 정의, 훈련 뷰 5개 전수 검증 결과 보호변수 0건, 분리 보관 테이블 구현 및 접근통제 적용 확인)
보완권고 (△)	보호변수 목록은 정의되었으나, 일부 훈련 테이블에 보호변수 컬럼이 잔존하거나, 분리 보관이 불완전하거나, 접근통제가 미설정된 경우	해당 항목 보완 요청 후 재확인 (사례: 훈련 테이블 1건에 BIRTH_DATE 잔존, 분리 보관 테이블 미생성, 접근통제 미설정)
부적합 (X)	보호변수 목록이 미정의, 훈련 스키마에 성별·인종 등 보호변수 컬럼이 다수 포함, 또는 분리 보관 설계 자체가 부재한 경우	시정조치 요구 및 이행 결과 확인 (사례: 보호변수 미정의, GENDER/RACE 컬럼이 훈련 테이블에 존재, 분리 보관 계획 부재)

5. 감리 시 유의사항 및 실무 Tip

1) 보호변수 목록의 '도메인 적합성'을 반드시 확인하라: 보호변수는 사업 도메인에 따라 달라진다. 의료 AI에서는 인종·연령·성별이 핵심이고, 채용 AI에서는 성별·연령·학력·출신학교·군복무 여부가 중요하며, 금융 AI에서는 성별·연령·거주지역이 핵심이다. 사업자가 정의한 보호변수 목록이 해당 도메인의 법적·윤리적 기준에 부합하는지 검토한다.

2) View 기반의 논리적 배제 여부를 철저히 검증하라: 개발자가 편의상 원본 테이블에는 보호변수를 남겨두고 학습 파이프라인에 제공하는 View에서만 해당 컬럼을 제외하는 경우가 있다. 종료단계 감리에서는 View의 DDL뿐만 아니라 기저(Base) 테이블의 스키마와 데이터 접근 권한을 확인하여, 권한 없는 ML 엔지니어나 자동화 파이프라인이 Base 테이블에 직접 접근할 수 없는지까지 검증한다.

3) BIRTH_DATE(생년월일)의 연령 추론 가능성에 주의하라: 생년월일 자체는 일반적인 개인정보이지만, AI 모델 학습에서는 현재 날짜와 결합하여 '연령'을 직접 추론할 수 있으므로 보호변수에 준하여 취급해야 한다. 훈련 스키마에 BIRTH_DATE가 포함된 경우, 연령 추론 가능성을 감안하여 배제 또는 연령대(Age Band)로의 변환 여부를 확인한다.

4) 분리 보관 테이블의 '접근통제'가 핵심임을 강조하라: 보호변수를 별도 테이블에 분리했더라도, 훈련 프로세스 DB 계정이 해당 테이블에 접근할 수 있으면 분리의 의미가 없다. GRANT/REVOKE 설정을 통해 훈련 계정의 접근이 물리적으로 차단되었는지를 반드시 확인한다. 이것이 보호변수 분리 보관의 실효성을 결정한다.

5) '보호변수 미정의'는 그 자체로 부적합 사유이다: 사업자가 '우리 시스템은 보호변수가 없다' 또는 '공정성 이슈가 없다'고 주장하는 경우가 있다. 그러나 모든 AI 시스템은 잠재적 편향 위험을 내포하고 있으므로, 보호변수 목록의 정의 자체가 필수이다. 보호변수를 정의하지 않은 것은 '문제가 없다'가 아니라 '검토하지 않았다'는 의미이므로, 부적합 판정을 검토한다.

AG_SD44-3-02 | 학습데이터 버전관리 점검

데이터베이스 | 시험 | 데이터 수집·처리 | 거버넌스/데이터계보

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 학습 데이터의 버전관리 체계가 정상 동작하는지 검증하는 것을 목적으로 한다.

단, 최신 DVC 환경은 순수 관계형 데이터베이스에 데이터를 담기보다 메타데이터만 DB(PostgreSQL 등)에 담고, 실제 데이터 페이로드는 S3 등 오브젝트 스토리지에 저장하는 하이브리드 구조를 가지므로, 데이터베이스 중심이 아닌 'DB와 스토리지를 연동한 아키텍처 관점'에서 구현 적정성을 점검해야 한다.

학습 데이터 버전관리는 AI 모델의 재현성(Reproducibility)과 추적성(Traceability)을 보장하기 위한 핵심 인프라로, 모델 성능 저하·장애 발생 시 원인이 된 데이터 버전을 신속히 식별하고 이전 정상 버전으로 복구할 수 있는 체계가 구축·검증되었는지를 확인한다.

2. 필요성

▶ AI 시스템은 일반 정보시스템과 달리 소스 코드뿐 아니라 학습 데이터가 모델 동작을 결정하는 핵심 자산이다. 동일한 모델 코드라도 학습 데이터가 변경되면 모델의 예측 결과가 완전히 달라질 수 있으므로, 학습 데이터에 대한 체계적인 버전관리는 AI 시스템 품질 보증의 전제 조건이 된다.

▶ 재현성(Reproducibility) 보장: 특정 모델 버전이 어떤 데이터 버전으로 학습되었는지 추적할 수 없으면, 모델 성능을 재현하거나 검증하는 것이 불가능하다. 이는 NIST AI RMF가 요구하는 AI 시스템의 유효성(Validity) 및 신뢰성(Reliability) 확보의 기본 전제이다.

▶ 장애 시 신속한 복구(Rollback): 새로운 학습 데이터로 모델을 재학습한 후 성능이 저하되거나 장애가 발생한 경우, 이전 정상 데이터 버전으로 신속히 복구할 수 있어야 서비스 연속성을 보장할 수 있다.

▶ 데이터 계보(Data Lineage) 추적: 학습 데이터의 변경 이력(언제, 누가, 무엇을, 왜 변경했는지)이 관리되지 않으면, 모델 성능 변화의 원인을 데이터 관점에서 분석할 수 없다.

▶ 감사 추적성 및 규정 준수: TTA 신뢰할 수 있는 인공지능 개발안내서는 AI 시스템의 전 생명주기에 걸친 데이터 이력 관리를 요구하고 있으며, 공공 AI 사업은 이를 준수하여 감사 시 근거를 제시할 수 있어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	데이터 리비전 식별 체계 구현 여부	해시값·버전 번호 등 데이터 리비전 식별 체계가 '데이터 버전관리 설계서'대로 구현되었는지 실제 메타 테이블 스키마({데이터베이스})와 대조
2	학습 데이터 변경 시 리비전 생성·기록 여부	학습 데이터 변경 시 새로운 데이터 리비전이 자동 또는 수동으로 생성·기록되는지 {테스트결과서} DVC 동작 시험 결과에서 확인
3	모델-데이터 리비전 연동 매핑 관리 여부	모델 리비전과 데이터 리비전의 연동 매핑이 '모델-데이터 연동 이력 대장'에 기록·관리되는지 확인
4	특정 리비전 복구(Rollback) 정상 동작 여부	특정 데이터 리비전으로의 복구(Rollback)가 정상 동작하는지 {테스트결과서} 복구 시험 결과에서 확인
5	리비전별 메타정보 기록 여부	생성 일시·생성자·변경 사유·데이터 건수 등 데이터 리비전별 메타정보가 '모델-데이터 연동 이력 대장'·메타 테이블에 기록되는지 확인
6	데이터 버전관리 운영 절차 문서화 여부	생성·조회·복구 등 데이터 버전관리 운영 절차가 '데이터 버전관리 운영 매뉴얼'에 문서화되었는지 확인
7	데이터 버전관리 요구사항 추적 가능 여부	{요구사항추적표}의 데이터 버전관리 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	데이터 리비전 식별 체계가 설계대로 구현되고, 모델-데이터 리비전 연동 매핑과 메타정보가 누락 없이 기록되며, 복구(Rollback) 시험 성공과 운영 매뉴얼 문서화까지 확인된 경우	현행 유지
보완권고 (△)	버전관리 체계는 구현되었으나, 연동 매핑 일부 누락, 복구 시험 미수행, 변경 사유 등 메타정보 기록 미흡, 운영 매뉴얼 미비 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	학습 데이터 버전관리 체계가 미구현되어 모델별 학습 데이터 버전을 추적할 수 없거나, 이전 데이터 버전으로의 복구가 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '연동 이력 대장'의 완전성을 최우선으로 점검하라: 모델-데이터 리비전 연동 매핑이 누락 없이 기록되어 있는지가 가장 핵심적인 상세 검토 항목이다. 현재 운영(또는 배포 예정) 모델 리비전을 기준으로 대응하는 데이터 리비전을 역추적해 보면, 매핑의 완전성을 빠르게 확인할 수 있다.

2) 복구 시험의 '시나리오 구체성'을 확인하라: {테스트결과서}에 '복구 시험 완료'라고만 기재된 경우, 실제 어떤 리비전에서 어떤 리비전으로 복구했는지, 복구 후 데이터 건수·내용이 원래와 일치하는지 확인할 수 없다. 시험 시나리오(복구 전 리비전 → 복구 후 리비전 → 건수 비교 → 소요 시간)가 구체적으로 기술되어 있는지 확인한다.

3) 메타 테이블의 '변경 사유' 컬럼 기록 여부를 확인하라: 데이터 리비전이 생성될 때 '왜' 데이터가 변경되었는지(신규 데이터 추가, 오류 데이터 수정, 라벨 변경 등)를 기록하는 변경 사유 컬럼은 향후 장애 원인 분석에 핵심적인 정보이다. 메타 테이블에 해당 컬럼이 존재하고 실제로 기록이 이루어지는지 확인한다.

4) DVC 도구 자체의 기능 검증은 범위 밖임을 인지하라: 감리원의 역할은 DVC 도구(DVC, MLflow, LakeFS 등)의 기능적 완전성을 검증하는 것이 아니라, 해당 도구를 활용한 버전관리 체계가 '설계대로 구현·동작하는지'를 산출물 기반으로 확인하는 것이다. 도구 자체에 대한 기술적 질문은 사업자에게 시연을 요청하는 것으로 갈음할 수 있다.

5) 운영 매뉴얼의 '독립 활용 가능성'을 점검하라: '데이터 버전관리 운영 매뉴얼'은 개발팀이 아닌 운영 인력이 독립적으로 활용할 수 있어야 한다. 매뉴얼만으로 리비전 생성·조회·복구를 수행할 수 있는 수준의 구체성(화면 캡처, 명령어 예시, 주의사항 등)을 갖추었는지 확인한다.

AG_SD44-3-03 | 개인정보 비식별화 및 파기 점검

데이터베이스 | 시험 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)·「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업에서, 학습·시험 목적으로 사용된 개인정보 및 민감정보가 계획된 절차에 따라 비식별화되었거나 완전히 파기·마스킹 처리되었는지, 그리고 그 이행 결과가 {보안점검결과서}에 적정하게 기록되었는지 여부를 점검하는 것이 목적이다.

감리원은 기술적 처리 방법 자체를 검증하는 것이 아니라, 비식별화·파기·마스킹 이행 계획의 수립 여부, 해당 계획 대비 실제 이행 여부, 그리고 책임자 서명 등 이행 근거의 존재 여부를 산출물 검토를 통해 확인한다.

2. 필요성

▶ AI 시스템은 일반 소프트웨어와 달리 모델 학습 단계부터 시험 단계까지 실제 개인정보(주민번호, 의료정보, 금융정보 등)가 포함된 데이터를 광범위하게 사용한다. 시스템 개발 완료 후 해당 데이터가 파기·마스킹되지 않고 잔존할 경우, 개인정보 침해 사고가 발생하거나 개인정보 보호법 위반에 따른 법적 제재를 받을 수 있다.

▶ 특히 생성형 AI 서비스는 학습 데이터가 모델 내부에 잔류할 수 있으며, 프롬프트 조작 등을 통해 학습 데이터에 포함된 개인정보가 외부에 노출될 수 있는 구조적 취약성이 존재한다. 이러한 이유로, 종료 단계 감리 시 학습·시험 데이터의 비식별화 및 파기 이행 여부를 반드시 확인하여야 한다.

▶ 또한, 공공 AI 서비스는 정보시스템 감리기준에 따라 보안점검 결과를 산출물로 남겨야 하며, 비식별화·파기 이행 사실이 {보안점검결과서}에 문서화되어 있지 않으면 향후 감사·감리에서 이행 여부를 소명하기 어려워진다.

3. 검토방법

No	점검 항목	확인 방법
1	비식별화 이행 계획 수립 여부	시험 착수 전 개인정보 비식별화 또는 파기·마스킹 계획이 수립되었는지, 계획서에 처리 대상 데이터 목록, 처리 방법(파기/마스킹/익명화), 처리 일정 및 담당자가 명시되어 있는지 확인

No	점검 항목	확인 방법
2	실데이터 사용 현황 파악	학습·시험 과정에서 실제 개인정보가 포함된 실데이터를 사용하였는지 여부를 'AI 학습·시험 데이터 관리대장' 또는 {보안점검결과서}를 통해 확인한다. 실데이터 사용이 있었던 경우, 사용 범위·기간·사유가 기록되어 있는지 검토
3	복구 불가 파기 이행 여부	실데이터를 사용한 경우, 사용 완료 후 복구 불가 수준으로 완전 파기되었는지 {보안점검결과서} 및 '비식별화·마스킹 이행확인서'를 검토한다. 단순 삭제가 아닌 덮어쓰기, 물리적 파기 등 구체적 파기 방법이 기재되어 있는지 확인
4	마스킹 처리 범위 및 완전성	파기 대신 마스킹을 적용한 경우, 마스킹 대상 항목(성명, 주민번호, 연락처 등 개인 식별 정보)이 누락 없이 처리되었는지 '비식별화·마스킹 이행확인서' 및 샘플 데이터를 검토하여 확인
5	이행 결과의 서명·승인 여부	{보안점검결과서} 및 '비식별화·마스킹 이행확인서'에 담당자, 개인정보보호 담당자, 프로젝트 관리자의 서명 또는 결재가 완료되어 있는지 확인한다. 서명 누락 시 이행 사실의 공식 인정이 어려우므로 보완 대상으로 기재
6	처리 결과와 산출물 일치 여부	비식별화·파기·마스킹 이행 내용이 {보안점검결과서}에 기재된 내용과 'AI 학습·시험 데이터 관리대장'의 내용이 상호 일치하는지, 모순이 없는지를 교차 검토
7	생성형 AI 모델 학습 데이터 처리 여부	생성형 AI(LLM, 이미지 생성 등) 서비스인 경우, 파인튜닝(Fine-tuning)에 사용된 학습 데이터 내 개인정보가 별도 관리되고 파기·마스킹 계획이 수립되어 이행되었는지를 관련 산출물로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	비식별화·파기·마스킹 계획이 수립되고 계획대로 이행되었으며, 처리 방법·범위와 책임자 서명이 포함된 산출물이 완비되고, 생성형 AI의 파인튜닝 학습 데이터에 포함된 개인정보가 별도 관리되어 파기·마스킹 계획이 수립·이행되었음이 확인되는 경우	현행 유지
보완권고 (△)	처리 이행은 확인되나 파기 방법 미기재, 마스킹 항목 일부 누락, 서명 미완료 등 일부 산출물이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	실데이터의 파기·마스킹이 이행되지 않았거나 이행 근거 산출물이 현저히 미흡하여 개인정보 잔존 위험이 있는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "삭제"와 "파기"는 다르다 — 방법 명시를 요구하라: {보안점검결과서}에 "삭제 완료"라고만 기재된 경우, 이를 복구 불가 파기로 인정하지 않는다. 구체적인 파기 방법(덮어쓰기 횟수, 물리적 파기 증빙, 암호화 키 폐기 증빙)이 기재되어 있는지 확인하고, 없다면 보완 사항으로 기재한다.

2) 마스킹 범위의 "누락 항목"을 반드시 확인하라: 마스킹 이행확인서에는 처리 항목 목록이 명시되어야 한다. "마스킹 완료"라는 포괄적 표현만 있는 경우, 이메일 주소, IP 주소, 간접 식별자 등 비표준 개인정보 항목이 누락되는 사례가 빈번하므로, 샘플 데이터 또는 처리 결과 목록을 요청하여 처리 범위를 교차 확인한다.

3) 생성형 AI 파인튜닝 데이터는 별도 확인 체크리스트를 적용하라: 파인튜닝에 사용된 학습 데이터는 모델 내부에 잔류할 수 있으므로 일반 시험 데이터와 별도로 관리대장·파기 계획·이행 결과를 확인한다. 파기 대상에서 파인튜닝 데이터셋이 누락된 경우 보완을 요구한다.

AG_SD44-3-04 | 데이터 옵트아웃 동의 처리점검

데이터베이스 | 시험 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」 · NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 데이터 활용에 대한 동의를 철회(옵트아웃, Opt-out)한 회원의 데이터가 AI 모델 훈련(학습) 파이프라인에서 실제로 제외되는지를 {시스템화면}과 관련 시험 결과를 중심으로 확인하고, 동의 철회 처리가 데이터베이스 수준에서 정상 반영되는 전 과정을 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

옵트아웃 처리는 개인정보 보호법상 정보주체의 권리를 보장하기 위한 핵심 기능이며, AI 시스템에서는 단순한 서비스 중단을 넘어 해당 회원의 데이터가 향후 모델 학습에 사용되지 않도록 데이터 파이프라인 수준에서 완전히 제외되어야 한다.

2. 필요성

▶ AI 시스템은 회원의 개인 데이터(행동 이력, 구매 내역, 의료 기록 등)를 학습에 활용하는 경우가 빈번하며, 회원이 데이터 활용 동의를 철회했을 때 해당 데이터가 학습 파이프라인에서 즉시·완전히 제외되지 않으면 개인정보 보호법 위반과 함께 AI 시스템에 대한 사회적 신뢰가 훼손된다.

▶ 정보주체의 권리 보장: 개인정보 보호법 제37조(개인정보의 처리정지)에 따라, 정보주체는 자신의 개인정보 처리 정지를 요구할 수 있으며, AI 학습 목적 활용도 이에 포함된다. 옵트아웃 시 학습 데이터에서의 제외가 기술적으로 보장되어야 한다.

▶ AI 학습 데이터 관리 체계 검증: 동의 철회가 단순히 서비스 화면에서만 반영되고 실제 학습 파이프라인에는 반영되지 않는 경우, 외형적으로는 처리된 것처럼 보이나 실질적으로 개인정보가 학습에 계속 사용되는 심각한 문제가 발생한다.

▶ NIST AI RMF 프라이버시 요건 충족: NIST AI RMF는 Privacy-Enhanced를 AI 신뢰성의 핵심 특성으로 정의하며, 정보주체의 동의 관리 및 데이터 활용 통제를 요구하고 있다.

▶ TTA 개발안내서 준수: TTA 신뢰할 수 있는 인공지능 개발안내서는 AI 서비스의 프라이버시 보호 요건으로 정보주체의 동의 철회 시 데이터 활용 중단 메커니즘의 구현을 요구하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	옵트아웃 신청 화면 구현 여부	옵트아웃(동의 철회) 신청 화면이 구현되어 회원이 직접 철회할 수 있는지 {시스템화면}에서 직접 확인
2	동의 상태 테이블 철회 상태 기록 여부	동의 상태 관리 테이블에 철회 상태가 정확히 기록되는지 {테스트결과서} 동의 상태 변경 시험 결과에서 확인
3	학습 데이터에서 옵트아웃 회원 제외 여부	학습 데이터 추출 쿼리·파이프라인에서 옵트아웃 회원이 실제로 제외되는지 {테스트결과서} 학습 데이터 제외 시험 결과에서 확인
4	옵트아웃 반영 시점 정의·구현 여부	옵트아웃 처리의 반영 시점(실시간/배치)이 '동의 관리 설계서'에 정의되고 구현되었는지 실제 동작과 대조
5	옵트아웃 처리 이력 DB 기록·관리 여부	철회 일시·대상 항목·처리 결과 등 옵트아웃 처리 이력이 '데이터베이스 설계서' 이력 테이블에 기록·관리되는지 확인
6	재동의 시 학습 데이터 재포함 프로세스 여부	옵트아웃 후 재동의(옵트인) 시 학습 데이터에 재포함되는 프로세스가 설계·구현되었는지 {테스트결과서} 재동의 시험 결과에서 확인
7	옵트아웃 요구사항 추적 가능 여부	{요구사항추적표}의 옵트아웃 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	옵트아웃 신청 기능과 동의 상태 관리가 구현되고, 옵트아웃 회원이 학습 데이터 추출 파이프라인에서 실제로 제외됨이 시험으로 확인되며, 처리 이력 관리와 재동의 프로세스까지 완비된 경우	현행 유지
보완권고 (△)	옵트아웃 신청과 동의 상태 기록은 구현되었으나, 학습 데이터 제외 시험 미수행, 반영 시점 미정의, 처리 이력 관리 미흡 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	옵트아웃 기능이 미구현되었거나, 동의를 철회한 회원의 데이터가 학습 파이프라인에서 제외되지 않고 계속 학습에 사용되는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '학습 데이터 제외 시험'의 수행 여부를 최우선으로 확인하라: 옵트아웃 점검의 핵심은 '철회 화면이 있는가'가 아니라 '철회 후 실제로 학습에 사용되지 않는가'이다. {테스트결과서}에 옵트아웃

회원의 학습 데이터 포함 건수가 0건임을 확인하는 시험이 포함되어 있는지를 최우선으로 점검한다.

2) '복수 추출 경로'의 존재 여부를 면담으로 확인하라: 학습 데이터가 정기 배치뿐 아니라 임시 추출, API 연계, 직접 쿼리 등 복수 경로로 추출되는 경우, 모든 경로에서 옵트아웃 필터가 적용되어야 한다. 사업자에게 학습 데이터 추출 경로가 몇 개인지 면담으로 확인하고, 각 경로에서의 필터 적용 여부를 점검한다.

3) 반영 시점의 '간격'이 합리적인지 확인하라: 동의 철회 후 다음 학습 실행까지 최대 얼마의 시간이 소요되는지 확인한다. 예를 들어 일배치(Daily Batch)라면 최대 24시간 후 반영되므로, 이 간격이 '개인정보 처리 방침서'에 고지된 처리 기간과 일치하는지 대조한다.

4) '이미 학습된 데이터'의 처리 방안을 확인하라: 옵트아웃 이전에 이미 학습에 사용된 데이터의 처리(모델 재학습, 머신 언러닝 등)는 본 점검의 직접 범위는 아니나, 해당 이슈에 대한 대응 방안이 '동의 관리 설계서'에 기술되어 있는지를 참고 확인하면 감리 보고서의 완전성을 높일 수 있다.

5) 생성형 AI 사업에서의 추가 점검 포인트: 생성형 AI(LLM)의 경우, Fine-tuning용 데이터에서의 옵트아웃 회원 제외뿐 아니라, RAG(검색 증강 생성)의 벡터 DB에서 해당 회원 관련 문서의 제외도 함께 확인해야 한다. 벡터 DB에서의 삭제·비활성화 처리가 별도로 수행되는지 면담으로 확인한다.

AG_SD44-3-05 | 대용량백업 스케줄러 실행점검

데이터베이스 | 시험 | 데이터 수집·처리 | 데이터 백업/복구

근거 | NIST AI RMF · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템의 대용량 데이터(학습 데이터, 모델 아티팩트, 추론 결과 등)에 대한 백업 스케줄러가 설계대로 구현되어 지정된 야간 시간에 자동으로 실행되고, 백업이 성공적으로 완료되는지를 {테스트결과서}를 중심으로 확인하고, 백업 체계의 구현·동작 적정성을 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 학습 데이터·모델·추론 이력 등 대용량 데이터 자산을 보유하며, 이들의 손실은 모델 재학습 비용, 서비스 중단, 데이터 복구 불가 등 치명적 피해를 초래한다.

백업 스케줄러의 정상 동작 검증은 AI 서비스의 비즈니스 연속성(Business Continuity)을 보장하기 위한 핵심 상세 검토 항목이다.

2. 필요성

▶ AI 시스템의 데이터는 일반 정보시스템과 비교하여 규모가 크고(수백 GB~수 TB), 재생성 비용이 높으며(학습 데이터 수집·가공·라벨링 비용), 모델 학습에 수 시간~수 일이 소요되므로 데이터 손실 시 복구 비용이 극히 높다.

▶ AI 데이터 자산 보호: 학습 데이터, 전처리 결과, 학습된 모델 파일(체크포인트), 추론 결과 이력 등은 AI 서비스의 핵심 자산이다. 백업 실패 시 이들 자산의 영구 손실이 발생할 수 있으며, 재구축에 수 주~수 개월이 소요될 수 있다.

▶ 서비스 연속성(Business Continuity) 보장: 장애 발생 시 백업 데이터로부터 신속한 복구(Recovery)가 가능해야 AI 서비스의 중단 시간을 최소화할 수 있다. 이는 RPO(Recovery Point Objective)와 RTO(Recovery Time Objective) 충족의 전제 조건이다.

▶ 야간 백업 시간대의 적정성 확보: 대용량 AI 데이터 백업은 서비스 부하가 낮은 야간 시간에 수행되어야 주간 서비스 성능에 영향을 미치지 않는다. 야간 백업 시간 내 완료 여부가 검증되어야 한다.

▶ 규정 및 감리 기준 준수: 정보시스템 감리 수행 가이드는 데이터 백업·복구 체계의 구현 적정성을 종료단계 상세 검토 항목으로 정의하고 있으며, NIST AI RMF도 AI 시스템의 복원력(Resilience)을 핵심 신뢰 특성으로 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	백업 스케줄러 야간 자동 실행 여부	백업 스케줄러가 지정된 야간 시간에 자동 실행되었는지 {테스트결과서} 스케줄러 실행 로그에서 확인
2	백업 오류 없이 완료 여부	백업이 오류 없이 성공적으로 완료되었는지 {테스트결과서} 백업 성공/실패 상태에서 확인
3	백업 소요 시간 윈도우 이내 여부	백업 소요 시간이 야간 백업 윈도우(시간대) 이내인지 {테스트결과서} 소요 시간과 대조
4	백업 대상 설계 일치 여부	DB·파일·모델 등 백업 대상이 '백업 설계서' 정의 대상과 일치하는지 {테스트결과서}와 대조
5	백업 복구(Restore) 시험 성공 여부	백업 데이터로부터의 복구(Restore) 시험이 수행되어 성공했는지 {테스트결과서} 복구 시험 결과에서 확인
6	백업 실패 시 알림·재시도 동작 여부	백업 실패 시 알림(Alert)·재시도(Retry) 메커니즘이 동작하는지 {테스트결과서} 장애 시험 결과 또는 '운영 매뉴얼'에서 확인
7	백업·복구 요구사항 추적 가능 여부	{요구사항추적표}의 백업·복구 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	백업 스케줄러가 지정된 야간 시간에 자동 실행되어 전체 대상 백업이 윈도우 내에 성공적으로 완료되고, 복구 시험 성공과 실패 시 알림·재시도 동작까지 확인된 경우	현행 유지
보완권고 (△)	백업 스케줄러는 정상 동작하나, 복구 시험 미수행, 일부 백업 대상 누락, 실패 알림 체계 미비 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	백업 스케줄러가 미구현되었거나 자동 실행되지 않고, 학습 데이터·모델 등 핵심 자산이 백업 대상에서 누락되어 데이터 손실 시 복구가 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '복구 시험' 수행 여부를 최우선으로 확인하라: 백업이 성공하더라도 복구가 안 되면 무의미하다. {테스트결과서}에 복구(Restore) 시험이 포함되어 있는지를 최우선으로 확인하고, 미수행 시 즉시 보완 권고한다. '백업 성공'만으로는 적합 판정이 어렵다.

2) 'AI 특화 백업 대상'의 누락 여부를 반드시 확인하라: 업무 DB 백업은 수행되었으나, AI 학습 데이터 파일, 모델 아티팩트(체크포인트), 벡터 DB 등 AI 특화 자산이 백업에서 누락되는 사례가 빈번하다. '백업·복구 설계서'의 대상 목록에 이들이 포함되어 있는지 반드시 별도로 확인한다.

3) '야간 윈도우 이내 완료'를 산술적으로 검증하라: {테스트결과서}의 백업 시작 시간과 종료 시간을 직접 확인하여, 전체 소요 시간이 야간 윈도우(예: 22:00~06:00 = 8시간) 이내인지 산술적으로 검증한다. 윈도우를 초과하면 주간 서비스에 영향을 미칠 수 있다.

4) 백업 실패 '알림 시험'의 수행 여부를 확인하라: 백업이 실패했을 때 아무도 인지하지 못하면, 장애 발생 시 복구할 백업이 없는 최악의 상황이 된다. {테스트결과서}에 의도적으로 백업을 실패시킨 후 알림이 발송되는지 확인하는 시험이 포함되어 있는지 점검한다.

5) 백업 보존 기간이 '무제한'이 아닌지 확인하라: 보존 정책이 정의되지 않으면 백업 데이터가 무한 축적되어 스토리지 비용이 폭증하고 관리가 불가능해진다. '백업·복구 설계서'에 보존 기간(예: 전체 30일, 증분 7일)과 만료 시 자동 삭제·아카이빙 정책이 정의되어 있는지 확인한다.

AG_SD44-3-06 | 악성노이즈-중독데이터 필터점검

데이터베이스 | 시험 | 데이터 수집·처리 | AI특화보안

근거 | NIST AI RMF · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 학습 데이터에 혼입될 수 있는 악성 노이즈(Malicious Noise) 및 데이터 중독(Data Poisoning) 공격 데이터를 탐지·차단하기 위한 필터링 스크립트가 설계대로 구현·동작하는지를 {테스트결과서}를 중심으로 확인하고, 특히 악성 라벨(Label)이 부여되었거나 의도적으로 왜곡된 패턴을 가진 데이터가 학습 파이프라인에 유입되지 않도록 차단되는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

악성 데이터 필터는 AI 모델의 안전성(Safety)과 보안성(Security)을 확보하기 위한 핵심 방어 메커니즘으로, 설계 시 수립한 필터링 규칙·탐지 기준이 스크립트로 구현되어 정상 동작하는지, 시험 결과가 산출물에 정량적으로 기록되었는지를 검증한다.

2. 필요성

▶ AI 시스템은 학습 데이터의 품질에 모델 동작이 직접적으로 좌우되므로, 의도적으로 오염된 데이터(Data Poisoning)가 학습에 사용되면 모델의 예측 결과가 공격자의 의도대로 왜곡될 수 있다. 특히 생성형 AI의 경우, 악성 데이터가 학습에 포함되면 유해 콘텐츠 생성, 개인정보 유출, 편향 증폭 등 심각한 사회적 위험을 초래할 수 있다.

▶ AI 시스템 안전성(Safety) 확보: NIST AI RMF는 AI 시스템의 안전성을 핵심 신뢰 특성으로 정의하며, 학습 데이터에 대한 의도적 공격(Data Poisoning)으로부터의 보호를 요구한다. 악성 데이터 필터가 동작하지 않으면 모델 자체가 공격 도구화될 수 있다.

▶ 악성 라벨 공격 방지: 정상 데이터에 잘못된 라벨(예: '정상'을 '사기'로 변조)을 부여하는 라벨 플리핑(Label Flipping) 공격은 분류 모델의 판단 기준 자체를 왜곡한다. 이를 탐지·차단하는 필터가 구현·검증되어야 한다.

▶ 백도어 트리거 데이터 차단: 특정 패턴(트리거)이 삽입된 데이터가 학습에 포함되면, 해당 패턴이 입력될 때만 공격자가 의도한 결과를 출력하는 '백도어(Backdoor)' 공격이 가능해진다. 이러한 이상 패턴 데이터의 탐지·차단이 필수적이다.

▶ 설계-구현 일관성 검증: 설계 단계에서 정의한 악성 데이터 탐지 규칙·필터링 기준이 구현 단계에서 스크립트로 정확히 구현되었는지 시험을 통해 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	악성 데이터 탐지 규칙의 필터 스크립트 구현 여부	라벨 이상·패턴 이상·통계적 이상 등 설계서 정의 악성 데이터 탐지 규칙이 필터 스크립트로 구현되었는지 '데이터 전처리 설계서' 규칙과 '필터링 스크립트 목록서'를 대조
2	악성 라벨 데이터 차단 시험 수행 여부	악성 라벨 데이터(Label Flipping 등) 차단 시험이 수행되었는지 {테스트결과서} 악성 라벨 차단 시험 항목에서 확인
3	이상 패턴 데이터 차단 시험 수행 여부	백도어 트리거 등 이상 패턴 데이터 차단 시험이 수행되었는지 {테스트결과서} 이상 패턴 차단 시험 항목에서 확인
4	필터 차단 건수·차단율 정량 기록 여부	필터 스크립트 실행 후 차단 건수·차단율이 '데이터 품질 검증 보고서'에 정량 기록되었는지 확인
5	정상 데이터 오차단 검증 수행 여부	정상 데이터가 필터에 의해 오차단(False Positive)되지 않았는지 {테스트결과서} 오차단율(FPR) 시험 항목에서 확인
6	필터 스크립트 파이프라인 자동 실행 여부	필터 스크립트가 데이터 파이프라인에 통합되어 자동 실행되는 구조인지 '필터링 스크립트 목록서' 또는 'ETL 설계서' 실행 시점·방식에서 확인
7	악성 데이터 필터 요구사항 추적 가능 여부	{요구사항추적표}의 악성 데이터 필터 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서의 악성 데이터 탐지 규칙이 필터 스크립트로 구현되고, 악성 라벨·이상 패턴 차단 시험과 오차단율(FPR) 검증이 수행되어 결과가 정량적으로 기록되었으며, 필터가 파이프라인에 자동 통합된 경우	현행 유지
보완권고 (△)	필터 스크립트는 구현되었으나, 일부 탐지 규칙 미구현, 오차단율 시험 미수행, 차단 결과의 정량 기록 미흡, 수동 실행 방식 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	악성 데이터 필터가 미구현되었거나 차단 시험이 전혀 수행되지 않아, 중독 데이터가 학습 파이프라인에 유입될 수 있는 상태인 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '시험 시나리오의 다양성'을 최우선으로 점검하라: {테스트결과서}에서 가장 중요한 것은 시험 시나리오가 설계서에 정의된 탐지 규칙을 충분히 커버하는지 여부이다. 라벨 이상, 패턴 이상, 통계적 이상 등 규칙 유형별로 최소 1개 이상의 시험 시나리오가 포함되어 있는지 확인한다.

2) '차단율 100%'도 의심하라: 모든 시험 시나리오에서 차단율이 100%로 보고된 경우, 시험에 사용된 악성 데이터가 너무 명확하여 실제 공격 상황을 반영하지 못할 가능성이 있다. 시험 데이터의 난이도(명확한 이상 vs. 경계선 이상)가 다양하게 구성되었는지 면담으로 확인한다.

3) '오차단율(FPR) 시험'의 수행 여부를 반드시 확인하라: 악성 데이터 차단만 시험하고 정상 데이터의 오차단 여부를 시험하지 않으면, 필터가 정상 데이터까지 과도하게 차단하여 학습 데이터 품질이 저하될 수 있다. FPR 시험이 수행되었는지, 허용 범위가 설계서에 정의되어 있는지 확인한다.

4) '수동 실행' 방식의 필터는 누락 위험이 높다: 필터 스크립트가 담당자의 수동 실행에 의존하는 경우, 데이터 적재 시 필터 실행이 누락되어 악성 데이터가 학습에 포함될 위험이 있다. 파이프라인 자동 통합 여부를 반드시 확인하고, 수동 실행인 경우 보완 권고한다.

5) 생성형 AI 사업에서의 추가 점검 포인트: 생성형 AI(LLM 등)에서는 프롬프트 인젝션(Prompt Injection), 탈옥(Jailbreak) 시도 데이터, 유해 콘텐츠 학습 데이터 등이 추가적인 악성 데이터 유형에 해당한다. 해당 유형에 대한 필터링 규칙과 시험이 포함되어 있는지를 추가로 확인한다.

AG_SD44-3-07 | 관리자-일반사용자 DB권한 통제점검

데이터베이스 | 시험 | 데이터 수집·처리 | 인증/인가/접근통제

근거 | 「개인정보 보호법」 · NIST AI RMF · ISO/IEC 27001 · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 관리자 계정과 일반 사용자 계정이 분리되어 각 계정별 DB 접근 권한이 최소 권한 원칙(Principle of Least Privilege)에 따라 설계·구현되었는지를 {데이터베이스} 관련 산출물을 중심으로 확인하고, 특히 일반 계정의 DML(INSERT, UPDATE, DELETE) 권한이 적절히 제한되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 학습 데이터, 모델 아티팩트, 추론 이력 등 고가치 데이터 자산을 보유하므로, DB 계정 권한이 적절히 통제되지 않으면 비인가 데이터 변경·삭제로 인한 학습 데이터 오염, 모델 무결성 훼손, 개인정보 유출 등 심각한 보안 사고가 발생할 수 있다.

2. 필요성

▶ AI 시스템의 데이터베이스에는 학습 데이터(원본·가공), 모델 메타정보, 추론 결과, 개인정보 등 다양한 민감 자산이 저장된다. 모든 사용자에게 동일한 권한을 부여하면 비인가 변경·삭제 위험이 증가하며, 특히 AI 학습 데이터의 무단 변경은 모델 성능에 직접적 영향을 미친다.

▶ AI 학습 데이터 무결성 보호: 일반 사용자(개발자, 분석가 등)에게 학습 데이터 테이블에 대한 INSERT/UPDATE/DELETE 권한이 부여되면, 의도치 않은 데이터 변경으로 학습 데이터가 오염될 수 있다. 학습 데이터 변경은 전담 관리자 또는 파이프라인 계정으로 제한되어야 한다.

▶ 최소 권한 원칙(Least Privilege) 준수: 각 계정에는 업무 수행에 필요한 최소한의 권한만 부여해야 한다. 이는 정보보안의 기본 원칙이며, NIST AI RMF와 정보시스템 감리기준 모두에서 요구하는 핵심 통제 항목이다.

▶ 개인정보 접근 통제: 개인정보가 포함된 테이블에 대한 접근 권한은 업무상 반드시 필요한 최소 계정으로 제한되어야 한다. 이는 개인정보 보호법 제29조의 안전조치의무에 해당한다.

▶ 감사 추적 및 책임 소재 명확화: 계정이 분리되어야 특정 데이터 변경이 어떤 계정(역할)에 의해 수행되었는지 추적할 수 있으며, 보안 사고 발생 시 책임 소재를 명확히 할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	역할별 DB 계정 분리 생성 여부	관리자·서비스·개발자·분석가 등 역할별 DB 계정이 분리 생성되었는지 'DB 계정 관리 대장' 또는 사업자 제출 계정 목록에서 확인

No	점검 항목	확인 방법
2	일반 계정 DML 권한 제한 여부	개발자·분석가 등 일반 계정의 DML(INSERT/UPDATE/DELETE) 권한이 제한되었는지 {테스트결과서} DML 차단 시험 결과에서 확인
3	권한 매트릭스대로 구현 여부	'보안 설계서'의 권한 매트릭스대로 권한이 구현되었는지 설계서 매트릭스와 실제 계정 권한을 대조
4	학습 데이터 테이블 변경 권한 제한 여부	AI 학습 데이터 테이블에 대한 변경 권한이 전담 계정(파이프라인/관리자)으로 제한되었는지 'DB 계정 관리 대장'에서 확인
5	공용 계정 금지·개인별 계정 발급 여부	공용 계정(Shared Account) 사용이 금지되고 개인별 계정이 발급되었는지 'DB 계정 관리 대장' 계정별 담당자에서 확인
6	DB 계정 비밀번호 정책 적용 여부	복잡도·변경 주기·잠금 등 DB 계정 비밀번호 정책이 설계서대로 적용되었는지 {테스트결과서} 비밀번호 정책 시험 결과에서 확인
7	권한 통제 요구사항 추적 가능 여부	{요구사항추적표}의 권한 통제 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	역할별 DB 계정이 분리 생성되고, 일반 계정의 DML 제한이 시험으로 확인되며, 권한 매트릭스와 실제 권한이 일치하고 1인 1계정 원칙과 비밀번호 정책이 적용된 경우	현행 유지
보완권고 (△)	계정 분리와 권한 통제가 대체로 구현되었으나, DML 차단 시험 미수행, 일부 계정의 권한 과다 부여, 비밀번호 정책 일부 미적용 등이 발견된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	관리자와 일반 사용자 계정이 분리되지 않았거나, 공용 계정이 사용되거나, 일반 계정에 학습 데이터 테이블 변경 권한이 부여되는 등 최소 권한 원칙이 지켜지지 않는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'DML 차단 시험'의 수행 여부를 최우선으로 확인하라: 권한 매트릭스가 설계서에 정의되어 있더라도, 실제로 비인가 DML이 차단되는지 시험으로 확인하지 않으면 구현의 정확성을 보장할 수 없다. {테스트결과서}에 일반 계정의 DML 실행 → 권한 오류 발생 시험이 포함되어 있는지를 최우선으로 점검한다.

2) 'AI 학습 데이터 테이블'의 권한을 별도로 확인하라: 일반 업무 테이블의 권한은 적절히 분리되었으나, AI 학습 데이터 테이블(training_dataset, feature_store 등)에 대해 개발자에게 DML 권한이 부여된 사례가 빈번하다. 'DB 계정 관리 대장'에서 AI 특화 테이블의 권한을 별도로 확인한다.

3) '공용 계정' 존재 여부를 반드시 확인하라: 'DB 계정 관리 대장'에서 담당자가 미지정이거나 복수 인원이 지정된 계정은 공용 계정일 가능성이 높다. 공용 계정은 책임 추적이 불가능하므로, 1인 1계정 원칙 위반 여부를 반드시 확인하고 보완 권고한다.

4) 추론 로그 테이블의 'DELETE 권한 미부여'를 확인하라: 추론 로그는 감사 추적의 핵심 근거이므로, DELETE 권한은 원칙적으로 어떤 계정에도 부여하지 않는 것이 권장된다. 로그 테이블에 DELETE 권한이 부여된 계정이 있으면 그 사유를 면담으로 확인한다.

5) 권한 매트릭스의 '테이블(스키마) 단위 분리'를 확인하라: 권한이 DB 단위로만 부여되고 테이블(스키마) 단위로 분리되지 않으면, 특정 역할에 불필요한 테이블까지 접근 가능해진다. 학습 데이터 스키마, 운영 데이터 스키마, 로그 스키마 등 스키마 단위로 권한이 분리되었는지 확인한다.

AG_SD44-3-08 | 파일-스토리지 민감컬럼 암호화점검

데이터베이스 | 시험 | 데이터 수집·처리 | 암호화

근거 | 「개인정보 보호법」 · NIST AI RMF · KCMVP · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 학습 데이터가 저장되는 파일 스토리지 및 데이터베이스 민감 컬럼에 대한 암호화가 AES-256 등 표준 알고리즘 기반으로 구현되어 정상적으로 암호화·복호화가 동작하는지를 {보안점검결과}를 중심으로 확인하고, 암호화 설계가 구현 단계에서 충실히 반영되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 학습 데이터에 개인정보·민감정보가 대량 포함되므로, 저장 시점(At-Rest)의 암호화는 데이터 유출 시 피해를 최소화하는 핵심 보호 조치이다.

특히 파일 스토리지(학습 데이터 파일, 모델 아티팩트 등)와 DB 민감 컬럼(주민등록번호, 카드번호 등) 모두에 대한 암호화 구현·동작 검증이 필수적이다.

2. 필요성

▶ AI 서비스는 학습·추론 과정에서 대규모 개인정보·민감정보를 처리·저장하며, 데이터 유출 시 피해 규모가 일반 정보시스템보다 크다. 저장 데이터의 암호화는 유출 사고 발생 시에도 데이터 내용을 보호하는 최후의 방어선이며, 개인정보 보호법과 관련 고시에서 의무화하고 있다.

▶ 개인정보 보호법 준수: 개인정보 보호법 제29조 및 '개인정보의 안전성 확보조치 기준'(개인정보보호위원회 고시)에 따라, 고유식별정보(주민등록번호 등), 비밀번호, 바이오정보 등은 반드시 암호화하여 저장해야 한다.

▶ AI 학습 데이터 보호: 학습 데이터 파일(CSV, Parquet, JSON 등)이 파일 스토리지에 평문으로 저장되면, 스토리지 접근 권한 탈취 시 대량 개인정보가 유출된다. 파일 수준 암호화(Storage-level Encryption)가 필수적이다.

▶ 표준 알고리즘 적용 검증: 자체 개발 암호화 알고리즘이나 취약한 알고리즘(DES, MD5 등) 사용은 보안 위험을 초래한다. NIST 승인 표준 알고리즘(AES-256, RSA-2048 이상 등)이 적용되었는지 확인이 필요하다.

▶ NIST AI RMF 보안 요건: NIST AI RMF는 AI 시스템의 보안(Secure)과 복원력(Resilient)을 핵심 신뢰 특성으로 정의하며, 저장 데이터 보호를 포함한 기술적 보안 조치를 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	암호화 대상 민감 컬럼 적용 여부	'보안 설계서' 정의 암호화 대상 민감 컬럼이 모두 암호화 되었는지 {보안점검결과} 암호화 적용 항목과 설계서를 대조
2	표준 암호화 알고리즘 적용 여부	적용된 암호화 알고리즘이 AES-256 등 표준 알고리즘인지 {보안점검결과} 알고리즘 명시 항목에서 확인
3	암호화 데이터 정상 복호화 시험 여부	암호화된 데이터의 정상 복호화가 시험으로 확인되었는지 {보안점검결과} 암·복호화 시험 결과에서 확인
4	파일 스토리지 암호화 적용 여부	AI 학습 데이터 파일이 저장되는 파일 스토리지에 암호화가 적용되었는지 {보안점검결과} 스토리지 암호화 항목에서 확인
5	암호키 관리 체계 수립·구현 여부	생성·저장·교체·폐기 등 암호키 관리 체계가 '암호키 관리 방안서'에 수립되고 구현되었는지 확인
6	암호화 후 성능 저하 SLA 이내 여부	암호화 적용 후 주요 쿼리의 성능 저하가 SLA 허용 범위 이내인지 {테스트결과서} 암호화 전후 성능 비교 결과에서 확인
7	암호화 요구사항 추적 가능 여부	{요구사항추적표}의 암호화 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 민감 컬럼과 파일 스토리지에 AES-256 등 표준 알고리즘 기반 암호화가 모두 적용되고, 암·복호화 시험과 성능 검증이 완료되었으며 암호키 관리 체계가 수립·구현된 경우	현행 유지
보완권고 (△)	암호화가 대체로 적용되었으나, 일부 대상 컬럼 누락, 파일 스토리지 암호화 미확인, 암호키 관리 방안 미흡 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	고유식별정보 등 법적 암호화 의무 대상이 평문으로 저장되어 있거나, 취약한 알고리즘(DES, MD5 등)이 사용된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 'AI 학습 테이블'의 민감 컬럼 암호화를 별도로 확인하라: 업무 테이블(고객_마스터 등)은 암호화되었으나, AI 학습 전용 테이블(학습_데이터셋, Feature Store 등)의 민감 컬럼은 암호화가 누락되는 사례가 빈번하다. AI 학습 테이블에 대한 별도 확인이 필수적이다.

2) '파일 스토리지'의 암호화를 반드시 점검하라: DB 컬럼 암호화만 확인하고 파일 스토리지의 암호화를 누락하는 경우가 많다. 학습 데이터 파일(CSV, Parquet 등)이 저장되는 NAS·S3·GCS 등에 암호화가 활성화되어 있는지 {보안점검결과}에서 별도로 확인한다.

3) 알고리즘명을 반드시 확인하라: {보안점검결과}에 '암호화 적용 완료'로만 기재되어 있으면, 적용된 알고리즘명(AES-256, ARIA-256, RSA-2048 등)을 반드시 확인한다. DES, 3DES, MD5, SHA-1 등은 취약한 알고리즘으로 분류되므로, 이러한 알고리즘이 사용된 경우 판정기준에 따라 부적합(시정조치) 검토 대상이다.

4) '암호키와 데이터의 분리 저장'을 확인하라: 암호키가 DB 테이블이나 애플리케이션 설정 파일에 평문으로 저장되면 암호화의 의미가 퇴색된다. 암호키가 별도 HSM, KMS, 키 서버 등에 분리 저장되는지 '암호키 관리 방안서'에서 확인한다.

5) 비밀번호는 '단방향 해시'인지 확인하라: 비밀번호는 복호화가 가능한 양방향 암호화(AES 등)가 아닌, 복호화 불가능한 단방향 해시(SHA-256 + Salt, bcrypt, scrypt 등)로 저장되어야 한다. {보안점검결과}에서 비밀번호의 암호화 방식이 단방향 해시인지 별도로 확인한다.

AG_SD44-3-09 | 성능평가 마트 OLAP 집계검증 대시보드 통계 정확성점검

데이터베이스 | 시험 | 데이터 수집·처리 | 모니터링/재학습

근거 | NIST AI RMF MEASURE · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 성능 평가 결과를 집계·분석하기 위해 구축된 성능 평가용 마트(OLAP) 테이블의 집계 데이터가 원천 데이터와 정합성을 유지하고 있는지, 그리고 실시간 모니터링 대시보드에 표출되는 통계 수치가 원천 집계 데이터를 정확하게 반영하고 있는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

AI 시스템에서는 모델 추론 결과의 정확도, 응답 시간, 처리량, 오류율 등 다양한 성능 지표가 주기적으로 수집·집계되어야 하며, 이러한 집계 데이터가 의사결정에 활용되는 대시보드에 왜곡 없이 전달되는지를 확인한다.

집계 과정에서의 데이터 누락, 중복 집계, 집계 기준 불일치, 시간대(Timezone) 변환 오류 등이 발생하면 AI 시스템의 성능을 잘못 판단하게 되어 배포 적합성 판정에 심각한 오류를 초래할 수 있다.

2. 필요성

▶ AI 성능 판정의 객관성 확보: AI 모델의 배포(Go/No-Go) 판정은 성능 지표 집계 결과에 근거한다. 마트 테이블의 집계가 부정확하면, 실제로는 성능 기준을 미달하는 모델이 배포되거나, 반대로 적합한 모델이 부당하게 탈락할 수 있다.

▶ 데이터 파이프라인 정합성 보장: AI 시스템의 성능 데이터는 추론 로그에서 ETL, 마트(OLAP), 대시보드의 다단계 파이프라인을 거친다. 각 단계에서 집계 기준(시간 단위, 필터 조건, 집계 함수 등)이 일관되게 적용되어야 최종 통계 수치의 신뢰성이 확보된다.

▶ 실시간 모니터링 신뢰성: 운영 배포 후 AI 시스템의 이상 탐지, 성능 저하 조기 경고 등은 대시보드 수치에 의존한다. 대시보드 수치가 원천과 불일치하면 장애 대응 시점을 놓치거나 불필요한 경보(False Alarm)가 발생하여 운영 효율성이 저하된다.

▶ 감사 추적성(Audit Trail) 확보: AI 시스템의 성능 이력은 향후 모델 재학습, 편향 검토, 규제 대응 시 핵심 근거 자료가 된다. 집계 데이터의 정확성이 보장되지 않으면 감사 추적의 신뢰성 자체가 훼손된다.

3. 검토방법

No	점검 항목	확인 방법
1	마트(OLAP) 스키마 설계 적정성	성능 평가용 마트 테이블의 차원(Dimension)·측정값(Measure) 정의가 AI 성능 KPI 항목과 일치하는지 확인
2	집계 로직(ETL) 정합성	추론 결과 원천 테이블에서 마트로의 ETL 변환 시 집계 함수, 필터 조건, 시간 단위가 요구사항과 일치하는지 확인
3	원천-마트 데이터 건수 일치	특정 기간·조건에 대해 원천 테이블의 건수와 마트 테이블의 집계 건수가 정합하는지 샘플링 검증
4	집계 수치 교차 검증	마트 테이블의 주요 집계 값을 원천 데이터로부터 직접 산출한 값과 비교하여 오차 범위 이내인지 확인
5	시간대(Timezone) 처리 일관성	추론 로그 타임스탬프, 마트 집계 기준 시간, 대시보드 표출 시간이 동일 시간대 기준으로 일관 처리되는지 확인
6	대시보드 표출 수치 일치	대시보드 화면 표출 통계 수치가 마트 테이블의 집계 결과와 동일한지 화면 캡처 대조로 확인
7	대시보드 갱신 주기 적정성	대시보드 갱신 주기(실시간/준실시간/배치)가 설계서 정의대로 동작하며 허용 지연 시간 이내인지 확인
8	예외 데이터 처리 관리	ETL 과정의 NULL, 이상치, 중복 레코드 등 예외 데이터에 대한 처리 규칙 정의 및 적용 여부 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	마트(OLAP) 스키마의 차원·측정값 정의가 AI 성능 KPI와 일치하고, 집계 로직(집계 함수·필터 조건·시간 단위)이 요구사항과 정합하며, 마트 집계 데이터가 원천 데이터와 정합하고, 대시보드 표출 수치가 마트 데이터와 일치하며 갱신 주기가 설계대로 동작하고, 예외 데이터 처리 규칙이 정의·적용되어 관련 시험 결과가 문서화된 경우 (예: 샘플링 검증 결과 오차 없음, 대시보드 수치-마트 일치 확인, Timezone 처리 일관성 확인)	현행 유지
보완권고 (△)	마트 집계 로직은 정의되어 있으나, 일부 KPI의 정합성 시험이 누락되었거나, 대시보드 표출 수치의 검증이 불완전한 경우 (예: 일부 KPI 교차 검증 미실시, 대시보드 갱신 주기 미검증, 예외 데이터 처리 규칙 미문서화)	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	마트 테이블 설계가 성능 KPI를 반영하지 않았거나, 원천-마트 간 데이터 불일치가 확인되었거나, 대시보드 수치가 마트 데이터와 상이한 경우 (예: 마트 테이블에 핵심 KPI 누락, SQL 검증 시 원천-마트 건수 불일치, 대시보드 수치 왜곡 확인)	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 마트 설계와 KPI 정의의 연결고리를 반드시 확인하라: 마트 테이블이 존재하더라도 정작 요구사항에 정의된 핵심 KPI가 누락된 경우가 빈번하다. 'AI 모델 성능 기준 정의서'의 KPI 목록과 마트 테이블의 Measure 컬럼을 1:1로 대조하는 체크리스트를 작성하면 누락 항목을 쉽게 발견할 수 있다.

2) SQL 샘플링 검증 시 경계값(Boundary)에 집중하라: 대부분의 집계 오류는 일자 변경 시점, 배치 실행 직전/직후, 시간대 전환 시점 등 경계 구간에서 발생한다. 자정(00:00) 전후, ETL 배치 실행 시점 전후의 데이터를 우선적으로 샘플링하면 잠재적 오류를 효과적으로 탐지할 수 있다.

3) 대시보드 검증 시 동일 시점을 명확히 고정하라: 대시보드와 마트 테이블의 수치를 비교할 때, 대시보드 캡처 시점과 SQL 조회 시점 사이에 데이터가 갱신되면 불일치가 발생할 수 있다. 반드시 대시보드 캡처 시점의 타임스탬프를 기록하고, SQL WHERE 조건에 동일 시점을 명시하여 비교해야 한다.

4) ETL 오류 처리 규칙의 문서화 여부를 반드시 확인하라: NULL 값이나 이상치를 자동으로 제외하는 로직이 코드에만 존재하고 문서화되지 않은 경우, 향후 모델 재학습이나 성능 재평가 시 동일한 집계를 재현할 수 없게 된다. 모든 예외 처리 규칙은 'ETL 매핑 정의서'에 명시되어야 하며, 이를 누락한 경우 보완 권고한다.

5) 과도한 기술적 검증보다는 설계-구현 일치 관점에 집중하라: 감리원이 ETL 코드를 직접 리뷰하거나 대시보드의 내부 쿼리를 분석하는 것은 과도한 기술적 개입이다. 감리의 핵심은 설계서에 정의된 집계 규칙이 실제 결과물에 올바르게 반영되었는지를 입력(원천)-출력(마트/대시보드) 비교를 통해 확인하는 것이다. 코드 수준의 검증은 사업자의 책임이며, 감리원은 결과의 정합성을 검증한다.

AG_SD44-3-10 | 샘플링클래스 불균형 해소점검

데이터베이스 | 시험 | 데이터 수집·처리 | 학습데이터

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 학습 데이터의 클래스 불균형(Class Imbalance) 문제를 해소하기 위해 설계 단계에 정의된 샘플링 알고리즘(오버샘플링, 언더샘플링 등)이 실제로 적용되었는지 여부를 '전처리결과서'를 중심으로 확인하고, 샘플링 적용 후 클래스 비율이 설계 목표에 부합하는 수준으로 균형을 달성하였는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

클래스 불균형 해소는 AI 모델의 공정성과 예측 정확도에 직접적 영향을 미치는 핵심 전처리 활동으로, 설계 시 수립한 샘플링 기법·목표 비율이 충실히 적용되었는지, 적용 전후의 클래스별 데이터 건수·비율이 산출물에 정량적으로 기록되었는지를 검증한다.

2. 필요성

▶ AI 시스템의 학습 데이터에서 특정 클래스(예: 정상/이상, 긍정/부정)의 데이터 건수가 현저히 불균형한 상태로 모델 학습이 수행되면, 소수 클래스(Minority Class)에 대한 예측 성능이 극단적으로 저하되는 문제가 발생한다. 특히 이상 탐지, 사기 감지, 질병 진단 등 소수 클래스가 핵심인 AI 서비스에서는 불균형 해소가 필수적이다.

▶ 모델 공정성 확보: 클래스 불균형이 해소되지 않으면 다수 클래스 편향(Majority Bias)이 발생하여, 모델이 소수 클래스를 거의 예측하지 못하는 문제가 발생한다. 이는 NIST AI RMF가 경고하는 통계적·계산적 편향(Computational and Statistical Bias)에 해당한다.

▶ 소수 클래스 예측 정확도 보장: 오버샘플링(SMOTE 등)이나 언더샘플링을 통해 클래스 비율을 조정하지 않으면, 모델의 전체 정확도(Accuracy)는 높아 보이지만 실제로 중요한 소수 클래스의 재현율(Recall)이 현저히 낮아지는 '정확도 역설(Accuracy Paradox)' 현상이 발생할 수 있다.

▶ 설계-구현 일관성 검증: 설계 단계에서 정의한 샘플링 기법과 목표 비율이 구현 단계에서 정확히 적용되었는지 확인하지 않으면, 설계와 구현 간 괴리로 인해 모델 성능이 기대에 미치지 못할 수 있다.

▶ 재현성 및 감사 추적성 보장: 샘플링 적용 과정(기법, 대상 클래스, 적용 전후 건수, 사용 파라미터 등)이 문서화되지 않으면 동일 실험의 재현이 불가능하며, 제3자 감사 시 검증 근거가 부재하게 된다.

3. 검토방법

No	점검 항목	확인 방법
1	샘플링 기법 적용 여부	SMOTE·Random Oversampling·Undersampling 등 '데이터 전처리 설계서' 정의 샘플링 기법이 실제로 적용되었는지 '전처리결과서' 적용 기법과 대조
2	샘플링 전후 클래스별 건수 정량 기록 여부	샘플링 적용 전후 클래스별 데이터 건수가 '전처리결과서'에 정량 기록되었는지 확인
3	샘플링 후 클래스 비율 목표 부합 여부	샘플링 적용 후 클래스 비율이 설계서 목표 비율에 부합하는지 '전처리결과서' 최종 비율과 대조
4	샘플링 주요 파라미터 기록 여부	k값·배율·시드 등 샘플링 기법 주요 파라미터가 '전처리결과서'에 기록되었는지 확인
5	합성 데이터 품질 검증 여부	오버샘플링으로 생성된 합성 데이터의 품질 검증이 수행되었는지 '데이터 품질 검증 보고서' 합성 데이터 분포 확인 항목에서 검토
6	샘플링의 데이터 분리 후 수행 여부	샘플링 적용이 학습/검증/시험 데이터 분리 후에 수행되었는지(Data Leakage 방지) '전처리결과서' 처리 순서 기술에서 확인
7	클래스 불균형 해소 요구사항 추적 가능 여부	{요구사항추적표}의 클래스 불균형 해소 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 샘플링 기법이 실제 적용되고, 적용 전후 클래스별 건수·비율과 주요 파라미터가 정량적으로 기록되었으며, 오버샘플링으로 생성된 합성 데이터의 품질 검증이 수행되고, 목표 비율 달성과 Split→Sampling 처리 순서(Data Leakage 방지)까지 확인되는 경우	현행 유지
보완권고 (△)	샘플링은 적용되었으나 전후 건수·파라미터 기록이 미흡하거나, 목표 비율과의 편차가 허용 범위를 벗어나거나, 합성 데이터 품질 검증·요구사항 추적 등 일부 항목이 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	설계서에 정의된 샘플링 기법이 적용되지 않아 클래스 불균형이 해소되지 않았거나, 샘플링을 데이터 분리 전 전체 데이터에 적용하여 Data Leakage가 발생하는 등 검증 근거가 부재한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 문서 검토와 병행하여 DB 내 클래스별 분포 쿼리를 요청하라: 가이드라인에서는 문서 기반 수작업을 권장하나, 종료단계 감리원이 데이터베이스 영역을 맡았다면 사업자에게 최종 학습 데이터 테이블을 대상으로 한 Group By 카운트 쿼리(SELECT class_label, COUNT(*) FROM training_table GROUP BY class_label;) 실행 결과를 현장에서 시연하도록 요구하여 산출물과 실제 DB 간의 무결성을 입증받아야 한다.

2) 'Data Leakage' 여부를 반드시 확인하라: 샘플링이 전체 데이터에 먼저 적용된 후 학습/시험으로 분리되었다면, 시험 결과가 과대 평가되어 실제 운영 시 성능이 크게 저하될 수 있다. '전처리 결과서'에서 ① Split → ② Sampling 순서가 명시되어 있는지 확인하고, 불명확한 경우 사업자 면담으로 확인한다.

3) '1:1 균형'이 항상 최적은 아님을 인지하라: 클래스 비율을 반드시 1:1로 맞출 필요는 없다. 도메인 특성에 따라 2:1, 3:1 등 다양한 비율이 최적일 수 있으므로, 감리원은 목표 비율 자체의 타당성보다 '설계서에 정의된 목표 비율과 실제 달성 비율의 일치 여부'를 중심으로 점검한다.

4) 다중 클래스 불균형도 동일 프레임으로 점검하라: 2개 클래스(이진 분류)뿐 아니라 3개 이상 클래스(다중 분류)에서도 불균형이 발생할 수 있다. 이 경우 클래스별로 건수와 비율을 개별 확인하고, 각 클래스의 목표 비율 달성 여부를 점검한다.

5) 설계서에 샘플링 기법이 명시되지 않은 경우의 대응: '데이터 전처리 설계서'에 클래스 불균형 해소 기법이 별도로 정의되지 않은 사업의 경우, {요구사항정의서}나 '데이터 구축 계획서'에서 관련 요구사항을 확인한다. 클래스 불균형이 존재하는데도 어떠한 해소 방안도 적용되지 않은 경우, 설계 단계 누락으로 판단하여 보완 권고를 할 수 있다.

AG_SD44-3-11 | 데이터 모델ERD물리DB 일치점검

데이터베이스 | 시험 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에서 수립된 데이터 모델(ERD)이 실제 물리 데이터베이스에 정확하고 완전하게 구현되었는지를 감리원이 제3자적 관점에서 검증하는 것을 목적으로 한다.

특히, AI 서비스에 특화된 학습 데이터 저장 테이블, 모델 추론 결과 테이블, 벡터 DB 연동 테이블, 프롬프트·응답 이력 테이블 등이 설계대로 생성되었는지, 그리고 참조무결성·제약조건(PK, FK, NOT NULL, CHECK, UNIQUE 등)이 설계서에 정의된 대로 100% 적용되었는지를 확인한다.

2. 필요성

▶ AI 서비스 구축 사업에서 데이터 모델과 물리 DB의 불일치는 시스템 안정성과 데이터 신뢰성에 심각한 영향을 미칠 수 있다. 특히 AI 시스템은 일반 정보시스템과 달리 아래와 같은 고유한 위험 요소를 내포하고 있어, 종료 단계에서의 정밀 점검이 필수적이다.

▶ AI 데이터 정합성 보장: 학습 데이터, 추론 결과, 피드백 이력 등 AI 파이프라인 전반에 걸친 데이터의 무결성이 물리 DB 수준에서 보장되지 않으면, 모델 재학습 시 오염된 데이터가 유입되어 AI 서비스 품질이 저하될 수 있다.

▶ 설계-구현 간 추적성 확보: ERD에 정의된 AI 특화 테이블(임베딩 저장, 토큰 사용량 로그, 편향 검출 결과 등)이 물리 DB에 누락되거나 변형되면, 설계 의도가 훼손되고 감리 시 추적이 불가능해진다.

▶ 참조무결성 미적용 위험: AI 시스템은 다수의 테이블 간 복잡한 참조 관계(학습 데이터셋 → 라벨링 이력 → 모델 버전 등)를 가지며, FK 제약조건이 누락되면 고아 레코드(Orphan Record)가 발생하여 데이터 품질이 심각하게 저하된다.

▶ 운영 안정성 사전 검증: 종료 단계에서 물리 DB의 설계 일치 여부를 검증하지 않으면, 운영 단계 진입 후 데이터 정합성 오류가 누적되어 장애 발생 시 원인 추적이 극히 어려워진다.

3. 검토방법

No	점검 항목	확인 방법
1	ERD-물리 테이블 일치	설계 ERD의 모든 엔터티가 물리 DB에 테이블로 생성되었는지, 테이블명·스키마가 일치하는지 확인

No	점검 항목	확인 방법
2	AI 특화 테이블 구현 완전성	학습 데이터셋, 모델 버전 관리, 추론 결과, 프롬프트/응답 이력, 임베딩 벡터, 편향 검출 결과 등 AI 특화 테이블의 누락 여부 점검
3	컬럼 정의 일치	각 테이블의 컬럼명, 데이터타입, 길이, NULL 허용 여부가 {테이블정의서}와 물리 DB DDL 간 100% 일치하는지 확인
4	PK/FK 제약조건 구현	설계된 기본키(PK), 외래키(FK) 제약조건이 물리 DB에 모두 생성되었는지 확인 (AI 테이블 간 참조관계 포함)
5	CHECK/UNIQUE/DEFAULT 제약조건	설계서에 정의된 CHECK, UNIQUE, DEFAULT 제약조건이 누락 없이 적용되었는지 점검
6	인덱스 설계 반영	설계서에 정의된 인덱스(일반, 복합, 부분 인덱스 등)가 물리 DB에 생성되었는지, AI 조회 패턴에 적합한지 확인
7	AI 데이터 파이프라인 연계 정합성	ETL/데이터 파이프라인 설계서의 소스-타겟 매핑이 물리 테이블 구조와 일치하는지 확인
8	시험 결과 기반 데이터 무결성 확인	DB 관련 단위·통합 시험에서 데이터 정합성(무결성 위반 건수 0건) 확인 결과를 점검

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	ERD의 모든 엔터티가 물리 DB에 테이블로 생성되고, 컬럼 정의·제약조건과 설계된 인덱스가 설계서와 100% 일치하며, AI 특화 테이블이 완전히 구현되고, ETL·데이터 파이프라인 설계서의 소스-타겟 매핑이 물리 테이블 구조와 일치하며, 단위·통합 시험에서 데이터 무결성 위반 0건이 확인된 경우	현행 유지 (사례: ERD 50개 엔터티 전수 대조 일치, FK 30건 전수 적용, AI 특화 테이블 12개 전수 구현 확인)
보완권고 (△)	대부분의 테이블이 구현되었으나, 일부 컬럼 정의 불일치 (데이터타입 변경, NULL 허용 변경 등) 또는 일부 제약조건 누락이 발견된 경우	해당 항목 보완 요청 후 재확인 (사례: 컬럼 5건 데이터타입 불일치, FK 3건 미적용, CHECK 제약 2건 누락)
부적합 (X)	ERD 대비 테이블 누락, AI 핵심 테이블 미구현, 또는 PK/FK 등 핵심 제약조건이 다수 누락되어 설계 의도가 훼손된 경우	시정조치 요구 및 이행 결과 확인 (사례: AI 특화 테이블 3개 이상 누락, FK 50% 이상 미적용, 설계서에 없는 테이블 다수 추가)

5. 감리 시 유의사항 및 실무 Tip

1) AI 특화 테이블 체크리스트를 사전에 준비하라: 일반 업무 테이블과 달리, AI 시스템에는 학습 데이터 관리, 모델 버전 관리, 추론 결과, 프롬프트/응답 이력, 임베딩 벡터, 편향 검출, 토큰 사용량 로그 등 AI 고유 테이블이 존재한다. 감리 착수 전 설계서를 기반으로 AI 특화 테이블 체크리스트를 작성해 두면, 현장에서 누락 여부를 체계적으로 점검할 수 있다.

2) DDL 스크립트 제출을 반드시 요구하라: 사업자가 '물리 DB를 ERD대로 구현했다'고 구두 설명하는 것만으로는 적합 판정이 불가하다. 반드시 실제 적용된 DDL 스크립트(CREATE TABLE, ALTER TABLE 등)를 산출물로 제출받아 설계서와 건별 대조한다. DDL이 없는 경우, DB 카탈로그 조회 결과 화면 캡처라도 요구해야 한다.

3) FK 미적용 사유에 대한 '성능' 주장을 경계하라: 사업자가 FK 제약조건을 '성능상의 이유'로 적용하지 않았다고 주장하는 경우가 빈번하다. 그러나 설계서에 FK가 정의되어 있는데 물리 DB에 미적용된 것은 설계-구현 불일치이므로, 반드시 변경 사유서와 대안(예: 애플리케이션 레벨 무결성 검증 로직)의 존재 여부를 확인한다. 대안이 없다면 부적합 판정을 검토한다.

4) ERD 버전과 DDL 버전의 동기화 여부를 확인하라: 개발 과정에서 ERD는 수정되었으나 DDL에 반영되지 않았거나, 반대로 DDL만 변경되고 ERD가 미갱신된 경우가 흔하다. 감리 시 ERD의 최종 수정일자와 DDL 스크립트의 최종 수정일자를 비교하여, 양쪽이 동기화되어 있는지 반드시 확인한다.

5) 벡터 DB와 관계형 DB 간 연계 구조에 주의하라: 생성형 AI 서비스에서는 관계형 DB(PostgreSQL, MySQL 등) 외에 벡터 DB(Milvus, Pinecone, pgvector 등)를 병행 사용하는 경우가 많다. 이 경우 양쪽 DB 간 데이터 매핑 키(예: 문서ID, 임베딩ID)가 설계서에 정의되어 있고 물리적으로도 구현되었는지 확인한다.

AG_SD44-3-12 | 실시간 스트리밍적재 파이프라인점검

데이터베이스 | 시험 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · NIST AI RMF MEASURE·MANAGE

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 카프카(Kafka) 등 메시지큐 기반 실시간 스트리밍 파이프라인이 설계대로 구현·동작하여 데이터가 지연(Latency) 및 유실(Loss) 없이 데이터베이스에 적재되는지를 {통합테스트} 결과를 중심으로 확인하고, 특히 메시지 전달 보장(Delivery Guarantee), 적재 지연 시간(Latency), 메시지 유실률 0% 달성 여부를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템에서 실시간 스트리밍 파이프라인은 사용자 행동 데이터, 센서 데이터, 로그 데이터 등을 실시간으로 수집·적재하여 모델 추론 또는 재학습에 활용하는 핵심 인프라이다.

파이프라인 장애 시 데이터 유실이나 지연이 발생하면 AI 서비스의 실시간성과 정확도가 직접적으로 훼손되므로, 통합 시험을 통해 파이프라인의 안정성을 검증한다.

2. 필요성

▶ 실시간 AI 서비스(추천, 이상 탐지, 실시간 분류 등)는 데이터의 실시간 수집·적재에 전적으로 의존하므로, 스트리밍 파이프라인의 장애는 곧 AI 서비스 장애를 의미한다. 특히 메시지 유실은 학습 데이터의 완전성을 훼손하고, 과도한 적재 지연은 실시간 추론의 정확도를 저하시킨다.

▶ 메시지 유실 방지: 카프카 등 메시지큐를 통해 전달되는 데이터가 유실 없이 데이터베이스에 적재되어야 AI 학습 데이터의 완전성이 보장된다. 유실된 데이터는 복구가 어려우며, 편향된 학습 결과를 초래할 수 있다.

▶ 적재 지연 최소화: 실시간 AI 추론에 사용되는 데이터의 적재가 지연되면, 모델이 과거 데이터를 기반으로 추론하여 정확도가 저하된다. 설계서에 정의된 SLA(Service Level Agreement) 기준의 지연 시간 준수가 필수적이다.

▶ 장애 복구(Fault Tolerance) 검증: 파이프라인 구성 요소(Producer, Broker, Consumer) 중 일부에 장애가 발생했을 때 데이터 유실 없이 복구되는지 시험을 통해 확인해야 한다.

▶ 설계-구현 일관성 검증: 설계 단계에서 정의한 스트리밍 아키텍처(토픽 구조, 파티션 수, 소비자 그룹 등)가 구현 단계에서 정확히 반영되었는지 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	메시지 발행-적재 건수 100% 일치 여부	메시지 발행 건수와 DB 적재 건수가 100% 일치(유실률 0%)하는지 {통합테스트} 발행-적재 건수 비교 결과에서 확인
2	적재 지연 시간 SLA 이내 여부	평균·P95·P99 적재 지연 시간(Latency)이 각각 기재되어 설계서 정의 SLA 기준 이내인지 {통합테스트}에서 대조
3	브로커 장애 시 메시지 유실 없는 복구 여부	브로커 장애(Failover) 시 메시지 유실 없이 복구되는지 {통합테스트} 장애 복구 시험 결과에서 확인
4	메시지 중복 적재 방지 동작 여부	메시지 중복 적재 방지 처리(Deduplication)가 동작하는지 {통합테스트} 중복 적재 시험 결과에서 확인
5	토픽 구조 구현 여부	토픽명·파티션 수·소비자 그룹 등 설계서 정의 토픽 구조가 구현되었는지 '데이터 파이프라인 설계서'와 실제 구성을 대조
6	부하 시험 목표 처리량(TPS) 달성 여부	부하(Load) 시험 시 목표 처리량(TPS)이 달성되는지 {통합테스트} 부하 시험 결과와 '비기능 요구사항'을 대조
7	실시간 적재 요구사항 추적 가능 여부	{요구사항추적표}의 실시간 적재 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	발행 건수와 적재 건수가 100% 일치(유실률 0%)하고, 평균·P95·P99 지연이 SLA 이내이며, 장애 복구(Failover)·중복 방지 시험이 통과되고, 토픽 구조가 설계서와 일치하며 목표 TPS가 달성된 경우	현행 유지
보완권고 (△)	핵심 시험은 수행되었으나 P95/P99 지연 미측정, 장애 시나리오·부하 시험 일부 누락, 시험 규모가 운영 환경을 대표하지 못하는 등 검증의 충분성이 미흡한 경우	해당 항목 보완(재시험) 요청 후 재확인
부적합 (X)	메시지 유실이 확인되거나 지연 시간이 SLA를 초과하는데도 조치가 없는 경우, 또는 스트리밍 적재에 대한 통합 시험 자체가 수행되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '발행 건수 = 적재 건수'의 산술 검증을 최우선으로 수행하라: {통합테스트}에서 가장 핵심적인 점검은 메시지 발행 건수와 DB 적재 건수의 100% 일치 여부이다. 두 수치의 차이를 직접 계산하고, 차이가 0건인지 확인하는 것만으로도 유실률 0%를 빠르게 검증할 수 있다.

2) '평균 지연'뿐 아니라 'P95/P99 지연'을 반드시 확인하라: 평균 지연이 SLA 이내여도 P99 지연이 극단적으로 높으면 실시간 AI 서비스에 간헐적 장애가 발생한다. {통합테스트}에 P95/P99 지연이 측정·기록되어 있는지 확인하고, 없으면 보완을 요청한다.

3) '장애 복구 시험'의 수행 여부를 반드시 확인하라: 정상 부하에서의 시험만으로는 파이프라인의 안정성을 보장할 수 없다. 브로커 장애, 컨슈머 장애 등 장애 시나리오별 시험이 {통합테스트}에 포함되어 있는지 확인하고, 미수행 시 보완 권고한다.

4) '시험 규모'가 운영 환경을 대표하는지 확인하라: 시험 시 100건만 발행하고 유실 0건을 보고하는 것은 충분한 검증이 아니다. 운영 예상 일일 건수 이상의 규모로, 일정 시간(최소 30분) 이상 지속적으로 수행된 시험인지 확인한다.

5) 카프카가 아닌 다른 메시지큐를 사용하는 경우에도 동일 프레임을 적용하라: RabbitMQ, AWS Kinesis, GCP Pub/Sub 등 다른 메시지큐를 사용하는 사업에서도 상세 검토 항목(유실률 0%, 지연 SLA, 장애 복구, 중복 방지)은 동일하게 적용된다. 기술 스택이 다르더라도 점검 관점은 동일하므로 본 가이드라인을 그대로 활용한다.

AG_SD44-3-13 | 인덱스-파티셔닝 효율성 점검

데이터베이스 | 시험 | 데이터 수집·처리 | 학습추론 및 저장

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「전자정부법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 대용량 AI 학습·추론 데이터를 처리하는 데이터베이스의 인덱스(Index) 및 파티셔닝(Partitioning) 설계가 구현 단계에서 효율적으로 적용되었는지를 {성능테스트결과서}를 중심으로 확인하고, 특히 대용량 쿼리에서 Full Table Scan이 억제되고 파티션 히트(Partition Hit)가 정상 동작하는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 수억~수십억 건 규모의 학습 데이터를 처리하므로, 인덱스와 파티셔닝이 적절히 구현되지 않으면 쿼리 응답 시간이 수 분~수 시간으로 증가하여 AI 서비스의 실시간 추론 및 배치 학습 모두에 치명적 성능 저하를 초래한다.

2. 필요성

▶ AI 시스템의 데이터베이스는 일반 OLTP 시스템과 비교하여 데이터 규모가 수십~수백 배 크고, 대량 범위 조회(Range Scan), 집계(Aggregation), 조인(Join) 쿼리가 빈번하게 발생한다. 인덱스와 파티셔닝이 적절히 설계·구현되지 않으면 매 쿼리마다 전체 테이블을 스캔(Full Scan)하여 심각한 성능 저하를 초래한다.

▶ AI 추론 응답 시간 보장: 실시간 AI 추론 서비스에서 Feature 조회, 임베딩 검색 등의 쿼리가 Full Scan으로 수행되면 응답 지연이 발생하여 SLA를 충족할 수 없다. 인덱스를 통한 빠른 검색이 필수적이다.

▶ 배치 학습 파이프라인 성능 확보: 수억 건의 학습 데이터에서 특정 기간·조건의 데이터를 추출하는 배치 쿼리에서 파티션 히트가 동작하지 않으면, 불필요한 파티션까지 스캔하여 학습 파이프라인 전체가 지연된다.

▶ 시스템 자원 효율성: Full Scan은 CPU, I/O, 메모리를 과도하게 소비하여 다른 서비스에도 영향을 미친다. 인덱스/파티셔닝을 통한 데이터 접근 범위 최소화는 전체 시스템 안정성의 기반이다.

▶ 설계-구현 일관성 검증: 설계 단계에서 정의한 인덱스 전략·파티셔닝 키·파티션 구조가 구현 단계에서 정확히 반영되었는지 성능 시험으로 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	주요 대용량 쿼리 Full Table Scan 미발생 여부	주요 대용량 쿼리에서 Full Table Scan이 발생하지 않는지 {성능테스트결과서} 실행 계획(Execution Plan) 분석 결과에서 확인
2	파티션 히트 정상 동작 여부	파티셔닝된 테이블에서 파티션 히트(Partition Pruning)가 정상 동작하는지 {성능테스트결과서} 파티션 접근 분석 결과에서 확인
3	주요 쿼리 응답 시간 SLA 이내 여부	주요 쿼리의 응답 시간이 '비기능 요구사항'의 SLA 이내인지 {성능테스트결과서} 응답 시간과 대조
4	설계 인덱스 실제 DB 생성 여부	'데이터베이스 설계서' 정의 인덱스가 실제 DB에 생성되었는지 설계서 인덱스 목록과 실제 DB 카탈로그를 대조
5	파티셔닝 키·분할 기준 적용 여부	설계서 정의 파티셔닝 키·분할 기준이 실제 DB에 적용되었는지 설계서 파티션 정의와 실제 파티션 구조를 대조
6	AI 학습 배치 쿼리 파티션 히트 동작 여부	AI 학습 배치 쿼리(대량 범위 조회)에서 파티션 히트가 동작하는지 {성능테스트결과서} 배치 쿼리 분석 결과에서 확인
7	DB 성능 요구사항 추적 가능 여부	{요구사항추적표}의 DB 성능 관련 요구사항이 구현 완료로 추적되는지 해당 항목 상태에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서에 정의된 인덱스·파티셔닝이 실제 DB에 구현되고, 주요 대용량 쿼리에서 Full Scan 없이 파티션 히트가 동작하며 AI 학습 배치 쿼리에서도 파티션 히트가 확인되고, 응답 시간이 SLA를 충족함이 성능 시험으로 확인된 경우	현행 유지
보완권고 (△)	인덱스·파티셔닝은 대체로 적용되었으나, 일부 쿼리의 Full Scan 발생, 일부 인덱스 누락, 배치 쿼리 성능 시험 미수행 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	설계된 인덱스·파티셔닝이 대부분 미구현되었거나, 주요 쿼리가 Full Scan으로 수행되어 SLA를 크게 초과하는 등 성능 요건을 충족하지 못하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) DB 딕셔너리를 조회하여 파티션 테이블 구성 여부(Range, Hash 등)를 확인하고, 주요 AI 추출 쿼리의 실행 계획(Explain)을 떠서 Partition Hit 동작을 물리적으로 확인한다. '문서 내 키워드 검색'이 아닌, 실제 DB 성능 뷰에서 Full Scan 쿼리를 추출하여 검증하는 것이 올바른 감리 방법이다.

2) 파티션 히트를 '접근 파티션 수'로 간단히 확인하라: {성능테스트결과서}에 파티셔닝 테이블 쿼리의 '접근 파티션 수'가 기재되어 있으면, 이 수치가 전체 파티션 수보다 현저히 적은지 확인한다. 예: 전체 12개 월별 파티션 중 1~2개만 접근하면 파티션 히트가 정상 동작하는 것이다.

3) 인덱스 '설계-구현 대조'는 건수 비교로 신속히 수행하라: '데이터베이스 설계서'의 인덱스 수(예: 45개)와 사업자가 제출한 DB 카탈로그의 인덱스 수를 비교하여, 건수가 일치하면 상세 대조를, 불일치하면 누락 인덱스를 즉시 식별할 수 있다.

4) 'AI 학습 배치 쿼리'의 성능을 별도로 확인하라: 온라인 추론 쿼리(1초 이내 SLA)와 배치 학습 쿼리(수분~수십 분 SLA)는 SLA 기준이 다르다. 배치 학습 쿼리가 {성능테스트결과서}에 포함되어 있는지, 해당 쿼리에서도 파티션 히트가 동작하는지를 별도로 확인한다.

5) 'SQL 튜닝 결과서'의 존재 자체를 확인하라: 대용량 AI DB에서 SQL 튜닝은 필수적이며, 'SQL 튜닝 결과서'가 별도 산출물로 작성되어 있는지 확인한다. 미작성 시 성능 최적화 노력의 부재로 판단하여 보완 권고할 수 있다. 튜닝 결과서에는 주요 쿼리의 실행 계획 변경 전후 비교, 응답 시간 개선율이 포함되어야 한다.

4.5. 시스템구조·보안

인프라·모델 보안, 성능, 배포 파이프라인 점검 가이드를 수록한다. (총 74건)

요구정의·분석 (24건)

AG_SD41-1-01 민감정보 유출차단 DLP 연동 요건

AG_SD41-1-02 백업-DR 체계요건

AG_SD41-1-03 모델성능 정확도 목표 요건

AG_SD41-1-04 목표AI시스템구성도완전성점검

AG_SD41-1-05 안전모드 작동 인프라 요건 점검

AG_SD41-1-06 시스템 도입 설치 환경분석 요건점검

AG_SD41-1-07 시스템 용량 산정기준 점검

AG_SD41-1-08 컨테이너 보안 요건점검

AG_SD41-1-09 CI/CD 파이프라인 보안 요건

AG_SD41-1-10 목적기반LegacyAPI설계점검

AG_SD41-1-11 DDoS_API_RateLimit 요건점검

AG_SD41-1-12 SecurityGuideline모델 공격 방어 요건

AG_SD41-1-13 모델저장소 Registry 보안요건

AG_SD41-1-14 비정상 API호출 모델추출 공격요건 점검

AG_SD41-1-15 AI 공급망 보안(AI-BOM) 요건 점검

AG_SD41-1-16 데이터 최적화 방어요건 및 적대적 데이터 최적화 공격방어 요건

AG_SD41-1-17 API 접근통제 인증요건 점검

AG_SD41-1-18 암호화 전송 요건

AG_SD41-1-19 오픈소스 라이브러리 보안 요건

AG_SD41-1-20 망분리 환경 모델 훈련 보안요건 점검

AG_SD41-1-21 AI신뢰성 테스트

AG_SD41-1-22 성능저하 임계값 요건점검

AG_SD41-1-23 성능저하 알림 모니터링 인프라 요건

AG_SD41-1-24 알림 라우팅체계 점검

설계 (19건)

AG_SD42-1-01 개인정보 비식별화 절차 설계

AG_SD42-1-02 성능지표설계

AG_SD42-1-03 성능 임계값 설계점검

AG_SD42-1-04 안전모드 Fail-safe 트리거 설계 점검

AG_SD42-1-05 CI/CD 배포 파이프라인 보안 설계점검

AG_SD42-1-06 모델 공격-방어 설계점검

AG_SD42-1-07 AI서비스 관리자페이지 접근통제 설계점검

AG_SD42-1-08 API 접근 인증-인가 통제설계점검

AG_SD42-1-09 모델저장소 접근통제점검

AG_SD42-1-10 HSM 연동 키관리 아키텍처 설계점검

AG_SD42-1-11 구간암호화 TLS 인증서 설계점검

AG_SD42-1-12 DDoS 공격방어 설계점검

AG_SD42-1-13 보안약점 예방 아키텍처 설계점검

AG_SD42-1-14 웹셸 악성파일 업로드 차단 설계점검

AG_SD42-1-15 클라우드 인프라 보안그룹 설계점검

AG_SD42-1-16 테스트환경 구축 계획

AG_SD42-1-17 자동화 테스트 설계점검

AG_SD42-1-18 SLM 경량 모델 활용 요건

AG_SD42-1-19 모델추론서버 스케일링(Auto-scaling) 점검

구현 (5건)

AG_SD43-1-01 적대적 사례 공격방어 점검

AG_SD43-1-02 모델저장소 비인가 접근차단 점검

AG_SD43-1-03 모델·파일스토리지 암호화 점검

AG_SD43-1-04 시스템 보안약점 예방필터 계층점검

AG_SD43-1-05 파일업로드 악성코드 방지점검

시험 (21건)

AG_SD44-1-01 민감정보 유출 차단점검

AG_SD44-1-02 성능 평가지표 달성여부

AG_SD44-1-03 시스템 목표용량 점검

AG_SD44-1-04 안전모드 이중화 전환점검

AG_SD44-1-05 API 게이트웨이 인증_인가 통제

AG_SD44-1-06 내부 서버간 제로트러스트 인증

AG_SD44-1-07 서비스 관리자페이지 IP 통제점검

AG_SD44-1-08 구간암호화 및 인증서 유효성

AG_SD44-1-09 DDoS_WAF 방어시험 점검

AG_SD44-1-10 OS미들웨어 보안패치_CVE취약점 점검

AG_SD44-1-11 망분리_폐쇄환경 모델 훈련점검

AG_SD44-1-12 클라우드 보안설정_방화벽 점검

AG_SD44-1-13 모델 신뢰성 테스트 점검

AG_SD44-1-14 성능테스트 통과여부

AG_SD44-1-15 APM 성능 저하 탐지 알림발송

AG_SD44-1-16 모니터링 요건 점검

AG_SD44-1-17 모델 백업버전 복구-롤백테스트

AG_SD44-1-18 서버 리소스 쿼터점검

AG_SD44-1-19 운영매뉴얼 목차 설계 점검

AG_SD44-1-20 오픈소스 라이선스 충돌점검

AG_SD44-1-21 위험관리대응매뉴얼 점검

전개 (5건)

AG_SD45-1-01 CI/CD 배포 파이프라인 자동화 점검

AG_SD45-1-02 컨테이너이미지 보안점검

AG_SD45-1-03 모델 드리프트 감지 및 재학습 트리거 기준 요건

AG_SD45-1-04 AI 서비스 로그 관리 및 프롬프트 감사 추적

AG_SD45-1-05 시스템 리소스 가용성 점검

AG_SD41-1-01 | 민감정보 유출차단 DLP 연동 요건

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 개인정보보호

근거 | 「개인정보 보호법」 · NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템의 프롬프트 입력 및 모델 응답 경로를 통한 민감정보·개인정보 유출 위험을 차단하기 위한 DLP(Data Loss Prevention) 연동 요건과 프롬프트 내 개인정보 유출 탐지 인프라 요건이 {보안요구사항}에 구체적이고 상세하게 정의되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 일반적인 네트워크 DLP 적용과 달리 AI 서비스 고유의 데이터 흐름(사용자 프롬프트 → AI 추론 엔진 → 모델 응답)에 맞춰 개인정보 유형별 탐지 규칙, 마스킹 정책, 탐지 로그 관리 방안이 보안요구사항으로 명확히 정의되었는지를 확인한다.

2. 필요성

▶ 생성형 AI 서비스는 사용자가 자연어로 프롬프트를 입력하고 모델이 자연어로 응답하는 구조상, 기존 IT 시스템에서 보기 어려운 민감정보 유출 경로가 존재한다. 사용자가 프롬프트에 주민등록번호, 카드번호, 의료정보 등을 의도 또는 비의도적으로 입력하거나, 모델 응답에 학습 데이터 내 개인정보가 노출될 위험이 있다.

▶ 프롬프트 경로의 신규 유출 위험: 생성형 AI 시스템은 사용자 입력이 외부 API(클라우드 LLM 등)로 전송되는 경우가 많아, 프롬프트에 포함된 개인정보가 외부로 유출될 수 있다. 이에 프롬프트 전송 전 실시간 탐지·차단·마스킹 인프라가 정의되어야 한다.

▶ 응답 경로의 학습 데이터 노출 위험: 모델이 학습 과정에서 기억한 개인정보(Memorization)가 응답에 포함되어 사용자에게 노출될 수 있다. 응답 경로에도 민감정보 필터링 요건이 반영되어야 한다.

▶ 개인정보 보호법 준수 의무: 공공 AI 사업은 개인정보 보호법상 안전성 확보 조치 의무를 준수해야 하며, AI 서비스 특유의 데이터 흐름에 맞춘 기술적 보호 조치가 요구사항에 선제적으로 반영되어야 한다.

▶ 규제 기준의 AI 프라이버시 요건: NIST AI RMF의 Privacy-Enhanced 원칙 및 TTA 생성AI 개발안내서는 AI 시스템 설계 초기부터 프라이버시 보호 기술(PETs) 적용을 권고하고 있으며, 요구정의 단계에서 이를 반영해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	프롬프트 입력 경로(사용자→AI 추론 엔진)의 DLP 연동 지점	{보안요구사항}에서 사용자 프롬프트가 추론 엔진에 전달되기 전 민감정보를 탐지·차단·마스킹하는 DLP 연동 지점이 명시되었는지 확인
2	AI 응답 경로(추론 엔진→사용자)의 DLP 연동 지점	{보안요구사항}에서 모델 응답이 사용자에게 전달되기 전 민감정보를 탐지·마스킹하는 DLP 연동 지점이 명시되었는지 확인하고, 학습 데이터 기억(Memorization)에 의한 개인정보 노출 차단 요건이 포함되었는지 함께 확인
3	외부 LLM API 호출 전 DLP 검사 요건	외부 LLM API로 프롬프트 데이터를 전송하기 전에 DLP 검사를 수행하도록 요건이 반영되었는지 {보안요구사항} 및 'AI 보안 아키텍처(안)'와 대조
4	DLP 차단 정책 유형 정의	탐지 시 차단·경고·마스킹 등 정책 유형과 유형별 적용 조건이 {보안요구사항}에 정의되었는지 확인
5	탐지 대상 민감정보 유형 열거	주민등록번호·카드번호·계좌번호·의료정보 등 탐지 대상 민감정보 유형이 {보안요구사항}에 열거되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	프롬프트 입력·응답 경로의 DLP 연동 지점, 외부 LLM API 호출 전 검사, 차단 정책 유형, 탐지 대상 민감정보 유형이 {보안요구사항}에 구체적으로 정의된 경우	현행 유지
보완권고 (△)	DLP 연동 요건이 일부 정의되었으나, 외부 API 호출 전 검사 요건 누락, 차단 정책 유형 미정의, 민감정보 유형 미열거 등 일부 항목이 미비하거나 선언적 수준에 그친 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 서비스 고유의 프롬프트·응답 경로에 대한 DLP 연동 및 개인정보 유출 탐지 요건이 {보안요구사항}에 전혀 정의되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 보안 요건의 수치화: '민감정보 유출을 방지한다', 'DLP를 적용한다'와 같은 모호한 문구는 분쟁의 원인이 된다. '프롬프트 내 개인정보 탐지 패턴 30건 이상 적용', 'DLP 탐지율 95% 이상', '탐지 로그 보관 기간 1년 이상' 등 측정 가능한 기준으로 구체화하도록 유도해야 한다.

2) AI 고유 데이터 흐름과 기존 DLP 경로 구분: 기존 DLP는 이메일·USB·웹 업로드 경로를 감시하지만, AI 서비스에서는 프롬프트 입력 → 전처리 → 외부 LLM API 호출 → 응답 후처리의 흐름이

추가된다. 감리원은 기존 DLP 경로가 아닌 AI 고유 데이터 경로에 대한 DLP 연동 요건이 별도로 정의되었는지에 중점을 두어야 한다.

3) 마스킹과 모델 품질의 트레이드오프: 프롬프트 내 개인정보를 과도하게 마스킹하면 모델 추론 품질이 저하될 수 있다. 요구사항에 마스킹 수준과 모델 품질 간 균형 분석 활동이 포함되어 있는지 확인하고, 이를 테스트할 수 있도록 권고하는 것이 바람직하다.

4) 외부 LLM API 사용 시 추가 위험 확인: 클라우드 기반 LLM API를 사용하는 경우, 프롬프트 데이터가 외부 서버로 전송되므로 DLP 검사가 API 호출 전에 반드시 수행되어야 한다. 이 경우 API 제공자의 데이터 처리·보관 정책도 요구사항에 포함되어야 하며, 감리원은 이 부분을 면담 시 확인해야 한다.

5) 개인정보 보호법과의 정합성 확인: AI 서비스의 프롬프트 처리 과정에서 개인정보가 수집·이용·제공되는 법적 근거가 존재하는지, 개인정보 영향평가 대상에 해당하는지를 확인하고, {보안요구사항}에 법적 근거가 기술되어 있는지 점검한다.

AG_SD41-1-02 | 백업-DR 체계요건

시스템구조·보안	요구정의·분석	시스템 구현	데이터 백업/복구

근거 | 「전자정부법」 · ISO/IEC 27001 · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템의 백업 및 재해복구(DR) 체계 요건이 {보안요구사항}에 구체적으로 명시되었는지, 특히 AI 모델 가중치(Weight)·하이퍼파라미터·설정 파일 등 AI 고유 자산에 대한 백업 요건이 포함되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템은 일반 정보시스템과 달리 모델 가중치 파일이 수십~수백 GB에 달하고, 학습 데이터·파인튜닝 이력·프롬프트 설정 등이 시스템 복원의 핵심 요소이다.

이러한 AI 고유 자산이 백업·복구 대상에서 누락될 경우 장애 발생 시 서비스 복원이 불가능하거나 모델 성능이 비가역적으로 저하될 위험이 있으므로, 요구정의 단계에서 이를 보안요구사항으로 사전 정의해야 한다.

2. 필요성

▶ 생성형 AI 시스템은 일반 정보시스템의 백업·복구 체계만으로는 서비스 연속성을 보장할 수 없는 구조적 특성을 갖고 있다. 기존 시스템은 애플리케이션 코드, 데이터베이스, 설정 파일 등을 백업하면 복원이 가능하지만, AI 시스템은 이에 더하여 모델 가중치·학습 이력·벡터 데이터베이스·프롬프트 설정 등 AI 고유 자산이 복원의 핵심 요소가 된다.

▶ AI 자산의 비가역적 손실 방지: 모델 가중치는 장기간의 학습·파인튜닝 결과물로서, 손실 시 원복에 수주~수개월의 재학습이 필요하며 동일한 결과를 보장할 수 없다. 요구정의 단계에서 이를 백업 대상으로 명시하지 않으면 개발 과정에서 누락될 가능성이 높다.

▶ GPU/NPU 인프라 장애 대응: AI 시스템 운영에 사용되는 GPU·NPU 등 가속기 장비는 고온·고전력 환경에서 운영되어 일반 서버 대비 하드웨어 장애 발생률이 높으므로, DR 체계 요건을 사전에 정의해야 한다.

▶ AI 서비스의 업무 연속성 확보: 생성형 AI 기반 서비스가 핵심 업무에 활용되는 경우, RTO(복구 목표 시간)·RPO(복구 목표 시점)를 AI 특성에 맞게 별도 정의해야 하며, 이를 요구사항에 명시하지 않으면 운영 단계에서 SLA 미달 사태가 발생할 수 있다.

▶ 법적·제도적 준수: 공공기관의 정보시스템은 전자정부법, 정보보호관리체계(ISMS) 등에 따라 백업·복구 계획 수립이 의무화되어 있으며, AI 시스템도 이 범위에 포함된다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 가중치(Weight) 파일의 백업 대상 명시	{보안요구사항}에서 모델 가중치 파일이 백업 대상 자산으로 명시되었는지 확인
2	파인튜닝 이력·학습 데이터셋·하이퍼파라미터의 백업 대상 포함	{보안요구사항} 및 {요구사항정의서}에서 파인튜닝 이력, 학습 데이터셋, 하이퍼파라미터 설정이 백업 대상에 포함되었는지 확인
3	벡터 DB 및 임베딩 인덱스의 백업 대상 포함	RAG 구성 시 벡터 데이터베이스와 임베딩 인덱스가 백업 대상에 포함되었는지 {보안요구사항}에서 확인
4	시스템 프롬프트·가드레일 등 AI 설정 파일의 백업 대상 포함	시스템 프롬프트, 가드레일 규칙, 금칙어 목록 등 AI 서비스 설정 파일이 백업 대상에 포함되었는지 {보안요구사항}에서 확인
5	AI 특성을 반영한 RTO/RPO의 정량 정의	{보안요구사항} 또는 {시스템요구사항}에서 RTO(복구 목표 시간)와 RPO(복구 목표 시점)가 수치로 정의되고, 모델 재로딩·GPU 자원 재확보 등 AI 특성에 따른 소요 시간이 반영되었는지 확인. '신속 복구' 등 정성 표현만 있으면 미흡으로 판정

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	모델 가중치, 파인튜닝 이력·학습 데이터셋·하이퍼파라미터, 벡터 DB·임베딩 인덱스, 시스템 프롬프트·가드레일 등 AI 고유 자산이 백업 대상으로 명시적으로 구분·열거되고, AI 특성을 반영한 RTO/RPO가 정량적으로 정의된 경우	현행 유지
보완권고 (△)	백업·DR 요건은 정의되어 있으나 AI 고유 자산 중 일부(벡터 DB, AI 설정 파일 등)가 백업 대상에서 누락되었거나, RTO/RPO가 '신속 복구' 등 모호한 표현에 그친 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	일반 시스템 백업 항목(OS, DB, 소스코드)만 정의되고 AI 고유 자산에 대한 백업·DR 요건이 전혀 정의되지 않아 장애 시 AI 서비스 복원이 불가능한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (RTO/RPO 포함 복구 계획 수립 요구)

5. 감리 시 유의사항 및 실무 Tip

1) AI 백업 대상 자산과 일반 시스템 백업 대상의 혼동 방지: 사업자가 일반 시스템 백업 항목(OS, DB, 소스코드)만 나열하고 AI 고유 자산(모델 가중치, 벡터 DB 등)을 누락하는 사례가 빈번하

다. {보안요구사항}의 백업 대상 자산 목록에서 AI 고유 자산이 명시적으로 구분·열거되어 있는지를 반드시 확인해야 한다.

2) 모호한 RTO/RPO 표현의 수치화 유도: "장애 시 신속하게 복구한다"와 같은 모호한 문구는 감사 단계에서 분쟁의 원인이 된다. "AI 추론 서비스 RTO: 모델 로딩 시간 포함 60분 이내", "모델 가중치 RPO: 최종 파인튜닝 완료 시점 기준 24시간 이내" 등 측정 가능한 수치로 구체화하도록 유도해야 한다.

3) AI 특화 데이터베이스(벡터 DB 등)의 RPO/RTO 기준 차등화 요건 확인: 수천만 건의 임베딩 데이터가 저장된 벡터 DB는 장애 시 재생성(Re-embedding)에 막대한 시간과 비용(API 호출 비용 등)이 소요된다. 단순 스냅샷 백업을 넘어, 벡터 DB와 모델 파일의 특성을 고려한 구체적인 목표복구시간(RTO)과 목표복구시점(RPO)이 요구사항에 정량적으로 명시되어 있는지 점검해야 한다.

4) 모델 가중치의 지적 재산권 관점 점검: 파인튜닝된 모델 가중치는 사업의 핵심 지적 자산이므로, 백업·보관 시 암호화 및 접근통제가 필수적이다. 특히 클라우드 기반 AI 서비스의 경우, 모델 가중치가 클라우드 사업자의 스토리지에 저장되므로 데이터 주권(Data Sovereignty) 및 반출 가능 여부를 사전에 확인해야 한다.

5) DR 전환 시 모델 일관성(Consistency) 확인: DR 사이트로 전환 시 백업된 모델 가중치와 벡터 DB의 시점이 일치하는지(예: 동일 시점의 스냅샷인지) 확인하는 절차가 필요하다. 시점 불일치 시 AI 서비스의 응답 품질이 저하될 수 있으므로, 이를 DR 전환 검증 기준에 포함하도록 유도해야 한다.

AG_SD41-1-03 | 모델성능 정확도 목표 요건

시스템구조·보안 | 요구정의·분석 | 인공지능 모델 개발 | 모델평가

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 {요구사항정의서}에 AI 모델의 성능 및 정확도 목표를 측정 가능한 수치 기준으로 구체적이고 명확하게 정의하였는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 '시스템은 정확한 답변을 제공한다'와 같은 선언적·추상적 요건이 아니라, AI 모델 유형(분류, 생성, 추천 등)에 적합한 정량적 성능 지표(F1-score, Precision, Recall, BLEU/ROUGE 등)와 합격 임계값이 요구사항에 명시되어 있는지, 그리고 해당 성능을 검증하기 위한 평가 데이터셋 및 검사 기준이 수립되었는지를 확인한다.

모호한 성능 요건은 검사 단계에서 발주기관과 사업자 간 분쟁의 핵심 원인이 되며, 이를 사전에 구체화하지 못하면 준공 검사 불통과·납기 지연 등 심각한 사업 위험을 초래할 수 있다.

2. 필요성

▶ AI 시스템은 확률적 출력을 생성하는 특성을 가지므로, 일반 소프트웨어와 달리 '정상 동작' 여부를 단순 통과/실패(Pass/Fail) 기준으로 판정하기 어렵다. 성능·정확도 목표 요건은 다음과 같은 이유로 반드시 점검되어야 한다.

▶ 분쟁 예방: 성능 목표가 수치로 명시되지 않으면 검사 단계에서 '적합'·'부적합' 판정 근거가 없어 발주기관과 사업자 간 극심한 분쟁이 발생한다. 예를 들어 '챗봇 응답 정확도 85% 이상' 등 검증 가능한 기준이 없으면 검수 완료가 불가능하다.

▶ 품질 확보: TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI)는 성능 목표를 도메인 맥락에 맞게 구체적으로 정의하고, 이를 테스트·검증 절차와 연계할 것을 요구한다. 기획 단계에서 성능 목표를 수립하지 않으면 개발 완료 후에도 품질 기준 없이 시스템이 배포되는 위험이 발생한다.

▶ 신뢰성 근거 확보: NIST AI RMF MEASURE 2.3은 'AI 시스템의 성능 또는 보증 기준이 배포 환경과 유사한 조건에서 정량적·정성적으로 측정되고 문서화되어야 한다'고 규정하고 있다. 요구사항에 측정 기준이 부재하면 이후 단계에서 이를 소급 적용하는 것은 사실상 불가능하다.

▶ 모델 유형별 적합 지표 선택: 분류 모델에는 F1-score, 생성형 AI에는 BLEU·ROUGE·응답 충실도(Faithfulness)와 같이 모델 유형에 따라 적합한 지표가 다르다. 요구사항에 모델 유형이 명시되지 않으면 개발자가 임의의 지표를 사용하여 성능을 과장할 위험이 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 유형의 요구사항정의서 명시 여부	{요구사항정의서}에서 분류·텍스트 생성·추천·객체 탐지 등 AI 모델 유형이 명시되었는지 직접 검토
2	모델 유형에 적합한 정량적 성능 지표 명시 여부	{요구사항정의서}에서 모델 유형별 성능 지표(분류→Precision/Recall/F1, 생성→BLEU/ROUGE/Faithfulness, 추천→nDCG 등)가 명시되었는지 확인
3	성능 지표별 합격 임계값(수치 기준) 명시 여부	{요구사항정의서}·'AI 모델 성능 목표 명세서'에서 각 지표의 합격 임계값(예: F1-score 0.85 이상)이 수치로 명시되었는지 확인
4	평가 데이터셋의 구성·규모·대표성 요건 포함 여부	{요구사항정의서} 검토와 담당자 면담으로 평가 데이터셋의 구성·규모·대표성 요건(예: 검증 1,000건 이상, 실서비스 분포 반영)이 포함되었는지 확인
5	기준 모델(Baseline) 또는 비교 벤치마크 명시 여부	{요구사항정의서}·{과업대비표}에서 기준 모델 또는 비교 벤치마크(예: 기존 대비 10%p 향상, 공개 벤치마크 상위 25% 이내)가 명시되었는지 검토
6	생성형 AI 환각 허용 기준 및 탐지·모니터링 요건 포함 여부	{요구사항정의서} 검토와 담당자 면담으로 환각 허용 기준(예: 사실 오류율 5% 미만)이나 탐지·모니터링 요건이 포함되었는지 확인
7	운영 환경 반영 성능 목표 설정 여부	{요구사항정의서}·'AI 서비스 품질 기준서'에서 성능 목표가 학습 데이터가 아닌 실서비스 데이터·부하 조건을 반영해 설정되었는지 확인
8	성능 목표와 RFP 요구사항 간 추적성 확보 여부	{요구사항추적표}·{과업대비표}에서 성능 목표가 RFP의 AI 요구사항과 일치하고 연계 관계가 추적되는지 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 모델 유형과 유형별 정량 성능 지표, 지표별 합격 임계값, 평가 데이터셋 요건과 기준 모델(Baseline)·비교 벤치마크가 {요구사항정의서}에 수치로 명시되고, 성능 목표가 실서비스 운영 환경 조건을 반영하며, 생성형 AI의 경우 환각 허용 기준까지 포함되어 RFP와 추적 가능한 경우	현행 유지
보완권고 (△)	성능 지표는 명시되었으나 합격 임계값·평가 데이터셋 구성 기준·운영 환경 반영 조건 등 일부 요건이 누락되었거나, RFP 목표치가 요구사항정의서에서 완화·생략된 경우	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	성능·정확도 목표가 'AI는 정확한 답변을 제공한다' 수준의 선언적 문구에 그치고, 측정 가능한 지표와 임계값이 전혀 정의되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요구사항정의서 구체성 기준 적용: 성능 요건은 정량적 허용 기준(예: F1 ≥ 0.85, 응답시간 ≤ 2초)으로 명시되어야 측정 가능한 점검이 가능하다.

2) 제안요청서에 제시된 AI 모델 관련 요건이 단순 선언적 기술로만 작성되어 있는지, 아니면 측정 가능한 기준(수치·범위·조건)으로 구체화되어 있는지를 구분하여 판단한다. 단순 "구현 예정" 표현은 적합 근거가 되지 못한다.

3) RFP와 요구사항정의서 교차 검토: RFP 성능 목표치가 요구사항정의서에서 완화·생략되는 경우를 집중 점검한다. 이는 가장 흔한 발주기관-사업자 분쟁 유형이다.

4) RFP(제안요청서)에 AI 모델 관련 요구사항이 명시된 경우, {과업대비표}를 통해 {요구사항정의서}로의 1:1 매핑 여부를 확인한다. 누락된 요구사항은 시정요구 대상으로 기록한다.

5) 성능 담당자 면담 시 실제 부하 시나리오와 병목 지점을 확인하고, 요건에 반영된 수치의 근거를 묻는다.

AG_SD41-1-04 | 목표AI시스템구성도완전성점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 아키텍처

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 작성한 시스템 아키텍처 상위 개념도({시스템구성도})와 목표 AI 시스템 구성도가 제안요청서(RFP)에 명시된 요구사항을 빠짐없이 반영하고, AI 서비스 특유의 구성요소(모델 서빙 계층, 데이터 파이프라인, GPU/추론 인프라, 외부 API 연계 등)를 식별 가능한 수준으로 구체적으로 표현하고 있는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 일반 정보시스템과 달리 AI 서비스는 학습 환경과 추론(서빙) 환경이 분리되고, 모델 버전 관리·재학습 파이프라인·벡터 데이터베이스 등 고유한 구성요소가 존재하므로, 이러한 AI 특화 계층이 상위 개념도에 누락 없이 도식화되었는지를 확인하는 것이 핵심이다.

2. 필요성

▶ AI 서비스 구축 사업에서 시스템 아키텍처 상위 개념도는 전체 시스템의 논리적·물리적 구조를 한눈에 파악할 수 있는 핵심 산출물이다. 요구정의 단계에서 이 구성도의 완전성이 확보되지 않으면, 이후 설계·개발 단계에서 구성요소 누락, 인터페이스 불일치, 인프라 과소 산정 등의 문제가 연쇄적으로 발생할 수 있다.

▶ AI 특화 구성요소 누락 방지: 일반 시스템 구성도에서는 표현되지 않는 모델 학습 환경, 추론 API Gateway, 프롬프트 관리 계층, RAG(검색증강생성) 파이프라인, 벡터 DB 등의 AI 고유 요소가 초기 구성도에서 누락되면, 설계 단계에서 아키텍처 전면 재수립이 필요해질 수 있다.

▶ 학습-추론 환경 분리 확인: 생성형 AI 시스템은 학습(Training)과 추론(Inference) 환경이 물리적·논리적으로 분리 운영되는 것이 일반적이며, 이 분리 구조가 구성도에 명확히 표현되어야 자원 산정, 보안 설계, 네트워크 구성의 정확성을 확보할 수 있다.

▶ 외부 AI 서비스 연계 구조 명시: 클라우드 기반 LLM API(예: GPT, Claude 등)를 활용하는 경우, 내부 시스템과의 연계 구간·데이터 흐름·보안 경계가 구성도에 가시적으로 표현되어야 보안 리스크 식별이 가능하다.

▶ 이해관계자 간 공통 이해 기반 확보: 상위 개념도는 발주기관, 사업자, 감리원 등 모든 이해관계자가 시스템 전체 구조를 동일하게 이해하기 위한 커뮤니케이션 기반이므로, 초기 단계의 완전성이 사업 전체의 품질을 좌우한다.

3. 검토방법

No	점검 항목	확인 방법
1	학습·추론 환경 구분 표현 여부	AI 모델 학습 환경(Training)과 추론 환경(Inference/Serving)이 구분되어 {시스템구성도}에 표현되었는지 확인
2	모델 서빙 계층 독립 표현 여부	API Gateway·추론 서버·로드밸런서 등 모델 서빙 계층(Model Serving Layer)이 독립된 구성요소로 표현되었는지 {시스템구성도}·'목표 AI 시스템 구성도'에서 대조
3	데이터 파이프라인 흐름 표현 여부	수집→전처리→학습데이터 저장→피드백 데이터 파이프라인 흐름이 {시스템구성도}에 표현되었는지 확인
4	생성형 AI 구성요소 포함 여부	생성형 AI 활용 시 RAG 파이프라인·벡터 DB·프롬프트 관리 계층이 구성요소로 포함되었는지 {시스템구성도}·'목표 AI 시스템 구성도'에서 확인
5	외부 AI 서비스 연계 구간·보안 경계 표현 여부	클라우드 LLM API 등 외부 AI 서비스 연계 시 연계 구간·프로토콜·보안 경계가 {시스템구성도}에 표현되었는지 {요구사항정의서}와 대조
6	모델 레지스트리·버전 관리 체계 식별 여부	모델 레지스트리(Model Registry) 또는 모델 버전 관리 체계가 구성요소로 식별되었는지 {시스템구성도} 검토·면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	학습·추론 환경 분리, 모델 서빙 계층, 데이터 파이프라인 흐름, RAG·벡터 DB 등 생성형 AI 구성요소, 외부 연계 구간의 보안 경계, 모델 버전 관리 체계가 구성도에 식별 가능한 수준으로 모두 표현된 경우	현행 유지
보완권고 (△)	주요 AI 계층은 표현되었으나 일부 구성요소(모델 레지스트리, 프롬프트 관리 계층, 외부 연계 구간의 보안 경계 표기 등)가 누락되었거나, 'AI 엔진' 등 모호한 단일 박스로 표현되어 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인 (생성형 AI 특화 요건 누락 시 추가 명시 요구)
부적합 (X)	구성도가 일반 시스템 수준(WEB-WAS-DB 3계층)에 그쳐 AI 특화 구성요소가 전혀 표현되지 않았거나, RFP 요구사항을 반영한 목표 AI 시스템 구성도가 제시되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 구성도의 '추상화 수준' 적정성 판단: 상위 개념도는 지나치게 상세하면 가독성이 떨어지고, 지나치게 추상적이면 구성요소 누락 식별이 어렵다. AI 서비스의 경우, 최소한 '학습 환경–추론 환

경–데이터 계층–외부 연계 구간'이 식별 가능한 수준의 추상화가 적정하다. 일반 시스템 구성도 수준(WEB-WAS-DB 3계층)만으로는 AI 아키텍처 점검이 불가능하므로, AI 특화 계층의 표현을 반드시 요구해야 한다.

2) 보안 아키텍처 관점의 플레인 분리 확인: 제로트러스트(Zero Trust) 원칙을 반영하여, 데이터가 처리되는 '데이터/추론 플레인(Data Plane)'과 시스템을 모니터링·관리하는 '컨트롤 플레인(Control Plane)' 영역이 논리적·물리적으로 명확히 분리되어 표현되었는지 점검한다. RAG 기반일 경우 벡터 DB가 아키텍처 상에 완전하게 식별되어야 한다.

3) '목표 AI 시스템 구성도' 미제출 시 대응: 사업자가 별도의 AI 시스템 구성도를 제출하지 않은 경우, {시스템구성도} 내에 AI 특화 구성요소가 충분히 표현되어 있는지를 기준으로 점검한다. 다만, 생성형 AI 서비스와 같이 복잡한 파이프라인이 필요한 경우에는 별도 산출물 작성을 권고하는 것이 바람직하다.

4) 외부 API 연계 구간의 '보안 경계' 필수 확인: 외부 클라우드 LLM API를 활용하는 경우, 내부 데이터가 외부로 전송되는 구간이 구성도에 명확히 표시되어야 한다. '단순 화살표'만으로는 부족하며, 암호화 프로토콜(TLS 등), 데이터 마스킹 적용 구간 등이 범례 또는 주석으로 부기되었는지 확인한다.

5) GPU/TPU 인프라 반영과 모호한 명칭 구체화: AI 추론 서버의 GPU 계층 존재와 예상 규모가 구성도 또는 '인프라 구성 계획서'에 표현되었는지 확인하고, 'AI 엔진', 'AI 플랫폼'과 같이 모호한 단일 박스로 표현된 경우 세부 구성요소(모델 서빙, 학습 파이프라인, 벡터 DB 등)를 면담으로 확인하여 구체화를 권고한다. 모호한 표현은 향후 설계·검사 단계에서 범위 분쟁의 원인이 된다.

AG_SD41-1-05 | 안전모드 작동 인프라 요건 점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 가용성/이중화

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 제안요청서(RFP)상의 요건이 {시스템요구사항}에 구체적이고 정량적으로 반영되었는지를 감리원이 점검하기 위한 것이다.

특히 AI 서비스 장애·오류 발생 시 서비스를 안전하게 유지하거나 복구하는 데 필요한 두 가지 핵심 요건, 즉 ① 안전모드 작동을 위한 인프라 구성 요건과 ② 폴백 서버/모델 이중화 구성 요건이 요구사항 문서에 누락·모호하게 기술되어 있을 경우, 이후 설계·구축 단계에서 해당 기능이 구현되지 않거나 검사 기준이 불분명해지는 위험이 직접적으로 발생한다.

이러한 요건 미확정 리스크를 사전에 식별·조치하는 것이 본 가이드라인의 목적이다.

2. 필요성

▶ 생성형 AI 서비스는 외부 API 연계, LLM 추론 서버, 거대한 모델 파라미터 등의 구조적 특성으로 인해 일반 정보시스템에 비해 장애 유형이 복잡하고 예측이 어렵다. 이에 따라 다음과 같은 이유로 안전모드 및 폴백 이중화 요건을 반드시 확인해야 한다.

▶ 안전모드 요건 미확정 리스크: AI 모델이 비정상 응답(환각, 유해 콘텐츠 등)을 생성하거나 추론 서버 과부하가 발생할 경우, 서비스를 자동으로 안전 상태로 전환하는 인프라 요건이 요구사항에 명시되지 않으면 구현 단계에서 해당 기능이 누락되는 구조적 원인이 된다.

▶ 폴백 이중화 요건 미확정 리스크: 주 LLM API 또는 추론 서버 장애 시 대체 경로(폴백)로 자동 전환하는 구성이 요구사항에 정량적으로 규정되지 않으면, 사업자는 이를 선택 사항으로 처리하여 구축 범위에서 제외할 가능성이 높다.

▶ 요구사항의 모호성이 사업 후반 분쟁의 원인: "안정적으로 운영해야 한다"와 같은 정성적 표현은 요건 충족 여부에 대한 분쟁을 야기하므로, 감리원은 정량화된 요건 명시 여부를 점검해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	안전모드 트리거 조건 정의 여부	안전모드가 작동해야 하는 조건(예: 유해 콘텐츠 임계값, 연속 오류 횟수, 응답 지연 기준 등)이 {시스템요구사항}에 수치·임계값으로 명시되어 있는지 확인

No	점검 항목	확인 방법
2	안전모드 전환 후 서비스 처리 방식 요건 명시 여부	안전모드 활성화 시 사용자에게 고정 메시지 제공, 기능 제한, 서비스 일시 중단 등 구체적인 처리 방식이 정의되어 있는지 확인
3	안전모드 해제 및 복귀 기준 정의 여부	안전모드 해제를 위한 조건(관리자 수동 해제, 자동 해제 기준 등)과 정상 서비스로의 복귀 절차가 요구사항에 포함되어 있는지 확인
4	폴백 대상 구성요소 및 우선순위 정의 여부	폴백이 적용될 대상(주 LLM API, 추론 서버 등)과 대체 경로의 우선순위(1차 폴백 → 2차 폴백 등)가 {시스템요구사항}에 명시되어 있는지 확인
5	폴백 전환 기준(임계값) 정량적 정의 여부	폴백 전환이 자동으로 발생하는 조건, 예: 응답 지연 N초 초과, 연속 오류 N회 이상, 가용성 N% 미만 등이 수치로 정의되어 있는지 확인
6	폴백 서버/모델의 성능·품질 요건 명시 여부	대체 서버·모델이 갖추어야 할 최소 성능 수준(응답 정확도, TPS, 레이턴시 등)이 정의되어 있어 "최소 기능 수준 유지"가 가능한지 확인
7	폴백 동작 여부 자동 확인(Health Check) 요건 정의	폴백 서버·모델이 상시 가동 상태인지 주기적으로 자동 점검(Health Check)하는 요건이 {시스템요구사항}에 정의되어 있는지 확인
8	요건의 정량화·측정 가능성 적정성	요구사항 전반에 걸쳐 "적절히", "충분히", "안정적으로" 등 비정량적 표현이 사용되지 않고, 감사·검사 시 측정 가능한 수치 기준으로 작성되었는지 확인
9	AI 특화 장애 유형 식별 및 시나리오 정의 여부	일반 IT 장애(서버 다운 등)뿐만 아니라 AI 특화 장애(LLM API 타임아웃, 토큰 한도 초과, 환각 비율 급증 등)가 'AI 위험·장애 시나리오 목록'에 포함되어 있는지 확인
10	이중화 아키텍처와 요구사항 정합성 확인	'AI 서비스 아키텍처 설계(안)'의 이중화 구성이 {시스템요구사항}의 폴백·안전모드 요건과 일치하는지 두 산출물을 대조 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	안전모드 트리거·처리·해제 기준과 폴백 대상·전환 임계값·성능 요건 및 폴백 서버·모델의 상시 가동 여부를 확인하는 Health Check 요건이 {시스템요구사항}에 구체적·정량적으로 명시되고, LLM API 타임아웃·토큰 한도 초과·환각 비율 급증 등 AI 특화 장애 유형이 AI 위험·장애 시나리오 목록에 포함되며, 아키텍처 설계의 이중화 구성과 일치하는 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	안전모드·폴백 요건이 일부만 명시되었거나 "충분히 구성한다" 등 비정량적 표현으로 기술되어 측정 가능성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	안전모드·폴백 요건 자체가 누락되었거나, 설계 내용이 요구사항과 상이한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요건의 수치화 유도가 감리원의 핵심 역할이다. "AI는 장애 시 적절한 방식으로 대응해야 한다"와 같은 모호한 표현은 향후 분쟁의 원인이 되므로, 안전모드 트리거 조건은 "응답 오류 연속 5회 이상", 폴백 전환은 "API 응답 지연 3초 초과" 등 측정 가능한 기준으로 구체화하도록 권고한다.

2) 안전모드와 폴백은 별개 요건으로 구분하여 점검한다. 안전모드는 서비스 자체를 제한·축소하는 동작이고 폴백은 다른 시스템으로 전환하는 동작으로, 두 요건은 서로 보완 관계이나 혼동되는 경우가 많다.

3) AI 특화 장애 유형(LLM API 타임아웃, 토큰 한도 초과, 환각 비율 급증 등)이 일반 IT 장애와 별도로 'AI 위험·장애 시나리오 목록'에 식별되어 있는지 확인하고, 누락 시 시나리오 보완을 권고한다.

AG_SD41-1-06 | 시스템 도입 설치 환경분석 요건점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 서버스토리지/네트워크

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 제안요청서에서 요구한 AI 시스템의 도입·설치 환경을 구체적으로 분석하였는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

생성형 AI 시스템은 대규모 GPU 서버, 고속 내부 네트워크, 대용량 스토리지, 특화된 AI 프레임워크 등 일반 정보시스템과 상이한 인프라 요구사항을 갖는다.

이러한 요구사항이 기획 단계에서 충분히 분석되지 않으면, 설계·구축 단계에서 환경 불일치로 인한 재설계 및 일정 지연이 발생할 위험이 높으며, 보안 취약점이 구조적으로 내재될 수 있다.

2. 필요성

▶ AI 시스템은 기존 정보시스템 대비 인프라 요구사항이 근본적으로 다르며, 이를 기획 단계에서 식별하지 못할 경우 사업 전반에 리스크가 전파된다.

▶ AI 전용 인프라 요구사항 식별: GPU/NPU 서버, 고속 내부 네트워크(NVLink, InfiniBand 등), 대용량 고속 스토리지 등 AI 특화 인프라 요건을 초기에 정의하지 않으면, 설계 단계에서 재설계 및 추가 예산이 발생한다.

▶ H/W·OS·프레임워크 호환성 분석: AI 프레임워크(PyTorch, TensorFlow 등)와 GPU 드라이버(CUDA 버전), OS 버전 간 호환성이 확보되지 않으면 개발·학습·추론 전 과정에서 장애가 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	GPU/NPU 서버 사양 분석 여부	모델·VRAM·수량 등 GPU/NPU 서버 사양이 AI 모델 요구치에 부합하도록 {시스템요구사항}·'환경 분석 보고서'에서 분석되었는지 확인
2	고속 내부 네트워크 요구사항 분석 여부	분산 학습·추론용 고속 내부 네트워크 요구사항이 {시스템요구사항}·'아키텍처 설계서'에서 분석되었는지 확인
3	대용량 스토리지 용량·성능 분석 여부	학습데이터·모델 저장용 대용량 스토리지의 용량·성능(IOPS·처리량)이 AI 모델 요구치에 부합하도록 {시스템요구사항}·'환경 분석 보고서'에서 정량 분석되었는지 확인

No	점검 항목	확인 방법
4	전력·냉각 설비 요구사항 분석 여부	전력 용량 및 냉각 설비 요구사항이 '환경 분석 보고서'에서 분석되었는지 확인하고 발주기관 면담으로 보완

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	GPU/NPU 서버 사양(모델, VRAM, 수량), 고속 내부 네트워크(NVLink, InfiniBand 등), 스토리지 용량·성능(IOPS), 전력·냉각 요구치가 AI 모델 요구치에 부합하도록 정량적으로 분석·문서화된 경우	현행 유지
보완권고 (△)	환경 분석은 수행되었으나 일부 항목(네트워크, 전력·냉각 등)의 분석이 미흡하거나, '적절한 GPU 서버 도입' 등 수치화되지 않은 모호한 수준으로 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 시스템 도입·설치 환경에 대한 분석이 수행되지 않았거나, 관련 산출물({시스템요구사항}, '환경 분석 보고서')이 부재하여 인프라 적정성 판단이 불가능한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) GPU 사양의 수치화 확인: '적절한 GPU 서버를 도입한다'와 같은 모호한 문구는 분쟁의 씨앗이 된다. 예: 'NVIDIA A100 80GB x 4기, 총 VRAM 320GB 이상'과 같이 측정 가능한 수치로 구체화하도록 유도해야 한다.

2) 호환성 매트릭스 요청: AI 프레임워크 버전 ↔ CUDA 버전 ↔ GPU 모델 ↔ OS 커널 버전 간 호환성을 한눈에 확인할 수 있는 매트릭스를 요청하면 점검 효율이 높아진다. 사업자가 제시하지 않은 경우 보완권고로 반영한다.

3) AI 인프라와 일반 인프라 점검의 구분: 본 가이드라인은 AI 워크로드 특화 환경 요소(GPU·가속기 사양, 전력·냉각, 대용량 스토리지, 학습·추론 네트워크 대역폭)만을 점검 대상으로 한다. 일반적인 서버·네트워크 도입 환경 점검 항목은 기존 감리 점검항목을 적용하며, 중복 점검을 피해야 한다.

4) 외부 AI 서비스 연계 시 보안 검토 강화: 외부 LLM API 또는 클라우드 AI 서비스를 연계하는 경우, 내부 데이터가 외부로 전송되는 경로가 존재하므로 데이터 유출 방지 방안(네트워크 분리, 암호화, 마스킹 등)이 환경 분석에 포함되었는지 특별히 주의해야 한다.

5) 전력·냉각 요구치 확인: GPU 서버는 일반 서버 대비 2~3배 이상의 전력과 냉각을 요구한다. 기존 전산센터의 전력 여유 용량, UPS 용량, 항온항습기 능력이 신규 GPU 서버를 수용할 수 있는지를 환경 분석에서 확인해야 한다. 필요 시 전산실 방문을 통해 실물 확인한다.

AG_SD41-1-07 | 시스템 용량 산정기준 점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 성능

근거 | 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의(기획/분석) 단계에서 도출된 성능 요구사항—목표 동시 접속자 수, 응답시간 기준, 처리량(TPS)—이 {시스템아키텍처설계서} 내 인프라 사이징 근거에 구체적이고 정합성 있게 반영되었는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

특히 일반 웹시스템과 달리 생성형 AI 서비스는 LLM(대형언어모델) 추론 요청 1건당 GPU 메모리·연산 비용이 수십~수백 배 높고, 동시 요청이 집중될 경우 응답 지연 또는 서비스 중단이 발생할 수 있다.

따라서 GPU 인스턴스 수량·메모리 용량, 추론 서버(Inference Server)의 배치 크기(Batch Size) 설정, 스케일아웃(Scale-out) 기준이 설계서에 명시되어 있는지 확인한다.

2. 필요성

▶ 생성형 AI 서비스는 일반 정보시스템과 근본적으로 다른 컴퓨팅 자원 특성을 가지며, 용량 산정이 부정확하면 서비스 가동 이후 수정이 매우 어렵고 비용이 크다.

▶ GPU 자원의 비선형적 소비 특성: LLM은 컨텍스트 길이(Context Length)에 따라 GPU VRAM 사용량이 비선형적으로 증가한다. 요청당 토큰 수를 설계서에 명시하지 않으면 동시 접속자 증가 시 메모리 부족(OOM)으로 서비스가 전면 중단될 수 있다.

▶ 공공 AI 서비스의 서비스 연속성 의무: 공공기관이 제공하는 AI 서비스는 민원·행정 처리와 직결되어 있어 서비스 중단 시 국민 불편이 발생한다. 요구사항에 명시된 가용성 목표(예: 99.9% 이상)를 충족하기 위한 이중화·스케일아웃 설계 근거가 설계서에 포함되어야 한다.

▶ 요구사항 → 설계 추적성 확보: 감리 기준상 요구사항이 설계 산출물에 빠짐없이 반영되었는지 추적성(Traceability)을 확인할 의무가 있다. 성능 요구사항이 산정 근거 없이 설계서에 누락되거나 모호하게 기술된 경우, 검사 단계에서 성능 기준 미달 여부를 판단할 수 없어 분쟁의 원인이 된다.

▶ AI 추론 서버 특유의 배치·큐 관리 설계 필요: 일반 웹서버와 달리 LLM 추론 서버(vLLM, TGI 등)는 요청을 배치로 묶어 처리하며, 큐(Queue) 대기 시간이 응답 지연에 직접 영향을 준다. 이에 대한 설계 방침이 없으면 사용자 체감 성능이 크게 저하될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	성능 요구사항 설계 반영 여부	{시스템요구사항}에 명시된 목표 동시접속자 수, 응답시간(평균/최대), TPS가 {시스템아키텍처설계서}의 인프라 사이징 근거 항목에 동일한 수치로 반영되었는지 확인
2	GPU 인스턴스 수량·사양 산정 근거	'AI 인프라 용량산정 보고서' 또는 {시스템아키텍처설계서}에서 GPU 유형(예: A100, H100 등), 수량, VRAM 용량이 목표 동시 요청 수를 처리할 수 있는 산정 산식과 함께 제시되었는지 확인
3	추론 서버 배치 크기 및 큐 설계	'AI 추론 서버 구성 설계서' 또는 {시스템아키텍처설계서}에서 배치 크기(Batch Size), 최대 대기 큐 길이, 큐 초과 시 처리 방침(거부/지연/우선순위)이 설계되어 있는지 확인
4	스케일아웃(Auto-scaling) 정책 설계	{시스템아키텍처설계서}에 CPU/GPU 부하 임계값 기반 스케일아웃 기준(예: GPU 활용률 80% 초과 시 인스턴스 추가)이 구체적으로 명시되었는지, 최소·최대 인스턴스 수가 기재되었는지 확인
5	네트워크 대역폭 및 스토리지 I/O 산정	{시스템아키텍처설계서}에 모델 가중치 로딩, RAG(검색증강생성) 데이터 조회, 로그 적재에 필요한 네트워크 대역폭 및 스토리지 I/O 용량이 산정되어 있는지 확인
6	부하 시나리오 및 피크 타임 가정	'AI 인프라 용량산정 보고서'에 일평균·피크 시간대 동시 요청 수, 요청당 평균 토큰 수(입력+출력), 피크 배율(Peak Factor) 가정치가 명시되어 있는지 검토
7	요구사항-설계 추적성(Traceability) 확보	{시스템요구사항}의 성능 요구사항 식별자(ID)가 {시스템아키텍처설계서}의 해당 항목에 참조(cross-reference)되어 있는지, 또는 별도의 요구사항 추적표가 존재하는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	동시접속·TPS·응답시간이 {시스템아키텍처설계서}에 요구사항과 동일한 수치로 명시되어 추적 가능하고, VRAM 산정 산식·GPU 수량 산출 근거·피크 배율 가정치, 배치 크기·최대 큐 길이·큐 초과 처리 방침, 스케일아웃 트리거 조건과 최소·최대 인스턴스 수, 모델 가중치 로딩·RAG 데이터 조회·로그 적재에 필요한 네트워크 대역폭과 스토리지 I/O 산정 결과가 모두 명시된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	성능 목표 수치는 반영되었으나 요구사항 추적표가 없거나 일부 성능 항목(가용성 등)이 누락된 경우, GPU 수량은 기재되었으나 산정 근거(산식, 가정치)나 피크 배율이 누락된 경우, 배치 설정이 '기본 설정'으로만 기술되거나 '필요 시 수동 확장' 등 임계값 없이 기술된 경우	산정 근거·정량 기준 보완 요청 후 재확인
부적합 (X)	{시스템아키텍처설계서}에 성능 목표 수치가 전혀 기재되지 않았거나, 인프라 사이징의 산정 근거가 확인되지 않는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 수치의 구체성 확인이 핵심: "충분한 GPU를 확보한다"와 같은 추상적 표현은 판단 기준이 없다. 성능 항목은 반드시 측정 가능한 수치(예: "동시 사용자 300명 기준 응답시간 5초 이내")로 기재되어야 한다. 수치가 없는 항목은 보완 권고 대상으로 즉시 지적한다.

2) 요구사항과 설계 수치의 불일치 패턴 주의: 요구정의 단계에서 "동시 접속 1,000명"으로 요건이 명시되었으나, 설계서에는 "동시 접속 500명"으로 축소된 사례가 발생한다. 이는 사업자가 비용 절감을 위해 요건을 임의로 조정한 경우이다. 반드시 양 문서를 병렬 대조하여 불일치를 확인한다.

3) 클라우드 vs. 온프레미스 구분 확인: 클라우드 기반 사업에서는 인스턴스 유형(예: AWS p4d.24xlarge)이 설계서에 명기되어야 하고, 온프레미스 사업에서는 물리 서버 사양(GPU 칩 모델, 수량, 서버 수)이 명기되어야 한다. 두 환경이 혼재된 하이브리드 구성의 경우, 각 구성 요소의 역할 분담과 자원 배분 기준이 설계서에 명확히 구분되어야 한다.

4) RAG(검색증강생성) 구성 여부 확인: 서비스에 RAG가 포함된 경우, 벡터 데이터베이스(Vector DB) 조회 시간과 I/O 용량이 추론 서버의 응답시간 산정에 포함되어야 한다. RAG 조회 지연이 전체 응답시간에서 차지하는 비중이 LLM 추론 자체보다 클 수 있으므로, 설계서에 RAG 조회 지연 목표값이 별도로 명시되어 있는지 확인한다.

5) 피크 배율(Peak Factor) 가정의 현실성 검증: "피크 배율 1.2×" 등 낮은 피크 배율을 가정한 경우, 공공기관의 특성(행정 처리 마감일, 특정 업무 시간대 집중 등)을 고려하여 적정성을 검토한다. 피크 배율 가정치와 그 근거가 명시되지 않은 경우는 보완 권고 대상이다.

AG_SD41-1-08 | 컨테이너 보안 요건점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 클라우드컨테이너

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · OWASP LLM Top 10 · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 컨테이너(Docker/Kubernetes) 기반 AI 시스템 인프라에 대한 보안요구사항이 구체적으로 정의되었는지, 컨테이너 이미지 취약점 점검 요건이 명시되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 컨테이너 환경은 기존 물리 서버 환경과 달리 이미지 공급망(Supply Chain) 취약점, 컨테이너 탈출(Container Escape), 오케스트레이션 플랫폼의 권한 오남용 등 AI 시스템 특유의 보안 위협이 존재하므로, 이러한 위험 요소가 요구사항으로 식별되고 보안요건으로 반영되었는지를 확인한다.

2. 필요성

▶ 생성형 AI 시스템은 GPU 클러스터, 모델 서빙 인프라, 데이터 파이프라인 등을 Docker/Kubernetes(K8s) 컨테이너 환경에서 구동하는 경우가 대부분이며, 컨테이너는 배포 속도와 확장성에서 장점이 있으나 보안 측면에서 기존 물리 서버와 다른 고유한 위험을 내포하고 있다.

▶ 공급망 공격 방지: 공개 컨테이너 레지스트리(Docker Hub 등)에서 내려받은 이미지에 악성코드나 알려진 취약점(CVE)이 포함될 수 있으며, 이를 요구정의 단계에서 점검 요건으로 명시하지 않으면 운영 단계에서 사고가 발생할 수 있다.

▶ 컨테이너 탈출 위험: 컨테이너가 호스트 OS 커널을 공유하는 구조상, 컨테이너 탈출(Container Escape) 공격으로 호스트 전체 및 동일 노드의 타 컨테이너가 침해될 수 있어, 기획 단계부터 격리(Isolation) 및 권한 최소화 요건을 정의해야 한다.

▶ K8s 플랫폼 보안: Kubernetes 클러스터는 API Server, etcd, kubelet 등 다수의 보안 포인트를 가지며, RBAC(역할 기반 접근제어), Network Policy, Pod Security 등의 보안 요건이 요구사항으로 반영되어야 한다.

▶ AI 모델/데이터 보호: AI 시스템의 학습 데이터와 모델 가중치(Weight)가 컨테이너 내부에 저장되므로, 컨테이너 이미지 무결성 검증 및 암호화 요건이 필수적이다.

3. 검토방법

No	점검 항목	확인 방법
1	컨테이너 베이스 이미지 출처 관리 요건 명시 여부	{보안요구사항}에서 공인 레지스트리 사용·신뢰 이미지만 허용 등 베이스 이미지 출처 관리 요건이 명시되었는지 직접 검토
2	컨테이너 이미지 취약점 스캔(CVE 점검) 수행 요건 명시 여부	{보안요구사항}·'사업수행계획서'에서 빌드·배포 전 필수 취약점 스캔(CVE 점검) 요건이 명시되었는지 검토
3	컨테이너 이미지 서명 및 무결성 검증 요건 명시 여부	{보안요구사항}에서 컨테이너 이미지 서명(Signing)과 무결성 검증 요건이 명시되었는지 검토
4	컨테이너 이미지 경량화 및 불필요 패키지 제거 요건 반영 여부	{보안요구사항}·'컨테이너 이미지 관리 기준서'에서 이미지 경량화(Minimal Base Image)와 불필요 패키지 제거 요건이 반영되었는지 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	베이스 이미지 출처 관리(신뢰할 수 있는 레지스트리 사용), 이미지 취약점 스캔(CVE 점검), 이미지 서명·무결성 검증, 이미지 경량화 요건이 {보안요구사항}에 구체적이고 측정 가능하게 명시되어 있는 경우	현행 유지
보완권고 (△)	컨테이너 보안 요건은 정의되어 있으나 이미지 서명·무결성 검증, 경량화 등 일부 항목이 누락되었거나, 스캔 수행 시점·차단 기준 등이 모호하게 기술되어 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	컨테이너(Docker/K8s) 보안 요구사항 및 이미지 취약점 점검 요건이 요구사항에 전혀 반영되지 않았거나, '컨테이너 보안을 적용한다' 수준의 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 요건의 수치화: '컨테이너 보안을 적용한다'와 같은 모호한 문구는 향후 분쟁의 씨앗이 된다. 'Critical/High CVE가 있는 이미지는 배포 차단', 'RBAC 정책은 최소 권한 원칙 적용' 등 측정 가능한 검사 기준으로 구체화하도록 유도해야 한다.

2) 솔루션 특정 점검 지양: 감리원은 특정 벤더의 보안 솔루션(예: Aqua Security, Prisma Cloud) 사용 여부를 점검하는 것이 아니라, 보안 요건이 요구사항으로 명시되었는지를 점검해야 한다. 사업자의 기술 선택권을 존중하되, 요구사항의 구체성과 측정 가능성에 집중한다.

3) 운영 산출물의 기획 단계 확인: 운영 과정의 점검사항은 본 가이드라인 범위가 아니지만, 개발 과정에서 만들어야 하는 운영 산출물(예: '컨테이너 보안 정책서', '컨테이너 이미지 관리 기준서')의 작성 계획이 사업수행계획서에 포함되었는지 확인한다.

4) AI 특유 컨테이너 보안 요건 구분: 일반적인 컨테이너 보안(취약점 스캔, RBAC 등) 외에 AI 시스템 특유의 보안 요건(GPU 격리, 모델 가중치 보호, 학습 데이터 암호화 등)이 요구사항에 반영되었는지에 특별히 주목한다. 이는 NIST AI RMF의 'Secure and Resilient' 특성과 직접 관련된다.

5) 면담 시 확인 사항: 사업자 면담 시 '컨테이너 이미지를 어디서 가져오는가?', '빌드 시 취약점 스캔을 수행하는가?', 'K8s 클러스터에 RBAC이 적용되어 있는가?' 등의 구체적인 질문을 통해 요구사항의 실행 가능성을 함께 확인한다.

AG_SD41-1-09 | CI/CD 파이프라인 보안 요건

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | CI/CD파이프라인

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「개인정보 보호법」 · NIST AI RMF · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 개발·학습·배포 과정을 자동화하는 CI/CD(Continuous Integration/Continuous Deployment) 파이프라인에 대한 보안 요건과, 학습 완료된 모델을 운영 환경에 안전하게 자동 배포하기 위한 환경 요건이 {시스템요구사항}에 구체적이고 상세하게 정의되었는지를 감리원이 점검하기 위한 기준을 제시하는 것을 목적으로 한다.

AI 시스템의 CI/CD 파이프라인은 일반 소프트웨어와 달리 모델 가중치(Weight) 파일, 학습 데이터셋, 하이퍼파라미터 설정 등 대용량·고민감도 자산이 파이프라인 상에서 이동하며, 모델 배포 시 무결성이 훼손되면 서비스 전체의 신뢰성이 저하된다.

따라서 기존 DevSecOps 관점의 일반 보안 점검과는 별도로, AI 특유의 파이프라인 보안 위험과 모델 배포 환경 보안 요건이 기획 단계에서부터 요구사항으로 명확히 정의되어야 한다.

2. 필요성

▶ 생성형 AI 시스템의 CI/CD 파이프라인은 모델 학습 코드, 학습 데이터셋, 모델 가중치, API 키, 추론 서비스 설정 파일 등 일반 소프트웨어 파이프라인보다 훨씬 민감한 자산을 다룬다. 파이프라인 보안이 미흡하면 모델 가중치 탈취, 학습 데이터 유출(개인정보 포함), 악성 코드 주입을 통한 모델 오염(Data Poisoning) 등 심각한 보안 사고로 이어질 수 있다.

▶ AI 파이프라인 자산 보호: 모델 가중치 파일, 학습 데이터, Fine-tuning 스크립트, 프롬프트 템플릿 등은 조직의 핵심 지식재산(IP)이자 보안 민감 자산이다. 파이프라인 단계별 접근통제·암호화·감사 로깅 요건이 기획 단계에서 정의되지 않으면, 이후 설계·구현 과정에서 보안 공백이 발생한다.

▶ 모델 무결성 및 공급망 보안: AI 모델 배포 시 모델 파일의 해시 검증, 서명 기반 무결성 확인, 모델 레지스트리 접근통제 등의 요건이 사전 정의되어야 한다. NIST AI RMF에서도 AI 시스템 공급망 전반에 걸친 보안 위험 관리를 강조하고 있다.

▶ 배포 환경 분리 및 롤백 체계: 개발·스테이징·운영 환경 간 명확한 분리와, 배포 실패 또는 모델 이상 감지 시 자동 롤백 절차가 요구사항으로 명시되어야 장애 발생 시 신속한 대응이 가능하다.

▶ 규제 준수 및 감사 추적성: AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 2026.1.22 시행), 개인정보 보호법 등 국내 규제와 NIST Secure Software Development Framework(S

SDF) 등 국제 기준에서 소프트웨어 공급망 보안을 요구하고 있으며, 이를 AI 파이프라인에 적용한 요구사항이 기획 단계에서 정의되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	코드 저장소 접근통제·인증 요건 명시 여부	파이프라인 코드 저장소(Git 등)에 대한 접근통제·인증 요건이 {시스템요구사항}에 명시되었는지 확인
2	시크릿 관리 방안 포함 여부	API 키·인증 토큰·DB 접속 정보 등 모델 학습·추론에 사용되는 시크릿의 관리 방안이 {시스템요구사항}·'아키텍처 설계서(초안)'에 포함되었는지 확인
3	파이프라인 실행 로그 기록·보관 요건 포함 여부	빌드·학습·배포 이력 등 파이프라인 실행 로그의 기록·보관 요건이 {시스템요구사항}에 포함되었는지 확인하고 사업자 면담으로 보완
4	공급망 보안(SBOM) 요건 포함 여부	서드파티 라이브러리·프레임워크의 공급망 보안(SBOM 등) 요건이 {시스템요구사항}·{과업대비표}에 포함되었는지 확인
5	학습 데이터 암호화 요건 명시 여부	파이프라인에서 처리되는 학습 데이터의 전송 중·저장 시 암호화 요건이 {시스템요구사항}에 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	파이프라인 코드 저장소 접근통제, 시크릿 관리, 실행 로그 기록·보관, 공급망 보안(SBOM), 학습 데이터 암호화 등 CI/CD 보안 요건이 {시스템요구사항}에 구체적으로 정의된 경우	현행 유지
보완권고 (△)	CI/CD 보안 요건이 일부 정의되었으나, 시크릿 관리 방안 미정의, 공급망 보안 요건 누락, 요건의 구체성 부족 등 일부 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	CI/CD 파이프라인 보안 및 모델 배포 환경에 대한 요건이 {시스템요구사항}에 전혀 정의되지 않아 설계·구현 단계의 보안 공백이 예상되는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 보안 요건의 수치화 유도: 'CI/CD 보안을 적용한다', '모델을 안전하게 배포한다'와 같은 모호한 문구는 이후 설계·구현 단계에서 해석 분쟁의 원인이 된다. '파이프라인 코드 저장소 접근 시 MFA 적용 필수', '모델 배포 전 SHA-256 해시 비교 검증 의무화', '배포 실패 시 5분 이내 자동 롤백 완료' 등 측정 가능한 보안 기준으로 구체화하도록 유도하여야 한다.

2) AI 파이프라인과 일반 CI/CD의 차이 인식: AI 파이프라인은 일반 소프트웨어 CI/CD와 달리 대용량 모델 파일(수 GB), 학습 데이터셋(개인정보 포함 가능), GPU 자원 접근 권한 등 추가적인 보안 고려사항이 존재한다. 일반 DevSecOps 관점의 점검만으로는 AI 특유의 보안 위험을 포괄할 수 없으므로, AI 파이프라인 고유의 보안 요건이 별도로 정의되었는지 확인하여야 한다.

3) 발주기관의 보안 정책과의 정합성 확인: 발주기관이 자체적으로 운영하는 정보보안 정책(망분리 환경, 클라우드 사용 제한 등)과 CI/CD 파이프라인 보안 요건 간의 충돌 여부를 면담을 통해 사전에 확인하여야 한다. 특히 공공기관의 경우 망분리 환경에서 외부 모델 허브(Hugging Face 등) 접근 제한 등이 파이프라인 설계에 중대한 영향을 미칠 수 있다.

4) 과도한 기술 구현 수준 요구 지양: 감리원은 특정 도구(Jenkins, GitLab CI, ArgoCD 등)나 기술 스택의 사용을 강제하는 것이 아니라, '역할 기반 접근통제가 요구사항에 정의되었는가', '모델 무결성 검증 절차가 명시되었는가' 등 요건의 정의 여부와 구체성 수준을 점검하는 데 집중하여야 한다.

5) 클라우드·온프레미스 환경별 차별 점검: 클라우드 기반 CI/CD(AWS CodePipeline, GCP Cloud Build 등)의 경우 CSP(Cloud Service Provider)의 공유 책임 모델에 따른 보안 경계가 요구사항에 반영되었는지 확인하고, 온프레미스 환경의 경우 파이프라인 서버 자체의 보안(패치 관리, 접근 로그 등) 요건이 포함되었는지 확인하여야 한다.

AG_SD41-1-10 | 목적기반LegacyAPI설계점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | MCP/Tool연계

근거 | 「개인정보 보호법」·「신용정보의 이용 및 보호에 관한 법률」·「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) ·「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 금융권 Agentic AI 서비스의 요구정의 단계에서, AI Agent가 접근하는 Legacy API가 단순 기술 인터페이스가 아닌 '업무 접근 통제 지점'으로서 목적기반(Purpose-driven)으로 설계되도록 요건이 정의되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 {요구사항정의서}와 'Agent 연계 API 명세서', '인터페이스설계서'를 근거로 ① 테이블·스키마 직접 노출 금지와 반환 필드 최소화, ② 업무 목적(purpose)의 API 명세 내재화, ③ 민감정보 마스킹의 응답 단계 내장, ④ 등록·발송·거래 실행의 Draft-승인 분리, ⑤ caseId·agentId·purpose 헤더 기반 E2E 감사 추적 요건을 확인한다.

2. 필요성

▶ 개인정보 과다 노출 차단: 테이블 전체나 원시 스키마를 노출하면 상담 수행에 불필요한 개인정보까지 Agent에 반출된다. 업무 수행에 필요한 식별·요약 필드만 반환하도록 API가 설계되어야 한다.

▶ 업무 목적의 내재화: '무엇을 위해 호출하는가(Purpose)'가 API 명세에 내재화되어야 보안 통제와 컴플라이언스를 동시에 달성할 수 있다. 범용 조회 API를 Agent에 그대로 개방하면 목적 외 이용을 통제할 수단이 없다.

▶ 등록·실행의 승인 강제: 등록·발송·거래 실행을 Agent가 단독으로 확정하면 프롬프트 인젝션 한 번으로 되돌릴 수 없는 처리가 발생한다. Draft 생성과 사람의 승인, 확정을 분리하는 프로세스가 강제되어야 한다.

▶ 감사 추적의 내재화: caseId·agentId·purpose를 헤더로 전달하여 Agent의 모든 호출이 End-to-End로 추적되어야 사고 시 책임 소재를 규명할 수 있다. 사후에 로그를 덧붙이는 방식으로는 추적 완전성이 확보되지 않는다.

▶ 요구정의 단계 확정의 필요성: API의 목적기반 재설계는 Legacy 시스템의 인터페이스 변경을 수반하므로, 설계·구현 이후에는 사실상 적용이 불가능하다. 요건으로 확정해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	테이블·원시 스키마 직접 노출 금지	'Agent 연계 API 명세서'에서 SELECT* 형태의 조회나 원시 테이블 스키마가 Agent에 노출되지 않도록 요건이 정의되었는지 확인
2	업무 목적(purpose)의 API 명세 내재화	API명과 파라미터에 업무 컨텍스트(purpose)가 반영되고, 목적 외 호출이 거부되도록 요건이 {요구사항정의서}에 정의되었는지 확인
3	반환 필드 최소화	업무 수행에 필요한 식별·요약 필드만 반환하도록 필드 목록이 'Agent 연계 API 명세서'에 API별로 정의되고, 불필요 필드 제거 요건이 명시되었는지 확인
4	민감정보 마스킹의 응답 단계 내장	계좌번호·주민등록번호 등 민감정보가 API 응답 단계에서 마스킹되도록 요건이 정의되었는지 '인터페이스설계서'에서 확인
5	등록·발송·거래 실행의 Draft-승인 분리	Agent가 등록·발송·거래를 단독 확정할 수 없도록 Draft 생성 → 사람 승인 → 확정의 3단계 프로세스가 {요구사항정의서}에 요건으로 정의되었는지 확인
6	caseId·agentId·purpose 헤더 기반 E2E 추적	모든 Agent 호출에 caseId·agentId·purpose 헤더가 필수로 포함되고 Legacy 시스템까지 전파되어 End-to-End 추적이 가능하도록 '인터페이스설계서'에 요건이 정의되었는지 확인
7	호출 권한과 목적의 정합성 검증	Agent에 부여된 API 호출 권한이 선언된 purpose 범위를 벗어나지 않도록 {요구사항정의서}에 검증 요건이 정의되었는지 확인하고, 권한 부여·회수 절차를 면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	원시 테이블·스키마의 직접 노출 금지, 업무 목적(purpose)의 API 명세 내재화, 반환 필드 최소화, 민감정보의 응답 단계 마스킹, 등록·발송·거래의 Draft-승인 분리, caseId·agentId·purpose 헤더 기반 E2E 추적, 호출 권한과 선언된 purpose 범위의 정합성 검증 요건이 구체적으로 명시되어 API 접근통제 요건이 완비된 경우	현행 유지
보완권고 (△)	API는 정의되었으나 마스킹·Draft 분리 요건이 일부 미흡한 경우	요건 구체화 요청 후 재점검
부적합 (X)	테이블 전체 조회·Agent 단독 등록을 허용하는 등 목적 기반 통제가 부재한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) API는 업무 통제 지점: API가 '테이블을 그대로 보여주는 창'이면 Agentic 보안은 성립하지 않는다. API는 '무엇을 위해' 호출하는지를 담는 업무 통제 지점이어야 한다.

2) purpose 의무화·E2E 추적: API 명세에 purpose 파라미터·헤더를 의무화하고, caseId·agentId·purpose 헤더 기반의 E2E 호출 추적이 보장되는지 확인한다.

3) 응답 단계 마스킹 확인: 마스킹이 화면단에서만 적용되고 API 응답은 원본을 반환하는 사례가 빈번하므로, 계좌번호는 은닉 후 자산군별 비중 등 가공 데이터로 응답하는지 점검한다.

4) Draft/승인 분리 시연 확인: 등록·발송·거래는 Draft 상태로만 생성한 후 승인 시 확정되도록 하고, 개발자에게 Agent 단독 거래 실행 차단을 시연으로 확인하며, 발주기관에 중요 액션 승인 권한 주체를 확인한다.

5) 가드레일 연계 점검: 목적기반 API의 마스킹·승인 요건이 가드레일의 Data 계층 및 HITL 승인 흐름과 일치하는지 함께 점검한다.

AG_SD41-1-11 | DDoS_API_RateLimit 요건점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | API보안

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템이 수반하는 대량 추론 요청(Inference Request)과 외부 API 연계 트래픽에 대한 과부하 방어 및 접근 통제 요구사항이 보안요구사항에 구체적이고 정량적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

생성형 AI 서비스는 사용자의 자연어 질의 한 건이 내부적으로 대규모 GPU 연산과 다단계 외부 API 호출(검색 증강 생성(RAG), 외부 LLM API, 벡터 DB 질의 등)을 유발하여, 일반 웹 서비스 대비 단일 요청당 자원 소모량이 수십~수백 배에 달한다.

이로 인해 소규모 트래픽 폭증만으로도 서비스 가용성에 치명적 영향을 줄 수 있으므로, DDoS 방어, 과부하 차단, API Rate Limit 등의 보안 통제 요건이 요구사항으로 명확히 정의되어야 하며, '시스템 구조 및 보안' 영역 감리원은 해당 상세 검토 항목을 적용하여 요구사항의 구체성과 정량성을 점검한다.

2. 필요성

▶ AI 추론 자원 보호: 생성형 AI 시스템은 추론 요청 1건당 GPU 연산 자원을 대량으로 소모하므로, DDoS 또는 비정상적 대량 요청이 발생하면 일반 웹 서비스보다 훨씬 적은 요청 수로도 서비스 전체가 마비될 수 있다. 요구사항에 AI 추론 엔드포인트 전용 트래픽 제어 정책이 명시되지 않으면 설계·구현 단계에서 방어 체계가 누락될 수 있다.

▶ 외부 API 과금 폭증 방지: 외부 LLM API, 검색 API 등을 연계하는 구조에서 Rate Limit이 요구사항에 정의되지 않으면, 비정상 트래픽 발생 시 과도한 API 호출로 인해 예산 초과 또는 외부 서비스 차단이 발생하며, 이는 AI 서비스 전체 가용성에 직접적 영향을 미친다.

▶ 사내 정보 유출 경로 차단: AI 시스템과 외부 API 간 데이터 전송 과정에서 프롬프트에 포함된 사내 기밀정보, 개인정보가 외부로 유출될 수 있다. 요구사항에 외부 연계 구간의 데이터 필터링 및 접근 통제 요건이 반영되지 않으면, 보안 사고의 근본 원인이 된다.

▶ 정량적 보안 기준 확보: "적절한 보안 조치를 수행한다"와 같은 모호한 요구사항은 설계·검수 단계에서 검증이 불가능하다. "API 호출 Rate Limit: 사용자당 분당 N회 이하", "DDoS 방어 임계값: 초당 N건 이상 자동 차단" 등 측정 가능한 요건을 정의해야, 후속 단계 감리에서 설계·구현의 적정성을 검증할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 특성 반영 DDoS 방어 요구사항 명시 여부	저빈도·고자원 소모형 AI 추론 엔드포인트 특성을 반영한 DDoS 방어 요구사항이 {보안요구사항}에 명시되었는지 확인
2	트래픽 구분 과부하 탐지·차단 정책 정의 여부	일반 웹 트래픽과 AI 추론 트래픽을 구분한 과부하 탐지·차단 정책이 {보안요구사항}·'아키텍처 설계서(초안)'에 정의되었는지 확인
3	외부 연계 서비스 API Rate Limit 정책 포함 여부	외부 LLM API·검색 API 등 외부 연계 서비스에 대한 API Rate Limit 정책이 {보안요구사항}·'외부 API 연계 명세서'에 포함되었는지 확인
4	사용자·역할별 접근 횟수 제한 요건 정의 여부	사용자별·역할별 AI 서비스 접근 횟수 제한(Rate Limit) 요건이 {보안요구사항}에 정의되었는지 확인
5	과부하 시 Graceful Degradation 전략 반영 여부	AI 추론 요청 과부하 시 단계적 서비스 축소(Graceful Degradation) 전략이 {보안요구사항}·{요구사항정의서}에 반영되었는지 확인
6	외부 API 구간 데이터 필터링 요건 포함 여부	외부 API 연계 구간의 사내 기밀정보·개인정보 유출 방지 데이터 필터링 요건이 {보안요구사항}에 포함되었는지 확인하고 면담으로 보완
7	Rate Limit·DDoS 로깅·모니터링·알림 요건 명시 여부	API Rate Limit 초과·DDoS 탐지 이벤트에 대한 로깅·모니터링·알림 요건이 {보안요구사항}에 명시되었는지 확인
8	외부 API 장애 시 대체 운영(폴백) 방안 반영 여부	외부 API 장애·차단 시 AI 서비스의 대체 운영(폴백) 방안이 {보안요구사항}·{요구사항정의서}에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 추론 엔드포인트 특성을 반영한 DDoS 방어·과부하 차단 정책, 외부 API 및 사용자별 Rate Limit, Graceful Degradation, 데이터 필터링, 로깅·모니터링, 폴백 요건이 측정 가능한 수치로 {보안요구사항}에 정의된 경우	현행 유지
보완권고 (△)	DDoS 방어·Rate Limit 등 주요 요건은 정의되었으나, 일부 요건(사용자별 호출 한도, 과부하 대응 전략, 데이터 필터링, 폴백 등)이 누락되었거나 정량 기준 없이 선언적으로만 기술된 경우	해당 항목 보완 요청 후 재확인 (개인정보 관련 요건 미준수 발견 시 즉시 보완 요구)

판정	판정 기준	주요 사례 / 조치
부적합 (X)	AI 추론 트래픽에 대한 과부하 방어·API Rate Limit 요건이 요구사항에 전혀 정의되지 않았거나, 일반 웹 서비스 기준만 존재하여 AI 서비스의 가용성·비용 통제가 불가능한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 보안요구사항의 수치화 유도: "적절한 보안 대책을 수립한다", "DDoS 방어를 적용한다"와 같은 선언적 문구는 후속 감리 단계에서 설계·구현 적정성 판단이 불가능하다. "AI 추론 엔드포인트: 초당 N건 초과 시 자동 차단", "외부 LLM API Rate Limit: 분당 N회, 초과 시 큐잉 처리" 등 측정 가능한 수치로 구체화하도록 보완을 권고한다.

2) AI 추론 엔드포인트와 일반 웹 엔드포인트 구분 확인: 사업자가 DDoS 방어 요건을 제시하더라도, 일반 웹 서비스 기준을 AI 추론 API에 동일하게 적용하면 실효성이 없다. AI 추론 요청의 자원 소모 특성을 고려하여 별도 임계값이 설정되었는지 반드시 확인한다.

3) 외부 API 과금 상한선(Budget Cap) 확인: 외부 LLM API를 활용하는 구조에서 Rate Limit만 정의하고 과금 상한선을 미설정하면, Rate Limit 이내의 정상 호출만으로도 예산을 초과할 수 있다. 과금 상한선이 요구사항에 포함되었는지 별도로 점검한다.

4) 영역 간 협의 사항 확인: 본 가이드라인은 시스템 구조 및 보안 영역의 관점이나, DDoS 방어·Rate Limit은 응용시스템(API Gateway 설계), 데이터베이스(로그 저장 구조), 사업관리(보안 요건 관리) 영역과도 연관된다. 타 영역 감리원과 사전에 점검 범위를 조율하여 중복·누락을 방지한다.

5) 운영 산출물 기획 여부 확인: 운영은 직접 점검 대상이 아니지만, 개발 과정에서 작성해야 할 운영 산출물('보안 모니터링 정책서', '보안 이벤트 대응 절차서', '외부 API Rate Limit 운영 기준서' 등)이 사업수행계획서의 산출물 목록에 포함되었는지 확인한다. 이들 산출물이 계획되지 않으면 운영 단계에서 보안 공백이 발생한다.

AG_SD41-1-12 | SecurityGuideline모델 공격 방어 요건

시스템구조·보안 | 요구정의·분석 | 인공지능 모델 개발 | AI특화보안

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)

1. 목적

AI 서비스 구축 사업에서 {보안요구사항}에 모델 공격(Adversarial Attack)에 대한 방어 요건이 충분한 구체성과 측정 가능성을 갖추어 정의되었는지를 제3자 관점에서 검토하는 것이 목적이다.

특히 회피 공격(Evasion Attack), 프롬프트 인젝션(Prompt Injection), 탈옥(Jailbreak) 등 생성형 AI에 고유한 보안 위협이 분석되었는지, 그리고 이에 대응하는 방어 요건이 향후 설계·구현·시험 단계에서 검증 가능한 형태로 명세되었는지를 확인하는 데 초점을 둔다.

2. 필요성

▶ 생성형 AI 시스템은 일반 소프트웨어와 달리 입력 데이터의 미세한 변형만으로도 모델이 오작동하거나 유해 콘텐츠를 생성하는 공격에 취약하다. 요구정의 단계에서 이러한 AI 고유 보안 위협이 구체적으로 분석되지 않으면, 설계 및 구현 단계에서 대응 수단이 누락되고, 시험 단계에서도 검증 기준이 모호해져 보안 결함이 운영 단계까지 이월될 수 있다.

▶ 또한 NIST AI RMF 1.0 및 TTA 2024 생성AI 개발안내서는 AI 시스템의 보안성(Secure & Resilient) 확보를 위해 개발 초기 단계부터 적대적 공격 유형을 식별하고, 이에 대한 방어 요건을 명세할 것을 권고하고 있다. 감리원은 이러한 기준에 비추어 보안 분석이 충실히 수행되었는지 확인하여야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	보안 위협 유형 식별 및 분류 여부	{보안요구사항}에서 회피 공격(Evasion), 데이터 포이즈닝 공격(Poisoning), 모델 추출 공격(Extraction), 멤버십 추론 공격(Membership Inference), 프롬프트 인젝션 등 AI 고유 공격 유형이 식별·분류되어 있는지 확인
2	공격 위협 시나리오의 구체성 및 서비스 적합성	위협 시나리오가 구축 대상 AI 서비스의 특성(업무 도메인, 사용자 유형, 외부 연계 여부 등)에 맞게 구체화되어 있는지 확인
3	방어 요건의 명세 수준 및 측정 가능성	식별된 각 위협에 대응하는 방어 요건이 {보안요구사항}에 명확히 기술되어 있는지 확인

No	점검 항목	확인 방법
4	입력값 검증 및 출력 필터링 요건 포함 여부	사용자 입력에 대한 유효성 검증(길이, 형식, 특수문자, 인코딩 우회 등)이 요건으로 정의되었는지 확인
5	위협 분석의 제3자 검토 또는 전문가 참여 여부	보안 위협 분석 활동에 보안 전문가 또는 제3자 검토자가 참여하였는지, 관련 이력(검토 회의록, 서명 등)이 산출물에 남아있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{보안요구사항}에 AI 모델 공격 유형이 구체적으로 식별되고 위협 시나리오가 구축 대상 AI 서비스의 특성에 맞게 구체화되며, 각 위협에 대응하는 방어 요건과 입력값 유효성 검증·출력 필터링 요건이 측정 가능한 기준으로 명세되고, 위협 분석에 보안 전문가 또는 제3자 검토자가 참여한 이력이 산출물에 남아 있는 경우	현행 유지
보완권고 (△)	일부 공격 유형이 누락되거나, 방어 요건이 선언적 수준에 그쳐 측정·검증 가능성이 낮은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안요구사항}에 AI 모델 보안 위협 관련 내용이 전혀 없거나, 위협 분석 산출물 자체가 존재하지 않는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) '공격 방어 기능을 구현한다'와 같은 선언적 표현이 있으면 구체적인 방어 메커니즘(입력 검증 방식, 탐지 임계값 등)을 기재하도록 보완을 요구한다. 감리원은 방어 요건이 측정 가능한 수치·기준과 함께 명세되었는지 확인한다.

2) 서비스 유형에 따라 위협 모델이 달라짐을 인지하고 점검한다: 분류 모델은 회피·포이즈닝 공격에, 생성형 AI는 프롬프트 인젝션·탈옥·모델 추출 공격에 취약하므로 서비스 특성에 맞지 않는 방어 요건은 실효성이 없다.

3) 방어 요건에 대한 검증 방법(수용 기준)이 명시되어 있는지 확인한다: '공격을 방어한다'는 요건에 대해 어떤 시험으로 합격 여부를 판정하는지가 없으면 감리·분쟁 소지가 생긴다.

4) 입력 검증과 출력 필터링을 모두 요건화했는지 확인한다: 입력 측 방어만 정의하고 유해 콘텐츠·민감정보에 대한 출력 필터링이 빠지면 탈옥이 성공했을 때 피해를 차단할 수 없다. 감리원은 입력 유효성 검증(길이·형식·특수문자·인코딩 우회)과 출력 필터링이 모두 {보안요구사항}에 정의되었는지 확인한다.

5) 위협 분석의 제3자 검토 이력을 확인한다: 개발팀이 스스로 수행한 위협 분석은 이미 인지된 위협만 담기 쉽다. 감리원은 보안 전문가 또는 제3자 검토자의 참여 이력(검토 회의록·서명 등)이 남아 있는지, 검토 시점이 요구사항 확정 이전인지를 확인한다.

AG_SD41-1-13 | 모델저장소 Registry 보안요건

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | AI특화보안

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서, 학습·추론에 사용되는 AI 모델을 저장·관리하는 모델 저장소(Model Registry)에 대한 보안 요구사항이 구체적으로 정의되었는지를 점검하기 위한 것이다.

모델 저장소는 모델 파일(.pt, .onnx, .safetensors 등), 가중치(Weight), 메타데이터, 버전 이력 등을 중앙에서 관리하는 핵심 인프라이며, 이 저장소의 접근통제·무결성 검증·배포 승인 절차가 기획 단계에서 보안요구사항으로 명확히 수립되지 않으면 비인가 모델의 운영 환경 투입, 모델 변조에 의한 추론 결과 왜곡 등 심각한 보안 사고로 이어질 수 있다.

따라서 본 가이드라인은 {보안요구사항} 산출물 내에 모델 저장소 보안과 관련된 접근통제 정책, 무결성 검증 메커니즘, 인가된 모델만 배포되도록 하는 승인 절차 등이 측정 가능하고 검증 가능한 수준으로 기술되어 있는지를 감리원이 체계적으로 점검할 수 있도록 검토 방향, 검토 항목, 점검 결과 작성 기준을 제시한다.

2. 필요성

▶ 생성형 AI를 포함한 AI 서비스 구축 사업에서 모델 저장소는 전통적인 소프트웨어 형상관리 대상과는 본질적으로 다른 보안 위협에 노출된다. 모델 파일은 수 GB~수백 GB에 달하는 대용량 바이너리로, 내부에 악성 코드가 삽입(Model Poisoning)되거나 가중치가 미세 변조되더라도 일반적인 코드 리뷰나 정적 분석으로는 탐지하기 어렵다.

▶ 모델 저장소에 대한 보안 요구사항이 기획 단계에서 누락되면 다음과 같은 위험이 발생할 수 있다.

▶ 비인가자가 모델 저장소에 접근하여 모델 파일을 무단 교체하거나 변조된 모델을 등록할 경우, 추론 서비스 전체의 신뢰성이 훼손된다.

▶ 모델 배포 시 무결성 검증(해시 비교, 디지털 서명 등) 절차가 없으면 공급망 공격(Supply Chain Attack)에 의해 변조된 모델이 운영 환경에 그대로 배포될 위험이 있다.

▶ 모델 버전 이력 및 승인 절차가 부재하면 실험 단계의 미검증 모델이 운영에 투입되어 예측 정확도 저하, 편향(Bias) 확대 등 서비스 장애를 유발할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	모델 저장소 접근통제 정책 정의	모델 저장소에 대한 역할 기반 접근통제(RBAC) 정책이 {보안요구사항}에 구체적 역할(모델 개발자, 검증자, 배포 관리자 등)과 권한(읽기/쓰기/삭제/배포 승인)으로 명시되어 있는지 확인
2	모델 파일 무결성 검증 요건	모델 등록·다운로드 시 해시값(SHA-256 이상) 비교 또는 디지털 서명 검증 절차가 {보안요구사항}에 정의되어 있는지 확인
3	모델 배포 승인 프로세스 정의	스테이징 → 프로덕션 배포 시 다단계 승인(최소 2인 이상) 절차, 승인 권한자, 승인 기록 보관 요건이 명시되어 있는지 확인
4	모델 버전 관리 및 이력 추적 요건	모델 버전별 메타데이터(학습 데이터셋, 하이퍼파라미터, 성능 지표), 변경 이력, 롤백 절차가 {시스템요구사항}에 정의되어 있는지 확인
5	모델 저장소 감사 로그 요건	모델 등록/수정/삭제/다운로드/배포 등 모든 주요 행위에 대한 감사 로그(행위자, 시각, 대상, 행위 유형) 기록 요건이 정의되어 있는지 확인
6	외부 모델 도입 시 보안 검증 요건	사전학습 모델, 오픈소스 모델 등 외부에서 도입하는 모델에 대한 보안 검증(출처 확인, 취약점 스캔, 라이선스 검토) 절차가 정의되어 있는지 확인
7	모델 저장소 암호화 및 통신 보안 요건	저장 시 암호화(Encryption at Rest) 및 전송 시 암호화(TLS 1.2 이상) 요건이 {보안요구사항}에 명시되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	검토방법의 상세 검토 항목 전반에 대해 {보안요구사항} 또는 {시스템요구사항}에 측정 가능하고 검증 가능한 수준으로 구체적 요건이 정의되어 있는 경우	현행 유지
보완권고 (△)	상세 검토 항목 중 일부(1~2개)가 누락되었거나 "모델 보안을 적용한다" 등 모호한 수준으로 기술되어 있어 보완이 필요한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	상세 검토 항목 중 3개 이상이 누락되었거나, 모델 저장소 보안에 대한 요구사항이 사실상 정의되지 않은 경우, 또는 접근통제·무결성 검증·배포 승인 중 핵심 요건이 모두 부재한 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 보안 요건의 구체화 유도: {보안요구사항}에 "모델 저장소에 접근통제를 적용한다" 수준의 모호한 문구가 있을 경우, "역할별 RBAC 정책 수립(최소 4개 역할 분리), 모든 접근 시 인증/인가 로그 기록" 등 측정 가능한 기준으로 구체화하도록 권고한다. 검사(시험) 단계에서 검증할 수 없는 요건은 사실상 요건이 아니다.

2) 무결성 검증의 자동화 수준 확인: 해시 비교나 서명 검증이 수동(사람이 매번 확인)인지 자동(CI/CD 파이프라인에 통합)인지를 반드시 구분하여 확인한다. 수동 검증은 휴먼 에러와 우회 가능성이 높으므로, 자동화 요건이 {보안요구사항}에 포함되어 있는지 점검한다.

3) 기획 단계 점검의 범위 인식: 본 가이드라인은 요구정의(기획/분석) 단계의 산출물 점검이므로, 실제 모델 저장소가 구축·운영되고 있는지를 검증하는 것이 아니라, 향후 구축을 위한 요구사항이 구체적이고 측정 가능하게 정의되었는지를 확인하는 것이다. "어떻게 구현할 것인가"보다 "무엇을 요구하고 있는가"에 초점을 맞추어야 한다.

4) 외부 모델 도입 시 공급망 위험 인식: Hugging Face, GitHub 등 공개 저장소에서 사전학습 모델을 도입하는 경우, 해당 모델의 출처·무결성·라이선스에 대한 검증 요건이 빠지기 쉽다. 특히 Pickle 형식(.pkl) 모델 파일은 역직렬화 시 임의 코드 실행이 가능하므로, SafeTensors 등 안전한 직렬화 포맷 사용 요건이 있는지도 확인한다.

5) 모델 공급망 보안(AI-BOM) 요건 확인: AI 모델의 공급망 보안(AI-BOM, SBOM) 요건이 정의되어 있는지 확인해야 한다. 승인되지 않거나 악성코드가 삽입된 모델 파일(Pickle 등)이 저장소에 업로드되는 것을 방지하기 위해, "업로드 시 자동화된 악성코드 스캔 및 해시 무결성 검증 절차 의무화" 요건이 {보안요구사항}에 포함되도록 유도한다.

AG_SD41-1-14 | 비정상 API호출 모델추출 공격요건 점검

시스템구조·보안 | 요구정의·분석 | 인공지능 모델 개발 | AI특화보안

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델이 외부에 API로 서비스되는 환경에서 발생할 수 있는 비정상 API 호출 및 모델 추출(Model Extraction) 공격에 대한 보안요구사항이 구체적이고 상세하게 정의되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템의 API 엔드포인트는 일반 웹 서비스와 달리, 대량의 추론(Inference) 요청을 통해 모델의 의사결정 경계(Decision Boundary)를 역추정하거나 모델 파라미터를 복제하는 공격에 노출될 수 있다.

이러한 공격은 기존 IT 보안 관점의 비인가 접근과는 성격이 다르며, 요구정의 단계에서부터 AI 특화 보안요구사항으로 명시되지 않으면 설계·구현 단계에서 대응 체계를 구축하기 어렵다.

따라서 본 가이드라인은 {보안요구사항}에 이러한 AI 고유 보안 위협에 대한 탐지 요건이 누락 없이 반영되었는지를 체계적으로 점검하기 위한 기준을 제공한다.

2. 필요성

▶ AI 서비스의 API 엔드포인트는 모델 추론 결과를 외부에 노출하는 인터페이스로서, 기존 정보시스템의 API 보안(인증·인가·트래픽 제어)과는 별도로 AI 고유의 보안 위협을 고려해야 한다. 요구정의 단계에서 이를 명확히 정의하지 않으면 개발 이후 보안 설계를 소급 적용하기 어렵고, 모델 자산(Intellectual Property)이 유출될 위험이 커진다.

▶ 모델 자산 보호: AI 모델은 대규모 투자(학습 데이터 확보, GPU 연산 비용 등)의 산출물이며, 모델 추출 공격을 통해 경쟁사 또는 악의적 행위자가 유사 모델을 복제할 수 있다. 요구정의 단계에서 추출 공격 탐지 요건을 보안요구사항으로 명시하지 않으면 이에 대한 기술적 대응이 설계에 반영되지 않는다.

▶ 비정상 API 호출 패턴 탐지 필요: 모델 추출 공격은 정상 사용자와 유사한 형태의 API 호출을 대량으로 수행하는 방식으로 이루어지므로, 일반적인 DDoS 방어나 Rate Limiting만으로는 탐지가 어렵다. 호출 빈도, 입력값 분포, 응답 패턴 등 AI 특화 이상징후 탐지 요건이 필요하다.

▶ 규제 및 기준 준수: NIST AI RMF는 AI 시스템의 보안·복원력 특성으로 적대적 사례(Adversarial Examples), 데이터 오염(Data Poisoning), 모델·학습데이터 유출(Exfiltration) 등을 주요 보안 우려사항으로 명시하고 있으며, 이러한 위협에 대한 요구사항 정의를 요구하고 있다.

▶ 공격 표면(Attack Surface) 사전 식별: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)는 AI 시스템의 보안 위협을 사전에 식별하고 대응 방안을 수립하도록 요구하며, API 엔드포인트를 통한 공격은 생성형 AI 서비스의 대표적 공격 표면에 해당한다.

3. 검토방법

No	점검 항목	확인 방법
1	API 호출 빈도 제한 요건 명시 여부	시간 단위·사용자 단위 등 구체적 기준과 함께 AI 모델 API 호출 빈도 제한(Rate Limiting) 요건이 {보안요구사항}에 명시되었는지 확인
2	비정상 대량 추론 요청 탐지 요건 포함 여부	동일 패턴 반복 요청·입력값 체계적 변조 등 단일 사용자·IP의 비정상 대량 추론 요청 탐지 요건이 {보안요구사항}·'AI 보안 위협 분석서'에 포함되었는지 확인
3	모델 내부 정보 노출 제한 요건 명시 여부	확률값·로짓·신뢰도 점수 등 API 응답의 모델 내부 정보 노출 범위를 제한하는 요건이 {보안요구사항}에 명시되었는지 확인
4	비정상 호출 탐지 시 대응 절차 반영 여부	차단·알림·로깅 등 비정상 API 호출 탐지 시 대응 절차가 {보안요구사항}·'사업수행계획서'에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	API 호출 빈도 제한, 비정상 대량 추론 요청 탐지, 응답 내 모델 내부 정보 노출 제한, 탐지 시 대응 절차가 {보안요구사항}에 구체적 기준과 함께 명시된 경우	현행 유지
보완권고 (△)	관련 요건이 일부 정의되었으나, 호출 제한 기준의 수치 미제시, 이상 패턴 탐지 요건 누락, 대응 절차 미정의 등 일부 항목이 미비하거나 선언적 수준에 그친 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	비정상 API 호출 및 모델 추출 공격에 대한 AI 특화 보안 요건이 {보안요구사항}에 전혀 정의되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 일반 API 보안과 AI 특화 API 보안의 구분: {보안요구사항}에 일반적인 API 인증·인가, 트래픽 제어(WAF, DDoS 방어) 요건만 포함되어 있고 AI 모델 특화 보안 요건(비정상 추론 패턴 탐지, 응답 정보 제한 등)이 누락된 경우가 빈번하다. 감리원은 'AI 모델 API' 보안과 '일반 API' 보안을 명확히 구분하여 점검해야 한다.

2) 모호한 보안 요건의 수치화 유도: {보안요구사항}에 'API 호출을 적절히 제한한다'와 같은 모호한 문구가 있는 경우, '시간당 사용자별 API 호출 100회 이내', '단일 IP 일일 누적 추론 요청 1,000건 초과 시 차단' 등 측정 가능한 기준으로 구체화하도록 보완을 권고해야 한다.

3) 모델 추출 공격의 위험 수준 판단: 모든 AI 모델이 동일한 수준의 추출 공격 위험을 갖지는 않는다. 모델이 외부에 공개 API로 서비스되는 경우, 모델이 고가의 투자로 개발된 핵심 자산인 경우, 모델 출력에 상세한 확률 정보가 포함되는 경우에 위험 수준이 높아진다. 감리원은 해당 AI 서비스의 특성을 파악한 뒤, 위험 수준에 비례하는 보안 요건이 정의되었는지 판단해야 한다.

4) 과도한 기술적 요구 지양: 감리원은 보안요구사항의 '존재 여부와 구체성'을 점검하는 것이지, 특정 기술 솔루션(예: 특정 WAF 제품, 특정 워터마킹 알고리즘)의 적용을 요구해서는 안 된다. 요구사항에 '무엇을' 방어해야 하는지가 명확히 기술되었는지를 점검하고, '어떻게' 구현할지는 사업자의 기술적 판단에 맡기는 것이 적절하다.

5) 개발 단계 산출물과의 연계: 요구정의 단계에서 AI 보안 요구사항이 적절히 정의되었더라도, 이를 설계 단계에서 구현하기 위한 'AI 보안 설계서(초안)' 또는 'API 보안 정책서' 작성 계획이 사업수행계획서(WBS)에 반영되어 있는지를 함께 확인해야 한다. 요구사항만 정의되고 후속 산출물로 연결되지 않으면 사문화될 위험이 있다.

AG_SD41-1-15 | AI 공급망 보안(AI-BOM) 요건 점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | AI특화보안

근거 | EU AI Act · NIST AI RMF GOVERN·MAP · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「저작권법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템을 구성하는 모든 소프트웨어·데이터·모델 구성요소에 대한 AI 소프트웨어 자재명세서(AI Bill of Materials, AI-BOM) 요건이 {요구사항정의서} 및 {보안요구사항}에 명시되고, 외부 파운데이션 모델·오픈소스 라이브러리·써드파티 데이터셋의 공급망 보안 위험이 식별·관리될 수 있도록 계획되어 있는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 'AI-BOM 초안', 데이터 카드, SBOM 도구 출력, '공급망 위험 평가서', 데이터 매니페스트를 근거로 구성요소 식별의 완전성, 라이선스·취약점 관리, AI-BOM의 운영 갱신과 공급망 사고 대응 절차를 확인한다.

2. 필요성

▶ AI 시스템은 외부 파운데이션 모델, 오픈소스 라이브러리, 써드파티 데이터셋에 광범위하게 의존하므로 공급망 보안 위험이 일반 정보시스템보다 크다. 구성요소를 식별하지 못하면 취약점 공지가 나와도 영향 범위를 판단할 수 없다.

▶ EU AI Act는 고위험 AI에 대해 구성요소·데이터·모델의 추적성 확보를 요구하며, NIST AI RMF의 GOVERN·MAP 기능은 제3자 공급망 의존성의 식별과 문서화를 요구한다.

▶ 오픈소스 LLM(Llama·Mistral 등)의 라이선스 조건 위반은 법적 분쟁의 원인이 된다. 상업적 이용·재배포·파인튜닝 결과물의 배포 조건은 모델별로 상이하므로 개별 확인이 필요하다.

▶ 모델 가중치와 학습 데이터는 배포 이후에도 갱신되므로 AI-BOM은 일회성 산출물이 아니라 운영 중 갱신되는 관리 대상이어야 한다. 갱신 주기와 책임자가 정의되지 않으면 즉시 노후화된다.

▶ 공급망 취약점이나 모델 변조 사고 발생 시 영향 범위를 신속히 판단하려면 구성요소의 버전과 해시가 기록되어야 한다. 사고 대응 절차가 AI-BOM과 연계되지 않으면 기록의 실효성이 없다.

3. 검토방법

No	점검 항목	확인 방법
1	파운데이션 모델명·버전·제공처	'AI-BOM 초안'에서 사용 모델의 명칭·버전·제공처와 라이선스가 항목으로 기재되었는지 확인
2	학습 데이터 출처·라이선스·개인정보 포함 여부	'AI-BOM 초안'과 데이터 카드에서 학습 데이터의 출처·라이선스·개인정보 포함 여부가 기재되었는지 확인
3	임베딩 모델·토크나이저 구성요소	시스템 구성도와 설정 파일에서 임베딩 모델·토크나이저의 버전이 식별되고 AI-BOM에 반영되었는지 확인
4	오픈소스 AI 라이브러리 버전·CVE 관리	SBOM 도구 출력에서 transformers·langchain 등 라이브러리의 버전과 알려진 취약점(CVE) 대응 상태를 확인
5	써드파티 API 의존성 및 위험 평가	'공급망 위험 평가서'에서 검색·번역 등 외부 API 의존성이 식별되고 위험 등급과 대응 방안이 기재되었는지 확인
6	파인튜닝·RAG 인덱스 데이터의 출처·해시	데이터 매니페스트에서 파인튜닝 데이터와 RAG 인덱스 원본의 출처와 해시가 기록되어 변조 여부를 검증할 수 있는지 확인
7	AI-BOM 운영 갱신 및 공급망 사고 대응	{요구사항정의서}와 {보안요구사항}에서 AI-BOM의 갱신 주기·책임자와, 공급망 취약점·모델 변조 사고 발생 시 대응 절차가 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI-BOM 명세·위험 평가·운영 갱신·사고 대응이 {요구사항정의서}와 {보안요구사항}에 구체적으로 반영되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	AI-BOM 일부 항목은 정의되었으나 라이선스·갱신·사고 대응이 일부 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI-BOM 개념·요건이 미반영되어 일반 SBOM 수준에 그치거나 단순 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 특화 항목의 명세 요구: AI-BOM은 SBOM 도구만으로는 생성되지 않으므로, AI 특화 항목(모델 가중치·학습 데이터·임베딩)을 수작업으로라도 명세하도록 요구사항에 명시해야 한다.

2) '오픈소스'와 상업적 사용의 구분: 오픈소스 LLM의 '오픈'이 '상업적 사용 자유'를 의미한다고 오해하지 않도록, 사용 모델별 라이선스 조건과 써드파티 데이터셋의 라이선스(CC-BY-NC 등)를 반드시 확인한다.

3) 표준 포맷·도구 조합 활용: CycloneDX의 AI/ML 확장(v1.5+)을 표준 포맷으로 권고하며, AI-BOM 자동화는 Trivy·Anchore와 자체 모델 메타데이터 스크립트 조합을 추천한다. 라이선스 충돌은 Black Duck·FOSSA 등 SCA 도구를 활용한다.

4) 인수자료 확보: 공급자의 SOC2 보고서·ISO27001 인증서 사본을 사업 종료 시 인수자료로 요구한다.

5) 갱신 지속성 점검: AI-BOM이 사업 종료 시 1회 작성으로 종결되지 않도록 갱신 주기(월 1회 이상 또는 구성요소 변경 시 즉시)와 CVE 스캔 도구·주기·담당자를 확인한다.

AG_SD41-1-16 | 데이터 최적화 방어요건 및 적대적 데이터 최적화 공격방어 요건

시스템구조·보안 | 요구정의·분석 | 데이터 수집·처리 | AI특화보안

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「개인정보 보호법」 · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 모델의 학습 데이터·추론 입력 데이터를 대상으로 한 적대적 예제 공격(Adversarial Example Attack)·데이터 중독 공격(Data Poisoning Attack)·모델 추출 공격(Model Extraction Attack)·멤버십 추론 공격(Membership Inference Attack) 등 적대적 데이터 최적화 공격 유형별 방어 요건이 {요구사항정의서}에 정의되었는지를 감리원의 제3자적 관점에서 점검하는 것을 목적으로 한다.

이를 통해 공격 방어 요건 미정의로 인해 발생할 수 있는, AI 서비스 출시 이후 적대적 공격에 의한 AI 모델 오동작·개인정보 유출·서비스 신뢰성 훼손 위험을 사전에 차단한다.

2. 필요성

▶ AI 모델은 전통적인 소프트웨어 보안 취약점과 달리 데이터 최적화 기반의 고유한 적대적 공격에 노출된다. 사람이 감지할 수 없는 미세한 데이터 조작으로 AI 모델의 판단을 완전히 뒤바꾸거나, 학습 데이터에 백도어를 삽입하여 특정 트리거 입력 시 의도된 오작동을 유발하거나, 모델 반복 조회를 통해 학습 데이터의 개인정보를 역추출하는 공격이 가능하다.

▶ 적대적 예제 공격으로 인한 AI 모델 판단 오류 및 안전 사고 방지: 이미지·텍스트·음성에 사람이 인지할 수 없는 미세한 변형을 가하여 AI 모델이 완전히 잘못된 판단을 내리도록 유도하는 적대적 예제 공격은 의료 영상 AI(악성 종양을 정상으로 오분류)·자율주행 AI(정지 신호 무시)·금융 사기 탐지 AI(사기 거래를 정상 거래로 오분류) 등에서 치명적 안전 사고로 이어질 수 있다.

▶ 데이터 중독 공격으로 인한 AI 모델 백도어 삽입 및 의도적 오작동 방지: 학습 데이터 수집·라벨링·전처리 과정에 공격자가 개입하여 특정 트리거 패턴(예: 이미지 내 특정 스티커·텍스트 내 특정 문구)이 포함되면 의도된 오분류를 유발하는 백도어 데이터를 삽입하는 데이터 중독 공격은, {요구사항정의서}에 방어 요건이 정의되지 않으면 모델 학습 완료 후에는 탐지가 매우 어렵고 방어 비용이 급증한다.

▶ 모델 추출 공격으로 인한 AI 모델 지식재산권 침해 및 기밀 데이터 유출 방지: 공격자가 API를 통해 AI 모델에 대량의 쿼리를 보내 모델의 입출력 패턴을 분석함으로써 상업적 가치가 있는 AI 모델의 구조·가중치를 복제하는 모델 추출 공격은 지식재산권 침해와 함께 학습 데이터 간접 노출 위험을 유발한다.

▶ 멤버십 추론 공격으로 인한 AI 학습 데이터 개인정보 역노출 방지: 특정 데이터가 AI 모델의 학습 데이터에 포함되었는지 여부를 모델의 출력 확률 분포 분석으로 판별하는 멤버십 추론 공격은, 의료·금융·공공 AI 서비스에서 특정 개인의 민감 정보(병력·신용등급 등)가 학습 데이터에 포함되었다는 사실 자체가 개인정보 침해가 될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 서비스 위험 등급 기반 적대적 공격 유형 식별 및 방어 요건 적용 범위 정의	AI 서비스의 위험 등급(고위험: 의료·안전·금융 / 중위험: 공공 서비스 / 저위험: 일반 정보 서비스)에 따라 방어가 필요한 적대적 공격 유형(적대적 예제·데이터 중독·모델 추출·멤버십 추론)과 방어 요건 적용 범위가 {요구사항정의서}에 정의되었는지 확인
2	학습 데이터 수집·라벨링 단계의 데이터 중독 공격 방어 및 무결성 검증 요건 정의	학습 데이터 오염 데이터 유입 차단(출처 신뢰도 검증·이상 데이터 탐지), 악의적 라벨 조작 방지(라벨러 접근 통제·교차 검수·IAA 기반 이상 라벨러 탐지) 등 데이터 무결성 검증 요건이 정의되었는지 확인
3	적대적 예제 공격 내성 요건 및 입력 데이터 검증·정제 요건 정의	운영 중 AI 모델에 입력되는 데이터에 대한 적대적 예제 공격 내성 목표 수치(예: PGD 공격 기준 정확도 70% 이상 유지), 입력 데이터 통계 분포 이상 탐지 요건, 의심 입력 차단·정제 요건이 정의되었는지 확인
4	모델 추출 공격 방어 요건: API 쿼리 제한·출력 결과 모호화 요건 정의	외부 API를 통한 AI 모델 쿼리에 대한 일정 시간 내 최대 쿼리 수 제한(Rate Limiting), 출력 확률값 반올림·범위 제한을 통한 출력 결과 모호화(Output Perturbation) 요건이 정의되었는지 확인
5	임베딩·벡터 대상 적대적 공격 방어 요건	벡터 인덱스 중독·임베딩 역추적·검색결과 변조 등 벡터 DB·RAG 대상 공격 유형별 방어 요건(OWASP LLM08 포함)이 정의되었는지 확인
6	멤버십 추론 공격 방어 요건: 차등 프라이버시·출력 정규화 요건 정의	AI 모델 출력 확률 분포 분석을 통한 학습 데이터 역추출 위험에 대한 차등 프라이버시(Differential Privacy) 적용 여부, L2 정규화를 통한 과적합 억제, 출력 확률 제한 요건이 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 서비스 위험 등급 기반 방어 대상 공격 유형 정의, 데이터 중독 방어·라벨러 접근 통제·무결성 검증 요건, 적대적 예제 내성 목표 수치·입력 탐지 요건, 모델 추출·멤버십 추론 방어 요건, 임베딩·벡터 대상 공격(벡터 인덱스 중독·임베딩 역추적·검색결과 변조) 방어 요건, 공격 탐지 기준 및 즉시 대응 절차가 {요구사항정의서}에 모두 정의됨	현행 유지
보완권고 (△)	일부 공격 유형 방어 요건은 정의되었으나 나머지 공격 유형 방어 요건 누락, 방어 수준 기준 수치 미정의, 공격 탐지 기준 미정의, 탐지 시 대응 절차 미정의 등 일부 미흡	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 적대적 공격 방어 요건이 {요구사항정의서}에 전혀 없거나 '보안을 강화하겠다'는 선언만 있고 AI 고유 적대적 공격 유형 식별·방어 요건·탐지 기준이 전무한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 적대적 공격 방어는 일반 IT 보안과 별개로 점검해야 한다: 방화벽·침입 탐지·암호화 등 일반 IT 보안 요건을 충족해도 AI 모델의 적대적 공격(적대적 예제·데이터 중독·모델 추출·멤버십 추론)은 완전히 별개의 위협 벡터로 일반 보안 체계로는 방어되지 않는다. 감리원은 {요구사항정의서}에 일반 IT 보안 요건 외에 AI 고유의 적대적 공격 유형별 방어 요건이 별도로 정의되었는지 반드시 확인해야 한다.

2) 서비스 위험 등급에 따라 방어 요건 강도를 차등 적용한다: 의료·안전·금융 AI 서비스(고위험)는 4대 공격 유형 전체에 대한 방어 요건이 수치 기준과 함께 정의되어야 한다. 반면 일반 정보 제공 AI 서비스(저위험)는 데이터 중독 방어와 기본적인 입력 검증 요건만으로 충분할 수 있다. 감리원은 서비스 위험 등급과 방어 요건의 강도가 비례하는지 확인하는 것이 핵심이다.

3) 데이터 중독 공격은 학습 데이터 구축 단계에서만 방어 가능: 데이터 중독 공격(백도어 삽입)은 AI 모델 학습이 완료되면 탐지가 매우 어렵고 방어 비용이 급증한다. 요구정의 단계에서 학습 데이터 수집·라벨링·전처리 단계별 중독 방어 요건이 {요구사항정의서}에 정의되어야만 데이터 구축 계획에 방어 절차가 반영될 수 있다. 감리원은 이 점에서 요구정의 단계 점검의 중요성을 강조해야 한다.

4) 외부 API 공개 AI 서비스는 모델 추출 공격 방어 요건 필수 확인: 외부에 REST API를 통해 AI 모델을 공개하는 서비스는 모델 추출 공격에 직접 노출되는 구조이므로 Rate Limiting·출력 결과 모호화 요건이 {요구사항정의서}에 반드시 포함되어야 한다. '내부망 서비스이므로 안전하다'는 주장도 내부 위협 경로(악의적 내부자·APT 공격 이후 내부망 침투)를 간과한 것이므로 수용하지 않는다.

5) 멤버십 추론 공격 방어는 개인정보 보호법 준수와 직결: 개인 민감 데이터(의료 기록·금융 거래·생체 정보)를 학습에 활용하는 AI 서비스에서 멤버십 추론 공격이 성공하면 '특정 인물의 민감 정보가 AI 학습 데이터에 포함됨'이 노출되어 개인정보 보호법 위반이 될 수 있다. 감리원은 개인 민감 데이터 기반 AI 서비스에서 멤버십 추론 공격 방어 요건(차등 프라이버시·출력 제한)이 {요구사항정의서}에 반드시 정의되었는지 확인해야 한다.

AG_SD41-1-17 | API 접근통제 인증요건 점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 인증/인가/접근통제

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · OWASP LLM Top 10 · 「국가·공공기관 AI 보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템의 추론(Inference) API에 대한 접근 통제 및 인증 체계가 {보안요구사항}에 구체적으로 정의되었는지, 비인가자의 추론 API 접근을 차단하기 위한 보안 요건이 누락 없이 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 추론 API는 학습된 모델에 외부 요청을 전달하고 결과를 반환하는 핵심 인터페이스로서, 기존 웹 API와 달리 모델 추출(Model Extraction) 공격, 프롬프트 인젝션(Prompt Injection), 과다 호출을 통한 서비스 거부(DoS) 등 AI 고유의 보안 위협에 직접 노출된다.

따라서 이러한 AI 특화 위협에 대응하는 접근 통제·인증 요건이 보안요구사항에 명확히 반영되었는지를 확인하는 것이 본 가이드라인의 핵심이다.

2. 필요성

▶ 생성형 AI 시스템의 추론 API는 일반적인 REST API와 달리 고가의 GPU 연산 자원을 직접 소비하며, 응답 결과에 학습 데이터의 패턴이 내포되어 있어 비인가 접근 시 모델 지식 탈취, 개인정보 추론, 서비스 악용 등 복합적 위험이 발생한다. 요구정의 단계에서 이를 체계적으로 정의하지 않으면 설계·구현 단계에서 보안 통제가 누락될 수 있다.

▶ AI 모델 자산 보호: 추론 API는 모델의 입출력 패턴을 외부에 노출하는 경로이므로, 비인가자가 반복 호출을 통해 모델을 복제(Model Extraction)하거나 학습 데이터를 추론(Membership Inference)할 수 있다. 이를 차단하기 위한 인증·접근 통제 요건을 정의해야 한다.

▶ 과다 호출 방지 및 자원 보호: GPU 기반 추론은 호출당 비용이 높으므로, Rate Limiting·Quota 관리 등 호출량 통제 요건이 요구사항에 미리 반영되지 않으면 비인가 대량 호출로 인한 서비스 장애와 과도한 비용 발생 위험이 있다.

▶ 프롬프트 인젝션 대응 요건 조기 확보: 생성형 AI 특유의 공격 벡터인 프롬프트 인젝션에 대비한 입력 검증·필터링 요건을 명시해야 설계 시 구현 방향이 확정된다.

▶ 내외부 연계 보안: AI 추론 API가 외부 서비스(SaaS LLM 등)와 연계되는 경우, API 키 관리·전송 구간 암호화·응답 데이터 필터링 등의 보안 요건이 요구정의 단계에서 누락되면 설계 이후 보완이 어렵다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 API 인증 방식 명시 여부	OAuth 2.0·API Key·JWT 등 AI 추론 API 호출 시 적용할 인증 방식이 {보안요구사항}에 명시되었는지 확인
2	인증 토큰 수명주기 정책 포함 여부	인증 토큰의 유효기간·갱신 정책·폐기 절차가 {보안요구사항}에 포함되었는지 확인
3	호출 주체별 인증 수준 차등 정의 여부	사용자·시스템·관리자 등 API 호출 주체별 인증 수준 차등 적용 요건이 {보안요구사항}·{요구사항정의서}에 정의되었는지 확인
4	외부 LLM API 키 관리 보안 요건 포함 여부	외부 LLM API 연계 시 API 키 관리(저장·전송·교체) 보안 요건이 {보안요구사항}에 포함되었는지 확인하고 사업자 면담으로 보완

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	추론 API의 인증 방식이 명시적으로 정의되고, 토큰 수명주기(유효기간·갱신·폐기) 관리 정책, 호출 주체별 인증 수준 차등, 외부 LLM API 키 관리 요건까지 {보안요구사항}에 구체적으로 반영된 경우	현행 유지
보완권고 (△)	인증 방식은 명시되었으나 토큰 수명주기 관리 정책, 호출 주체별 차등 인증, API 키 관리 절차 등 일부 요건이 누락되었거나 모호한 선언 수준으로 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	추론 API에 대한 접근 통제·인증 요건이 {보안요구사항}에 정의되지 않았거나, 일반 웹 API 보안 요건만 존재하여 비인가 추론 API 접근 차단이 불가능한 경우	부적합 사항 시정조치 요구 및 이행 결과 확인 (암호화·접근제어 요건 미적용 발견 시 즉시 보완 요구)

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 보안 요건의 수치화: {보안요구사항}에 'API 접근을 통제한다'와 같은 모호한 문구가 있으면, 'OAuth 2.0 기반 인증 적용, 토큰 유효기간 30분, 사용자당 분당 최대 호출 60회' 등 측정 가능한 기준으로 구체화하도록 개선을 유도한다. 특히 OAuth 2.0, JWT, API Key 등 적용할 인증 방식과 그 조합이 명시적으로 선정되었는지 확인한다.

2) AI API의 서비스 간 인증(Service-to-Service Authentication) 요건 분리 확인: AI API의 특성상 내부 시스템 간의 통신(예: Backend와 LLM 간)에 대한 '서비스 간 인증(Service-to-Service Authentication)' 요건이 일반 사용자 인증과 분리되어 정의되었는지 점검할 것을 권고한다.

3) 일반 API 보안과 AI API 보안의 구분: 일반적인 웹 API 보안(HTTPS, 인증 등)만 기술된 경우, AI 추론 API 고유의 위협(모델 추출, 프롬프트 인젝션, 과다 호출에 의한 GPU 자원 고갈 등)에 대한 보안 요건이 추가로 필요한지 확인한다. 기존 보안 요건만으로는 AI 특화 위협에 대한 대응이 불충분할 수 있다.

4) 외부 LLM API 연계 시 특별 주의: 사업자가 OpenAI, Google 등 외부 LLM API를 연계 사용하는 경우, API 키가 소스코드·환경변수·로그 등에 하드코딩되지 않도록 관리 요건이 정의되었는지 확인한다. 외부 API 호출 시 사내 민감정보가 프롬프트에 포함되어 전송되는 위험도 함께 점검한다.

5) 보안 아키텍처 설계서(초안) 점검: 'API Gateway + 인증 서버 + 추론 서버'의 보안 구성요소 배치가 개념 수준으로라도 설계되어 있는지 확인한다. 이 산출물이 없으면 보안 요구사항이 설계 단계에서 구현 가능한지 판단하기 어렵다.

AG_SD41-1-18 | 암호화 전송 요건

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 암호화

근거 | 「개인정보 보호법」 · CSAP · KCMVP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템의 데이터 및 모델 파라미터가 내·외부 통신구간을 통해 전송될 때 적용되는 암호화 요건이 {보안요구사항}에 구체적으로 정의되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 AI 시스템은 학습 데이터, 추론 결과, 프롬프트·응답 페이로드, 모델 가중치 등 일반 정보시스템보다 다양하고 민감한 데이터가 API 호출 및 내부 마이크로서비스 간 통신을 통해 빈번하게 전송된다.

이 과정에서 TLS(Transport Layer Security) 적용 범위와 버전, 파라미터(프롬프트·응답 본문) 암호화 방식이 요구사항으로 명시되지 않으면, 설계·구현 단계에서 보안 공백이 발생할 위험이 높다.

본 가이드라인은 이러한 AI 특유의 암호화 전송 요건이 요구정의 산출물에 빠짐없이 반영되었는지 확인하기 위한 점검 방향과 기준을 제시한다.

2. 필요성

▶ 생성형 AI 시스템은 일반 정보시스템과 달리, 프롬프트에 포함된 사내 기밀정보·개인정보가 외부 LLM API로 전송되거나, 학습 데이터가 내부 네트워크를 경유하여 GPU 클러스터로 이동하는 등 민감 데이터의 전송 경로가 복잡하고 다양하다. 이에 따라 통신구간 암호화 및 파라미터 암호화 요건을 구체적으로 정의해야 한다.

▶ 프롬프트·응답 데이터 보호: 사용자가 입력하는 프롬프트에 민감 정보(개인정보, 사내 기밀 등)가 포함될 수 있으며, AI 응답에도 비식별화 이전의 원본 데이터가 노출될 수 있다. 통신구간 암호화(TLS) 및 페이로드 수준의 파라미터 암호화가 적용되지 않으면 중간자(Man-in-the-Middle) 공격에 의한 정보 유출 위험에 노출된다.

▶ 모델 가중치 및 학습 데이터 유출 방지: AI 모델의 가중치(weights)와 학습 데이터셋은 조직의 핵심 지적자산에 해당하며, 전송 과정에서 비암호화 채널을 통해 유출될 경우 모델 탈취(Model Extraction) 및 학습 데이터 역추론(Membership Inference) 공격에 노출된다.

▶ 외부 API 연계 구간 보안: 서드파티 LLM API(예: OpenAI, Anthropic 등) 호출 시 API 키, 프롬프트 내용, 응답 데이터가 공용 인터넷을 경유하므로, 이 구간의 TLS 버전·인증서 검증·API 키 암호화 보관 요건이 요구사항에 명시되지 않으면 설계 단계에서 보안 허점이 고착된다.

▶ 법적·규정 준수: 개인정보 보호법, 전자금융거래법, 클라우드 보안 인증제(CSAP) 등 국내 법규는 개인정보의 암호화 전송을 의무화하고 있으며, AI 시스템도 이러한 규정의 적용 대상에 해당한다. 법적 암호화 요건을 명확히 반영해야 후속 단계에서의 준수 여부를 점검할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	통신구간 식별·구간별 TLS 요건 명시 여부	사용자↔프론트엔드·프론트엔드↔AI 추론서버·추론서버↔외부 LLM API·학습데이터 파이프라인 등 주요 통신구간이 식별되고 구간별 TLS 적용 요건이 {보안요구사항}에 명시되었는지 확인
2	TLS 버전·암호 스위트 기준 포함 여부	TLS 프로토콜 버전(1.2 이상)과 허용 암호화 스위트(Cipher Suite) 기준이 {보안요구사항}·'AI 보안 아키텍처 설계서(초안)'에 포함되었는지 대조
3	인증서 관리 방안 포함 여부	발급 주체·갱신 주기·인증서 고정(Certificate Pinning) 적용 여부 등 인증서 관리 방안이 {보안요구사항}에 포함되었는지 확인하고 사업자 면담으로 보완
4	내부 마이크로서비스 mTLS 요건 명시 여부	AI 시스템 내부 마이크로서비스 간 통신(East-West 트래픽)에 대한 mTLS(상호 TLS) 적용 요건이 {보안요구사항}에 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	주요 통신구간이 모두 식별되고, 구간별 TLS 적용 범위·버전·암호화 스위트, 인증서 관리 방안, 내부 mTLS 적용 요건이 {보안요구사항}에 구체적으로 명시된 경우	현행 유지
보완권고 (△)	암호화 전송 요건은 정의되어 있으나 일부 통신구간 누락, TLS 버전·암호화 스위트 미특정, 인증서 관리 방안 미비 등 요건이 부분적으로만 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안요구사항}에 암호화 전송 요건이 누락되었거나 "암호화를 적용한다" 등 선언적 문구만 있어 적용 구간·기준을 확인할 수 없는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선언적 문구 경계: {보안요구사항}에 '암호화를 적용한다', '보안을 강화한다' 등 선언적 문구만 기술된 경우, 설계·구현 단계에서 해석이 분분해져 보안 수준이 사업자 재량에 맡겨지는 결과를 초래한다. 감리원은 반드시 구체적인 TLS 버전, 암호화 알고리즘, 키 길이, 적용 구간이 수치·명칭으로 특정되었는지 확인해야 한다.

2) AI 특유 통신구간에 집중: 일반 IT 시스템에서도 TLS 적용은 기본 요건이므로, 감리원은 일반적 TLS 요건보다 AI 시스템 고유의 통신구간(추론 서버↔외부 LLM API, 학습 데이터 파이프라인, 모델 가중치 전송 등)에 대한 암호화 요건이 별도로 식별·명시되었는지에 중점을 두어야 한다.

3) 외부 LLM API 호출 시 정보 유출 위험 강조: 사용자 프롬프트가 외부 LLM API로 전송되는 구조에서는, 프롬프트에 포함된 사내 기밀·개인정보가 서드파티 클라우드에 저장·학습될 가능성이 있다. 이에 대한 요구사항(프롬프트 비식별화, 외부 전송 차단 정책, 서드파티 데이터 처리 방침 확인 등)이 {보안요구사항}에 포함되었는지 반드시 점검해야 한다.

4) 국내 암호화 기준 적용 확인: 공공기관 AI 사업의 경우 국가정보원 암호모듈 검증(KCMVP) 기준이 적용될 수 있으므로, {보안요구사항}에 KCMVP 인증 모듈 사용 요건이 포함되었는지 확인한다. 민간 사업의 경우에도 개인정보 보호법에 따른 암호화 기준(예: KISA 암호 알고리즘 기준)이 준수되었는지 확인이 필요하다.

5) 운영 단계 산출물 계획 확인: 개발 과정에서 만들어야 하는 운영 관련 산출물(예: 'AI 시스템 암호화 운영 절차서', '인증서 관리 대장', 'API 키 관리 대장')의 작성 계획이 '사업수행계획서'에 포함되었는지 확인한다. 이러한 운영 산출물 계획이 요구정의 단계에서 수립되어야 설계·구현 단계에서 누락 없이 작성된다.

AG_SD41-1-19 | 오픈소스 라이브러리 보안 요건

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 시큐어코딩

근거 | NIST AI RMF GOVERN·MAP · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

AI 서비스 구축 사업에서, AI 시스템 개발에 사용되는 오픈소스 라이브러리에 대한 보안 요건과 의존성 취약점(CVE) 점검 도구 도입 요건이 {보안요구사항}에 구체적·측정 가능한 형태로 반영되었는지를 확인하는 것이 목적이다.

요구정의 단계에서 오픈소스 보안 요건이 누락되거나 모호하게 기술될 경우, 개발·테스트 단계에서 취약 라이브러리가 검증 없이 사용되고, 최종 서비스 배포 후 알려진 보안 취약점(CVE)을 통한 공격에 노출될 위험이 크다.

이 가이드라인은 그러한 위험의 원인을 분석 단계에서 선제적으로 식별하고 보완 요구 근거를 확보하는 데 초점을 둔다.

2. 필요성

▶ AI 시스템은 LLM 프레임워크(LangChain, Hugging Face 등), 벡터 데이터베이스, 데이터 전처리 라이브러리 등 다수의 외부 오픈소스 패키지를 복합적으로 사용한다. 이러한 패키지들은 빠른 버전 주기와 함께 신규 CVE가 지속적으로 등록되며, 특정 라이브러리 하나의 취약점이 전체 AI 파이프라인의 보안을 위협할 수 있다.

▶ 보안 요건 미정의: 발주기관·사업자 간 오픈소스 허용 범위와 취약점 허용 기준을 사전에 합의하지 않으면, 검사 단계에서 기준 부재로 인한 분쟁이 발생한다.

▶ CVE 점검 자동화 요건 부재: 수동 점검만으로는 수백 개의 패키지 의존성을 체계적으로 관리하기 어렵다. 점검 도구 도입 요건이 정의되어야 도구 선정·예산 편성이 적시에 이루어진다.

▶ AI 특유의 공급망 위험: 사전훈련 모델(Pre-trained Model) 및 전이학습 컴포넌트 사용은 전통적 소프트웨어와 다른 공급망 취약점을 내포한다(NIST AI RMF MAP 4.1, GOVERN 6.1 참조).

▶ 법·제도적 근거: 국가정보원 소프트웨어 보안 개발 가이드라인, 과기정통부 소프트웨어 공급망 보안 가이드라인은 사용 오픈소스 목록 관리 및 취약점 점검 절차 수립을 명시하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	허용 라이선스 및 출처 제한	허용 오픈소스 라이선스 유형(MIT, Apache 2.0)과 허용 불가 라이선스(GPL 등) 목록 정의 여부 확인

No	점검 항목	확인 방법
2	CVE DB 연동 요건	NVD·OSV 등 CVE 데이터베이스 연동 요건 및 DB 업데이트 주기 명시 여부 확인
3	점검 자동화 및 CI/CD 연계 요건	CI/CD 파이프라인 자동 취약점 점검 요건·Break Build 기준 포함 여부 확인
4	오픈소스 사용 목록(안) 작성 여부	도입 예정 라이브러리 목록(명칭·버전·라이선스·사용 목적) 별도 산출물 존재 여부 확인
5	점검 결과 보고 및 관리 요건	취약점 점검 결과 보고서 산출 주기·형식(SBOM 포함)·조치 기한 및 책임자 지정 여부 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	허용·불허 오픈소스 라이선스 유형과 출처 제한이 정의되고, CVE 데이터베이스 연동 요건과 DB 갱신 주기, CI/CD 자동 취약점 점검 및 Break Build 기준이 명시되며, 도입 예정 라이브러리 목록(명칭·버전·라이선스·사용 목적)이 별도 산출물로 작성되고, 점검 결과 보고서의 산출 주기·형식(SBOM 포함)과 조치 기한·책임자가 지정된 경우	현행 유지
보완권고 (△)	{보안요구사항}에 오픈소스 요건은 있으나, CVSS 기반 허용 기준 모호, CI/CD 연계 미명시 등 도구 요건 일부가 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안요구사항}에 오픈소스·CVE 관련 요건이 전혀 없거나, "보안 취약점 없어야 함" 등 측정 불가 표현만 있는 경우 (버전 없는 라이브러리 목록 포함)	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 요건의 수치화를 유도한다. "보안 취약점이 없는 라이브러리를 사용한다"와 같은 모호한 표현은 향후 분쟁의 원인이 되므로, "CVSS 7.0 이상 취약점 포함 라이브러리 사용 금지", "분기 1회 이상 CVE 스캔 실시" 등 측정 가능한 기준으로 구체화하도록 유도한다.

2) AI 프레임워크의 간접 의존성(Transitive Dependency)에 주의한다. LangChain, Hugging Face Transformers 등 AI 프레임워크는 수십~수백 개의 간접 의존성 패키지를 포함하므로, '오픈소스 사용 목록(안)'이 직접 선언한 패키지만 열거된 경우 간접 의존성 관리 계획 포함 여부를 추가로 확인한다.

3) 점검 도구 요건과 기술 스택의 정합성을 확인한다. 도입 예정 점검 도구가 사업의 기술 스택(Python/pip, Node.js/npm, 컨테이너 이미지 등)과 CVE 데이터베이스(NVD·OSV) 연동을 실제로 지원하는지 확인하고, 미지원 시 도구 재선정을 권고한다.

AG_SD41-1-20 | 망분리 환경 모델 훈련 보안요건 점검

시스템구조·보안 | 요구정의·분석 | 시스템 구현 | 망분리/폐쇄망

근거 | 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · 「개인정보 보호법」 · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 망분리 환경 내 AI 모델 훈련(학습)에 필요한 보안 요건이 {보안요구사항}에 구체적으로 반영되었는지를 감리원이 검토하기 위한 기준을 제공하는 데 목적이 있다.

생성형 AI 모델의 훈련 과정은 대량의 내부 데이터(업무 문서, 고객 정보, 기밀 자료 등)를 학습 데이터로 활용하므로, 외부 인터넷과 물리적·논리적으로 격리된 폐쇄망(내부망) 환경에서의 학습이 요구되는 경우가 많다.

이 과정에서 학습 데이터의 외부 유출, 모델 파라미터의 비인가 접근, 훈련 파이프라인을 통한 데이터 오염(Data Poisoning) 등의 보안 위협이 존재하며, 이러한 위협이 요구정의 단계에서 보안 요구사항으로 식별·정의되지 않으면 설계·개발 단계에서 보안 통제가 누락되는 결과를 초래할 수 있다.

따라서 본 가이드라인은 '망분리 환경에서의 모델 훈련'이라는 AI 고유의 보안 위험에 초점을 맞추어, 감리원이 요구정의 산출물의 보안요구사항을 체계적으로 점검할 수 있도록 검토 기준을 제시한다.

2. 필요성

▶ 생성형 AI 시스템의 모델 훈련은 일반 정보시스템 개발과 근본적으로 다른 보안 위협 구조를 가진다. 대량의 원본 데이터가 학습 파이프라인에 직접 투입되고, 훈련된 모델 가중치(Weight) 자체가 학습 데이터의 정보를 내재하고 있어 모델 유출이 곧 데이터 유출로 이어질 수 있다. 특히 공공기관이나 금융기관과 같이 망분리 규제가 적용되는 환경에서는 훈련 환경 전반의 보안 요건을 명확히 정의하여야 한다.

▶ 학습 데이터 유출 방지: 내부망에 존재하는 고객 정보, 영업비밀, 비공개 업무 문서 등이 모델 학습 과정에서 외부로 유출될 경우 개인정보 보호법 위반 및 기관 신뢰도 실추로 직결된다. 요구정의 단계에서 학습 데이터의 반입·반출 통제 요건이 명시되지 않으면 이후 단계에서 보안 공백이 발생한다.

▶ 모델 가중치 보호: 훈련이 완료된 모델 파일(가중치)은 학습에 사용된 데이터의 통계적 특성을 내포하고 있어, 멤버십 추론 공격(Membership Inference Attack) 등을 통해 원본 데이터가 역

추출될 위험이 있다. 모델 파일의 접근통제·암호화·반출 승인 절차가 요구사항으로 정의되어야 한다.

▶ 훈련 인프라 격리: GPU 서버, 학습 파이프라인, 하이퍼파라미터 관리 시스템 등 훈련 인프라가 업무망이나 인터넷망과 분리되지 않으면, 외부 공격자가 훈련 프로세스에 개입하여 모델을 조작하거나 데이터를 탈취할 수 있다.

▶ 공급망 보안: 사전 학습 모델(Pre-trained Model), 오픈소스 라이브러리, 학습 프레임워크 등 외부에서 도입하는 소프트웨어 구성요소의 무결성 검증이 요구사항에 반영되지 않으면, 악성코드가 포함된 구성요소를 통해 내부망이 침해될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	학습 데이터 반입 승인·매체 통제 요건 명시 여부	학습 데이터 반입 시 승인 절차 및 매체 통제 요건이 {보안요구사항}에 명시되었는지 확인
2	학습 데이터 저장 시 암호화 요건 포함 여부	학습 데이터 저장 시 암호화(At-Rest Encryption) 요건이 {보안요구사항}에 포함되었는지 확인
3	개인정보 학습 데이터 비식별화·가명처리 요건 명시 여부	개인정보 포함 학습 데이터에 대한 비식별화·가명처리 요건이 {보안요구사항}·{요구사항정의서}에 명시되었는지 교차 검토
4	학습 완료 후 데이터 안전 폐기 절차 포함 여부	학습 완료 후 학습 데이터(원본·중간 산출물)의 안전한 폐기 절차가 {보안요구사항}에 포함되었는지 확인
5	학습 데이터 무결성 검증 요건 포함 여부	해시 비교 등 학습 데이터 무결성 검증 요건이 {보안요구사항}에 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	학습 데이터 반입 승인·매체 통제, 저장 시 암호화, 비식별화·가명처리, 안전한 폐기, 무결성 검증 등 망분리 환경 모델 훈련의 보안 요건이 {보안요구사항}에 구체적으로 명시된 경우	현행 유지
보완권고 (△)	보안 요건이 대체로 정의되었으나, 폐기 절차·무결성 검증 등 일부 항목이 누락되었거나 선언적 수준으로 기술되어 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	망분리 환경에서의 모델 훈련 보안 요건이 {보안요구사항}에 사실상 정의되지 않았거나, 개인정보 포함 학습 데이터의 비식별화 요건이 부재한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선언적 문구와 구체적 요구사항의 구분: '망분리를 적용한다', '보안을 강화한다'와 같은 선언적 문구는 요구사항으로서의 구체성이 부족하다. 감리원은 {보안요구사항}에 망분리 방식(물리적/논리적), 암호화 알고리즘(AES-256 등), 접근통제 방식(RBAC, MFA 등)이 측정 가능한 수준으로 기술되어 있는지를 확인해야 한다.

2) AI 훈련 환경과 일반 서버 환경의 보안 차이 인식: AI 모델 훈련 환경은 일반 서버와 달리 대용량 데이터 I/O, GPU 클러스터 간 고속 통신(NVLink, InfiniBand 등), 장시간 연속 가동이라는 특성이 있다. 이러한 특성을 반영한 보안 요건(고속 통신 구간 암호화, GPU 메모리 보호 등)이 누락되지 않았는지 유의해야 한다.

3) 운영 단계 산출물의 사전 정의 확인: 요구정의 단계에서 직접 운영 환경을 점검하지는 않으나, 운영 단계에서 필요한 산출물(모델 반출 승인 대장, 학습 데이터 반입 이력 관리 대장, 보안 감사 로그 보관 정책 등)이 요구사항으로 사전 정의되어 있는지를 확인해야 한다.

4) RFP 원본과의 교차 검증: 발주기관이 제안요청서(RFP)에 명시한 망분리·보안 요구사항이 사업자의 {보안요구사항}에 빠짐없이 반영되었는지 {과업대비표}를 활용하여 역추적 점검을 수행한다. 특히 RFP에 '내부망 폐쇄 환경에서의 학습'이 명시된 경우, 이에 대응하는 구체적 보안 요건이 {보안요구사항}에 존재하는지 반드시 확인한다.

5) 대용량 AI 자산의 망연계(망간 자료전송) 요건 구체화: 망분리 환경에서는 외부의 사전학습 모델(수십~수백 GB)이나 최신 오픈소스 패키지를 내부망으로 반입해야 한다. 이에 대해 기존 일반 망연계 솔루션의 용량 한계를 극복할 수 있는 '안전하고 검증된 대용량 파일 반입 절차 및 시스템 요건(악성코드 스캔 포함)'이 요구사항에 정의되어 있는지 확인해야 한다.

AG_SD41-1-21 | AI신뢰성 테스트

시스템구조·보안 | 요구정의·분석 | 인공지능 모델 개발 | AI QA체계

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서 사업자가 수립한 {테스트계획서}에 AI 모델의 신뢰성 검증을 위한 특화 테스트 항목과 측정 가능한 합격 기준이 명시되어 있는지, AI 모델 특화 테스트 일정이 전체 사업 일정에 구체적으로 반영되어 있는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

이 단계에서의 점검은 후속 개발·시험 단계의 품질 분쟁을 사전에 예방하는 데 실질적인 효과가 있으며, 모호한 테스트 기준이 계획서에 포함될 경우 검사(인수) 단계에서 발주기관과 사업자 간의 수용 여부를 둘러싼 분쟁이 발생할 수 있다.

측정 불가능한 테스트 기준과 AI 모델 특화 항목의 누락은 이 가이드라인의 핵심 점검 대상이다.

2. 필요성

▶ AI 시스템은 동일한 입력에 대해 항상 동일한 출력을 보장하지 않는 비결정론적(Non-deterministic) 특성을 가지며, 환각(Hallucination), 편향(Bias), 탈옥(Jailbreak), 적대적 입력(Adversarial Attack) 등 일반 정보시스템에는 존재하지 않는 고유한 신뢰성 위험을 내포한다.

▶ 테스트 기준의 조기 확정 필요: AI 모델의 출력 품질(정확도, 유해성 차단율 등)에 대한 합격 기준은 개발 완료 이후에 설정하면 분쟁의 원인이 된다. 수치화된 기준을 확정해야 개발자와 발주기관 모두 명확한 목표를 공유할 수 있다.

▶ AI 특화 테스트 일정의 독립 배정 필요: 생성형 AI 서비스의 모델 성능 평가, 편향 검증, 레드팀 테스트 등은 기존 단위·통합 테스트와 성격이 상이하며 별도의 전문 인력과 시간이 요구된다. 이를 일반 테스트 일정에 통합할 경우 일정 부족으로 인한 품질 저하 및 미흡한 검증이 발생할 수 있다.

▶ TTA·NIST 기준의 사업 반영 의무화: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야)와 NIST AI RMF는 AI 시스템 생명주기 전 단계에서의 TEVV(테스트·평가·검증·확인) 활동 수행을 요구하고 있으며, 공공 AI 사업은 해당 기준을 준수할 의무가 있다. 감리원은 이 기준이 테스트 계획에 반영되었는지 확인해야 한다.

▶ 인수 단계 분쟁 예방: 모호한 테스트 기준과 AI 특화 항목의 누락은 사업 최종 단계인 검사(인수) 단계에서 "AI가 제대로 동작하는가"에 대한 발주기관과 사업자 간 이견을 초래한다. 이를 점검·시정하는 것이 가장 비용 효율적인 품질보증 방법이다.

3. 검토방법

No	점검 항목	확인 방법
1	{테스트계획서}에 AI 신뢰성 테스트 계획 반영 여부	{테스트계획서}에 환각 탐지, 편향 검증, 탈옥 방어, 프롬프트 인젝션 탐지 등 AI 모델 특화 테스트 항목이 명시되어 있는지 문서 검토로 확인
2	AI 신뢰성 테스트의 합격 기준(성능 지표) 수치화 여부	"정확도 ○○% 이상", "금칙어 차단율 ○○% 이상" 등 측정 가능한 수치 기준이 {테스트계획서} 또는 {요구사항정의서}에 명시되어 있는지 확인
3	AI 모델 특화 테스트 일정 포함 여부	{사업수행계획서}의 전체 테스트 일정 내에 AI 특화 테스트 일정이 별도로 구분·배정되어 있는지 일정 비교로 확인
4	테스트 케이스 설계 계획의 AI 특성 반영 여부	AI 모델의 비결정론적 출력 특성을 고려한 테스트 케이스 설계 방법(다중 실행 결과 비교, 통계적 합격 기준 등)이 계획되어 있는지 문서 검토로 확인
5	제3자 독립 테스트 또는 레드팀(Red Team) 계획 포함 여부	레드팀 테스트 또는 외부 독립 평가 계획이 {테스트계획서} 또는 {사업수행계획서}에 포함되어 있는지 확인
6	테스트 데이터셋의 대표성 및 편향 방지 계획 여부	테스트 데이터셋의 구성 기준(성별·연령·지역 등 다양성 고려)과 학습·테스트 데이터 분리 원칙이 {테스트계획서}에 명시되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{테스트계획서}에 AI 특화 테스트 항목(환각·편향·탈옥·적대적 입력 등)과 측정 가능한 수치 합격 기준이 명시되어 있고, AI 특화 테스트 일정이 전체 일정에 별도 배정되어 있으며, 레드팀 계획과 테스트 데이터셋 구성 기준까지 반영됨	현행 유지
보완권고 (△)	AI 특화 테스트 항목은 일부 반영되어 있으나 합격 기준이 정성적 문구에 그치거나, AI 테스트 일정이 일반 테스트 일정에 흡수되어 있는 등 일부 요건이 미흡함	해당 항목 보완 요청 후 재확인
부적합 (X)	{테스트계획서}에 AI 모델 특화 테스트 항목·합격 기준·일정이 전혀 반영되어 있지 않거나 "적절하게 동작해야 한다" 수준의 모호한 문구만 존재함	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 합격 기준의 수치화 유도: "적절하게 동작해야 한다" 등 정성적 문구는 인수 단계 분쟁의 원인이 되므로, "정확도 90% 이상", "금칙어 차단율 99% 이상"과 같이 측정 가능한 수치 기준으로 구체화하도록 유도한다.

2) AI 특화 테스트 일정의 독립 배정 확인: 모델 성능 평가·편향 검증·레드팀 테스트가 일반 단위·통합 테스트 일정에 흡수되어 있으면 일정 부족으로 검증이 부실해지므로, WBS에서 별도 활동으로 구분·배정되었는지 확인한다.

3) 비결정론적 특성 반영 여부 점검: 동일 입력에 대해 출력이 달라질 수 있는 AI 특성상, 다중 실행 결과 비교나 통계적 합격 기준(예: N회 실행 중 합격률) 적용 계획이 테스트 케이스 설계에 반영되었는지 확인한다.

4) 레드팀·제3자 평가 계획 확인: 적대적 공격 대응 검증은 개발팀 자체 테스트만으로는 한계가 있으므로, 레드팀 테스트 또는 외부 독립 평가 계획의 포함 여부를 점검하고 미포함 시 감리의견 작성을 검토한다.

5) TTA·NIST 기준 매핑 확인: TTA 신뢰할 수 있는 인공지능 개발안내서와 NIST AI RMF의 TEVV 요구사항이 테스트 계획에 어떻게 반영되었는지 매핑표 등 근거 자료를 요구하여 확인한다.

AG_SD41-1-22 | 성능저하 임계값 요건점검

시스템구조·보안 | 요구정의·분석 | 운영·모니터링 | AI QA체계

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 시스템의 성능 저하 임계값 기준 및 허용 범위를 {요구사항정의서}에 구체적이고 측정 가능한 형태로 반영하였는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, "AI가 적절히 응답해야 한다"와 같은 정성적·추상적 문구만 기술된 경우에는 이후 검사·인수 단계에서 성능 분쟁의 원인이 되므로, 응답시간 상한선, 정확도 최솟값, 서비스 가용률, 모델 성능 재훈련 기준 등 측정 가능한 임계값이 요구사항으로 명시되었는지를 확인한다.

2. 필요성

▶ AI 서비스는 일반 정보시스템과 달리 모델 드리프트(Model Drift), 데이터 분포 변화, 추론 부하에 따른 응답 지연 등 비결정적(non-deterministic) 성능 저하 패턴을 가진다. 이로 인해 요구정의 단계에서 성능 임계값 기준을 명확히 규정하지 않으면 다음과 같은 리스크가 발생한다.

▶ 인수 기준 불명확: 성능 검사·인수 시 발주기관과 사업자 간 "적합 여부" 해석 차이로 분쟁 발생

▶ 품질 확인 불가: 감리원이 성능 수준의 적정성을 판단할 객관적 기준이 없어 점검 실효성 저하

▶ 운영 기준 미비: 개발 완료 후 모니터링·재훈련 기준이 불명확하여 서비스 품질 관리 실패

▶ 책임 소재 불명확: 성능 저하 발생 시 원인 규명 및 계약 이행 책임 판단 근거 부재

3. 검토방법

No	점검 항목	확인 방법
1	AI 응답 시간 상한 기준 명시 여부	평균 응답 3초 이내·95% 요청 5초 이내 등 AI 서비스 응답 시간 상한 기준이 {요구사항정의서}에 수치로 명시되었는지 확인
2	AI 성능 지표 최솟값 기준 명시 여부	정밀도·재현율·F1 등 AI 모델 정확도·성능 지표의 최솟값 기준이 {요구사항정의서}·'AI 품질 기준서'에 명시되었는지 확인
3	가용률·RTO/RPO 명시 여부	AI 서비스 가용률(Availability)·장애 복구 목표시간(RTO/RPO)이 {요구사항정의서}에 명시되었는지 확인

No	점검 항목	확인 방법
4	성능 저하 임계값·대응 절차 포함 여부	성능 저하 발생 기준(임계값)과 초과 시 대응 절차(알림·재훈련·서비스 제한 등)가 {요구사항정의서}·'성능 목표 정의서'에 포함되었는지 확인
5	성능 측정 시나리오 정의 여부	정상 부하·최대 부하·동시 접속자 수 등 성능 측정 기준 시나리오가 {요구사항정의서}·'사업수행계획서'에 정의되었는지 확인
6	모델 재훈련 주기·기준 반영 여부	재훈련 주기 또는 성능 저하 %·데이터 누적량 등 재훈련 기준이 {요구사항정의서}에 반영되었는지 확인하고 사업자 면담으로 보완
7	AI 성능 요구사항 추적 연결 여부	RFP상 AI 성능 요구사항이 {요구사항정의서}의 해당 항목과 빠짐없이 연결되어 있는지 {요구사항추적표}·'제안요청서(RFP)'와 교차 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	응답 시간 상한, 정확도 최솟값, 가용률·RTO/RPO, 성능 저하 임계값과 초과 시 대응 절차, 재훈련 기준, 측정 시나리오가 {요구사항정의서}에 측정 가능한 수치로 명시되고 RFP 요구사항과 빠짐없이 추적 연결된 경우	현행 유지
보완권고 (△)	일부 성능 지표는 수치화되었으나 임계값 초과 시 대응 절차, 재훈련 기준, 측정 시나리오 등 일부 항목이 누락되었거나 정성적으로만 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	성능 저하 임계값 기준이 전혀 정의되지 않았거나, "AI가 적절히 응답해야 한다" 수준의 추상적 문구만 존재하여 인수·검사 기준으로 활용할 수 없는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 성능저하 임계값이 서비스별로 차별화되어 있는지 확인한다: '응답 시간 3초 이상 시 경고'처럼 단일 기준을 모든 기능에 적용하면, AI 추론 지연이 큰 기능과 단순 조회 기능의 특성 차이를 반영하지 못한다. 감리원은 기능 유형별 임계값이 구분 정의되었는지 확인한다.

2) 임계값 모니터링 주기와 알림 대상이 명확히 정의되었는지 확인한다: 임계값이 있더라도 실시간 감지·알림 체계가 없으면 장애 인지가 늦어진다. 감리원은 측정 주기, 알림 수신자, 초과 시 대응 절차(알림·재훈련·서비스 제한)가 요구사항에 포함되었는지 확인한다.

3) 정성적 표현은 요건이 아니다: "AI가 적절히 응답해야 한다" 수준의 문구는 인수·검사 단계에서 반드시 분쟁을 유발한다. 감리원은 응답시간 상한, 정확도·정밀도·재현율의 최솟값, 가용률과 RTO/RPO가 모두 수치로 명시되었는지 확인하고, 전면 누락 시 부적합 판정을 검토한다.

4) 재훈련 기준이 요구사항에 포함되었는지 확인한다: 모델 드리프트는 개발이 아니라 운영 단계에서 현실화되므로, 요구정의 단계에서 재훈련 주기 또는 재훈련 트리거(성능 저하 비율, 데이터 누적량 등)가 정의되지 않으면 운영 품질 관리의 근거가 사라진다.

5) 측정 시나리오 없이 정의된 임계값은 검증할 수 없다: 동일한 정확도 목표라도 정상 부하와 최대 부하에서 결과가 달라진다. 감리원은 측정 시나리오(정상·최대 부하, 동시 접속자 수)가 정의되었는지 확인하고, RFP의 성능 요구사항이 {요구사항추적표}에 빠짐없이 연결되었는지 교차 검토한다.

AG_SD41-1-23 | 성능저하 알림 모니터링 인프라 요건

시스템구조·보안 | 요구정의·분석 | 운영·모니터링 | 모니터링/재학습

근거 | NIST AI RMF · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 모델의 추론(Inference) 성능 저하를 실시간으로 감지하여 알릴 수 있는 모니터링 인프라 요건이 {시스템요구사항}에 구체적으로 정의되었는지, 그리고 APM 등 성능 감시 도구와의 연동 요건이 명시되었는지를 감리원이 제3자 관점에서 검토하기 위한 기준을 제시하는 것을 목적으로 한다.

AI 시스템은 GPU/NPU 자원 경합, 모델 버전 교체, 동시 요청 폭증 등 일반 정보시스템에서는 발생하지 않는 고유한 성능 변동 요인을 가진다.

이러한 요인이 요구정의 단계에서 모니터링 대상 지표·임계값·알림 채널로 구체화되지 않으면, 이후 설계·구현 단계에서 모니터링 공백이 발생하고 운영 중 장애 탐지가 지연되는 위험이 있다.

2. 필요성

▶ 생성형 AI 시스템의 성능 모니터링 요건을 검토해야 하는 필요성은 다음과 같다.

▶ AI 추론 지연(Latency) 특성: 생성형 AI 모델은 토큰 단위로 순차 생성하므로 첫 토큰 응답 시간(TTFT)과 초당 토큰 처리량(TPS)이라는 전통적 웹 응답시간과는 다른 성능 지표가 필요하다. 이를 요구정의 시점에 정의하지 않으면 일반적인 HTTP 응답시간만 측정하게 되어 실제 사용자 체감 품질을 포착할 수 없다.

▶ GPU/NPU 자원 포화 위험: AI 워크로드는 GPU 메모리(VRAM), 연산 유닛 점유율, 배치 큐 대기 길이 등 전용 하드웨어 자원에 크게 의존한다. 이 자원이 포화되면 응답 지연이 급격히 증가하거나 OOM(Out of Memory) 오류가 발생하며, 이를 사전에 감지하려면 AI 전용 리소스 모니터링 요건이 요구사항에 포함되어야 한다.

▶ APM 도구 연동의 조기 확보: AI 추론 파이프라인은 전처리·모델 추론·후처리·가드레일 등 다수의 컴포넌트로 구성된다. 각 구간의 병목을 식별하려면 분산 트레이싱 기반 APM 도구와의 연동 요건이 요구정의 시점에 명시되어야 계측 포인트(Instrumentation Point)를 반영할 수 있다.

▶ 모델 드리프트 조기 탐지 기반 확보: 생성형 AI는 입력 데이터 분포 변화에 따라 출력 품질이 저하될 수 있다. 성능 모니터링 인프라에 모델 품질 지표(예: 환각률, 응답 적합도 점수)를 포함하도록 정의해 두면, 운영 중 모델 드리프트를 조기에 탐지하는 기반이 된다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 전용 성능 지표 정의 여부	생성형 AI 모델의 고유 성능 지표(TTFT, TPS, GPU 사용률, VRAM 점유율, 배치 큐 길이 등)가 {시스템요구사항}에 구체적 수치(임계값)와 함께 정의되어 있는지 확인

No	점검 항목	확인 방법
2	알림 임계값 및 에스컬레이션 절차 정의 여부	성능 저하 시 알림을 발송할 임계값이 단계별(Warning/Critical)로 수치화되어 있고, 에스컬레이션 절차(담당자, 채널, 대응 시한)가 명시되어 있는지 확인
3	APM 도구 연동 요건 정의 여부	AI 추론 파이프라인의 각 구간(전처리→모델 추론→후처리→가드레일)에 대한 분산 트레이싱 연동 요건이 {시스템요구사항}에 포함되어 있는지 확인
4	GPU/NPU 자원 모니터링 요건 정의 여부	GPU 메모리, 연산 유닛 사용률, 온도, 전력 소비 등 AI 전용 하드웨어 자원 모니터링 요건이 정의되어 있으며, 기존 서버 모니터링과 분리 또는 통합 방안이 명확한지 확인
5	모델 품질 지표 모니터링 요건 정의 여부	모델 출력 품질 관련 지표(환각률, 응답 적합도, 거부율 등)의 모니터링 방안이 요구사항에 포함되어 있는지 확인
6	모니터링 대시보드 구성 요건 정의 여부	실시간 대시보드 구성 요건(갱신 주기, 시각화 항목, 접근 권한)이 요구사항에 포함되어 있는지 확인
7	로그 수집·보존 요건 정의 여부	AI 추론 로그(입력 프롬프트 해시, 토큰 수, 응답 시간 등)의 수집 범위, 보존 기간, 개인정보 비식별화 방안이 정의되어 있는지 확인
8	부하 테스트 기준 연계 정의 여부	성능 모니터링 임계값 도출의 근거가 되는 부하 테스트 시나리오(동시 사용자 수, 프롬프트 길이 분포 등)가 요구사항에 연계 정의되어 있는지 확인
9	운영 이관용 산출물 정의 여부	개발 완료 후 운영 조직에 인계할 모니터링 설정 문서, 알림 규칙, 대시보드 템플릿 등 운영 산출물이 요구정의에 포함되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 전용 성능 지표(TTFT, TPS, GPU 사용률 등)가 수치화된 임계값과 함께 {시스템요구사항}에 정의되어 있고, APM 도구 연동 요건과 단계별 알림 임계값·에스컬레이션 절차가 구체적으로 기술되며, GPU/NPU 자원 모니터링, 모델 품질 지표 모니터링, 대시보드 구성, AI 추론 로그의 수집 범위·보존 기간·비식별화 방안, 임계값 도출 근거가 되는 부하 테스트 시나리오 연계, 운영 이관용 산출물이 모두 정의된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	AI 전용 성능 지표는 일부 정의되어 있으나, 임계값이 수치화되지 않았거나 APM 연동 요건이 개괄적 수준에 그치는 경우, 또는 모델 품질 지표·운영 이관 산출물 중 일부가 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 전용 성능 지표가 전혀 정의되지 않고 일반적인 서버 응답시간만 기재되어 있거나, APM 도구 연동 요건 자체가 누락된 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 성능 요건의 수치화 유도: "AI는 빠르게 응답해야 한다"와 같은 모호한 문구는 분쟁의 원인이 된다. "TTFT ≤ 2초, 동시 100명 기준 TPS ≥ 15"와 같이 측정 가능한 수치로 구체화하도록 유도하는 것이 바람직하다.

2) 기존 서버 모니터링과 AI 모니터링의 구분: CPU/Memory 등 일반 인프라 모니터링은 별도의 감리 항목(일반 IT 감리)에서 점검하므로, 본 가이드라인에서는 GPU/NPU, 모델 추론 지연, 토큰 생성 속도 등 AI 고유 지표에 집중하여 점검한다. 두 영역이 혼재되지 않도록 유의한다.

3) APM 도구의 AI 파이프라인 지원 여부 확인: 기존 APM 도구가 AI 추론 파이프라인의 분산 트레이싱을 지원하는지는 도구마다 차이가 크다. 감리원이 특정 제품을 평가하는 것은 지양하되, 요구사항에 "AI 추론 구간별 트레이싱 지원"이라는 기능 요건이 기술되어 있는지를 확인하는 것이 핵심이다.

4) 운영 산출물 점검의 범위: 본 가이드라인은 개발 과정에서 작성하여 운영에 인계해야 할 모니터링 관련 산출물(설정 문서, 알림 규칙, 대시보드 템플릿)의 요건 정의 여부를 점검하는 것이며, 실제 운영 과정의 모니터링 활동 자체는 점검 범위에서 제외한다.

5) AI 특화 성능 및 자원 임계값 모니터링 지표 명확화: 일반적인 CPU/RAM 외에 AI 특화 지표인 '초당 토큰 생성 수(Tokens Per Second)', 'GPU VRAM 사용량 및 누수(Leak)', 'OOM(Out of Memory) 에러 발생 빈도' 등이 모니터링 대상 요구사항에 명시적으로 포함되어 있는지 점검하도록 추가한다.

AG_SD41-1-24 | 알림 라우팅체계 점검

시스템구조·보안 | 요구정의·분석 | 운영·모니터링 | 모니터링/재학습

근거 | NIST AI RMF MANAGE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 생성형 AI 시스템의 장애·이상 상황 발생 시 담당자에게 자동으로 알림(Alert)을 전달하는 라우팅 체계가 {시스템요구사항}에 구체적으로 정의되어 있는지 검토하는 것을 목적으로 한다.

AI 시스템은 모델 추론 지연(Latency Spike), GPU 메모리 부족(OOM), 외부 API 연계 장애, 환각(Hallucination) 빈도 급증 등 일반 정보시스템에서는 발생하지 않는 고유한 장애 유형을 가지며, 이러한 이상 징후를 조기에 탐지하여 적절한 담당자에게 즉시 전달하는 자동 통지 인프라가 요구사항 단계에서 명시되지 않으면 설계·구현 단계에서 알림 체계가 누락되거나 부실하게 구축될 위험이 있다.

본 가이드라인은 '시스템 구조 및 보안' 감리원이 이러한 요건의 적정성을 점검하기 위한 검토 방향과 세부 기준을 제시한다.

2. 필요성

▶ 생성형 AI 시스템은 대규모 언어모델(LLM) 기반의 추론 처리, GPU 클러스터 운용, 외부 데이터 소스 연계 등 복합적인 인프라 구성요소를 포함하며, 각 구성요소에서 발생하는 장애의 유형과 전파 경로가 일반 시스템과 상이하므로, 개발 초기 단계인 요구정의 시점에서 알림 라우팅 체계를 요구사항으로 명확히 정의할 필요가 있다.

▶ AI 고유 장애의 조기 감지 필요: GPU OOM, 모델 추론 타임아웃, 토큰 생성 속도 저하, 외부 API 연계 실패 등 AI 시스템 특유의 장애 유형은 일반 서버 모니터링(CPU/Memory)만으로 탐지할 수 없으므로, AI 전용 메트릭 기반의 알림 체계가 정의되어야 한다.

▶ 에스컬레이션 경로의 사전 정의: AI 시스템 장애는 인프라 담당자뿐 아니라 모델 운영자, 데이터 엔지니어, 보안 담당자 등 다수의 역할에 걸쳐 전파되므로, 장애 유형별 에스컬레이션 경로가 요구사항에 명시되지 않으면 운영 단계에서 대응 지연이 발생한다.

▶ 보안 이벤트 자동 통지 체계: 프롬프트 인젝션(Prompt Injection), 모델 탈옥(Jailbreak) 시도, 비정상 API 호출 패턴 등 AI 보안 이벤트는 실시간 자동 통지가 필수적이며, 이를 위한 인프라 요건이 정의되어야 한다.

▶ 국내외 기준의 요구: NIST AI RMF의 MANAGE 4.1은 배포 후 모니터링 계획에 인시던트 대응 및 복구 메커니즘을 포함하도록 하고 있으며, TTA 신뢰할 수 있는 인공지능 개발안내서는 AI 시스템의 안전성 확보를 위한 실시간 감시 체계 구축을 권고하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 전용 메트릭 모니터링 명시 여부	추론 지연(Latency)·GPU 사용률·메모리(VRAM) 사용량 등 AI 전용 메트릭이 모니터링 대상으로 {시스템요구사항}에 명시되었는지 확인
2	AI 출력 품질 메트릭 포함 여부	환각(Hallucination) 빈도·토큰 생성 속도·응답 품질 지표 등 AI 출력 품질 메트릭이 모니터링 대상에 포함되었는지 {시스템요구사항} 검토·사업자 면담으로 확인
3	외부 AI API 연계 상태 모니터링 요건 포함 여부	외부 AI API(LLM API·벡터 DB 등) 연계 상태 모니터링 요건이 {시스템요구사항}·'아키텍처 설계서(초안)'에 포함되었는지 대조
4	메트릭별 임계값 수치 정의 여부	응답시간 3초 초과 시 알림 등 각 메트릭별 임계값(Threshold) 기준값이 {시스템요구사항}에 수치로 정의되었는지 확인
5	장애 유형별 알림 라우팅·에스컬레이션 경로 정의 여부	장애 유형·심각도별 알림 수신 담당자(인프라·모델 운영·보안 등)와 단계적 확대(에스컬레이션) 경로·시간 기준, 보안 이벤트의 분리 통지 경로가 {시스템요구사항}에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 전용 메트릭(추론 지연, GPU 사용률, VRAM 사용량 등), 출력 품질 관련 메트릭, 외부 AI API 연계 상태 모니터링 요건, 메트릭별 임계값 수치 기준 및 장애 유형별 알림 라우팅·에스컬레이션 경로가 {시스템요구사항}에 구체적으로 명시되어 있는 경우	현행 유지
보완권고 (△)	알림·모니터링 요건은 정의되어 있으나 일부 AI 전용 메트릭이 누락되었거나, 임계값 기준값 등 정량적 기준이 수치로 정의되지 않아 구체성이 부족한 경우	구체적 임계값·기준 수립 등 해당 항목 보완 요청 후 재확인
부적합 (X)	알림 라우팅 체계 및 자동 통지 인프라 요건이 요구사항에 정의되지 않았거나, CPU·메모리 등 일반 인프라 메트릭만 나열되어 AI 시스템 특성이 전혀 반영되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 요건의 수치화: '장애 발생 시 즉시 알림'과 같은 모호한 문구는 향후 검사 단계에서 분쟁의 원인이 된다. '추론 응답시간 3초 초과 시 1차 알림 발송', '1차 알림 후 15분 미대응 시 2차 에스컬레이션' 등 측정 가능한 기준으로 구체화하도록 사업자에게 유도해야 한다.

2) AI 전용 메트릭과 일반 메트릭의 구분: 요구사항에 모니터링 항목이 기술되어 있더라도, CPU·메모리 등 일반 인프라 메트릭만 나열된 경우 AI 시스템 특성이 반영되지 않은 것이다. 반드시 AI 전용 메트릭(추론 지연, GPU VRAM, 환각 빈도 등)의 포함 여부를 별도로 확인해야 한다.

3) 보안 이벤트 알림의 분리 확인: 프롬프트 인젝션, 모델 탈옥 등 AI 보안 이벤트에 대한 알림 경로가 일반 장애 알림 경로와 분리되어 정의되었는지 확인해야 한다. AI 보안 이벤트는 보안팀으로의 즉각 통지 경로가 별도로 필요하다.

4) 외부 API 연계 장애의 특수성: 생성형 AI 시스템은 외부 LLM API, 벡터 데이터베이스, 검색 엔진 등 다수의 외부 서비스와 연계되는 경우가 많다. 외부 API 장애는 내부 시스템과 다른 알림·에스컬레이션 경로가 필요하므로, 요구사항에 외부 연계 구간별 알림 체계가 구분되어 있는지 확인해야 한다.

5) 요구정의 단계와 운영 단계의 구분: 본 가이드라인은 요구정의 단계에서 '요구사항으로 정의되었는지'를 점검하는 것이지, 실제 알림 시스템이 작동하는지를 검증하는 것이 아니다. 감리원은 요구사항 문서의 완전성과 구체성에 집중하여야 하며, 구현 적정성은 설계 또는 검사 단계 감리에서 확인한다.

AG_SD42-1-01 | 개인정보 비식별화 절차 설계

시스템구조·보안 | 설계 | 데이터 수집·처리 | 비식별화

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서, 기획/분석(요구사항 정의) 단계에 수립된 개인정보 비식별화 요구사항이 설계 산출물에 누락 없이 구체적으로 반영되었는지, 그리고 개발·테스트 환경에서의 개인정보 처리(파기·마스킹) 절차가 설계 수준에서 명문화되었는지를 감리원이 점검하기 위한 것이다.

개인정보 비식별화 절차 및 마스킹 설계는 단순 개인정보보호 준수 여부를 넘어, AI 모델 학습·검증 품질과 직결되는 핵심 설계 항목으로, 설계 단계에서 미반영 시 구현·테스트 단계에서의 재작업 비용이 급증하고, 실 서비스에서 개인정보 침해 사고로 이어질 위험이 존재한다.

감리원은 기술적 구현 방식이 아닌, 설계 산출물의 완결성·구체성·일관성 관점에서 점검한다.

2. 필요성

▶ AI 서비스는 학습 데이터, 테스트 데이터, 운영 데이터에 실제 개인정보가 혼입될 가능성이 높아, 일반 IT 시스템 대비 개인정보 처리 위험이 구조적으로 높다. 특히 생성형 AI는 학습 과정에서 민감정보를 암기하거나 재현할 수 있어, 비식별화 설계의 엄밀함이 요구된다.

▶ 설계 단계에서 다음과 같은 문제가 발생하면 이후 단계에서 수정 비용이 급증한다.

▶ 비식별화 대상 항목이 미정의된 채 개발이 시작될 경우, 학습 데이터셋에 개인정보가 포함되어 AI 모델이 민감정보를 학습·생성하는 사고로 이어진다.

▶ 테스트 환경 마스킹 절차가 미설계될 경우, 개발·테스트 과정에서 실 개인정보가 무단 노출될 수 있으며, 이는「개인정보 보호법」위반에 해당한다.

▶ 요구사항 단계에서 정의된 비식별화 요건이 설계서에 반영되지 않으면 요구사항-설계 불일치(traceability gap)로 인해 검수 분쟁 및 재설계 비용이 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	비식별화 대상 개인정보 항목 명시 여부	비식별화 대상 개인정보 항목이 {상세설계서} 또는 {사업수행계획서}에 명시되었는지 개인정보 항목 목록에서 확인

No	점검 항목	확인 방법
2	항목별 비식별화 기법 구분 설계 여부	가명처리·익명처리·마스킹·삭제 등 개인정보 항목별 비식별화 기법이 구분하여 설계되었는지 기법별 항목 매핑표에서 확인
3	비식별화 처리 시점 명시 여부	수집 시·저장 시·학습 전·테스트 시 등 비식별화 처리 시점이 설계서 데이터 흐름도 또는 처리 절차에 명시되었는지 확인
4	테스트 데이터 마스킹·파기 절차 설계 여부	개발·테스트 환경 사용 데이터의 마스킹 또는 파기 절차가 '테스트 데이터 관리 계획서' 또는 {상세설계서}에 설계되었는지 확인
5	비식별화 요건 설계 추적 연결 여부	요구사항 정의서의 비식별화 요건이 {상세설계서}에 추적 가능한 형태로 연결(traceability)되었는지 요구사항-설계 추적표 또는 요건 인용에서 확인
6	재식별 가능성 평가 절차 반영 여부	가명처리 데이터셋 등 비식별화 결과물의 재식별 가능성 평가 절차가 설계서에 반영되었는지 재식별 위험 평가 기준·절차에서 확인
7	개인정보 처리 담당자·역할 지정 여부	개인정보 처리·비식별화 관련 담당자(책임자)·역할이 {사업수행계획서} 또는 관련 설계서에 지정되었는지 역할·책임(RACI) 정의에서 확인
8	비식별화 설계의 법령 부합 검토 여부	비식별화 설계 내용이 「개인정보 보호법」 제28조의2~5·가명정보 처리 가이드라인에 부합하는지 법령 검토 기록(회의록·검토 의견서)에서 확인
9	학습 데이터 개인정보 스크리닝 계획 반영 여부	AI 모델 학습 데이터셋에 대한 개인정보 포함 여부 스크리닝(검사) 계획이 설계서에 반영되었는지 스크리닝 도구·방법·기준·주기 기술에서 확인
10	테스트 후 개인정보 파기 절차·확인 방법 포함 여부	테스트 완료 후 개인정보(마스킹 전 원본 포함)의 파기 절차·파기 확인 방법이 설계서에 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	비식별화 대상 항목 목록, 항목별 기법 매핑, 처리 시점, 테스트 데이터 마스킹·파기 절차, 요구사항-설계 추적성, 재식별 위험 평가 절차, 담당자 지정, 법령 부합 검토, 학습 데이터 스크리닝 계획이 설계 산출물에 모두 구체적으로 반영됨	현행 유지
보완권고 (△)	비식별화 설계는 존재하나 항목별 기법 매핑·처리 시점·재식별 평가 절차·파기 확인 방법·요건 추적 연결 중 일부가 누락되거나 모호하게 기술됨	해당 항목 보완 요청 후 재확인

판정	판정 기준	주요 사례 / 조치
부적합 (X)	개인정보 비식별화 절차 및 테스트 데이터 마스킹·파기 설계가 설계 산출물에 전혀 반영되지 않았거나 '개인정보를 적절히 처리한다' 수준의 선언적 문구만 존재함	시정조치 요구 및 이행 결과 확인 (개인정보 보호법 준수 촉구)

5. 감리 시 유의사항 및 실무 Tip

1) 비식별화 요건의 수치화·구체화 유도: "개인정보를 적절히 처리한다"와 같이 모호하게 기술된 설계 내용은 향후 분쟁의 원인이 된다. 감리원은 "주민등록번호 전체 마스킹(****-*******)", "테스트 데이터셋 내 개인정보 스크리닝 통과율 100%" 등 측정 가능한 기준으로 구체화하도록 지적 의견을 작성한다.

2) 요건 추적성(Traceability) 검토는 교차 확인 방식으로: {요구사항정의서}의 비식별화 관련 요건 목록을 먼저 추출한 후, 해당 요건 ID가 {상세설계서}에 인용되었는지를 역추적한다. 설계서에서 비식별화 내용을 먼저 확인하는 방식은 누락을 발견하기 어려우므로 반드시 요건 목록 → 설계서 대조 순으로 진행한다.

3) 테스트 데이터 마스킹 절차 확인 시 범위 주의: 개발 환경뿐 아니라 스테이징 환경, 외부 협력사 전달 데이터 등 실 개인정보가 유입될 수 있는 모든 경로가 마스킹 절차의 적용 범위에 포함되었는지 확인하고, 테스트 완료 후 마스킹 전 원본 데이터의 파기 절차와 파기 확인 방법까지 설계에 포함되었는지 점검한다.

AG_SD42-1-02 | 성능지표설계

시스템구조·보안 | 설계 | 인공지능 모델 개발 | 모델평가

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 단계에서 정의된 AI 성능 요구사항이 설계 산출물—특히 {테스트계획서}—에 측정 가능한 수치 지표로 구체화되어 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 시스템의 성능은 "충분히 정확해야 한다"와 같은 추상적 표현으로는 사업 완료 시 수용 기준을 판단할 수 없다.

정밀도(Precision), 재현율(Recall), F1-Score 등의 지표에 대해 산식과 목표값이 설계 단계에 문서화되지 않으면, 이후 검사·인수 단계에서 분쟁이 발생하거나 AI 시스템의 품질을 객관적으로 검증하기 어렵다.

본 가이드라인은 이러한 위험을 사전에 통제하기 위한 점검 기준을 제공한다.

2. 필요성

▶ AI 시스템은 일반 소프트웨어와 달리 "정답이 하나가 아닌" 확률적 출력을 산출하므로, 기능 요구사항의 충족 여부를 단순 통과/실패(Pass/Fail) 기준으로만 판단하기 어렵다. 이 때문에 AI 특유의 성능 지표를 설계 단계에서 명확히 정의하는 것이 반드시 필요하다.

▶ 요구사항 추적성 확보: 요구사항 단계에서 정의된 AI 성능 요구사항(예: "문서 분류 정확도 90% 이상")이 {테스트계획서}에 측정 가능한 지표와 목표값으로 연결되어야만 추적 가능한 품질 관리가 이루어진다.

▶ 검사·인수 기준의 객관성: TTA 신뢰할 수 있는 인공지능 개발안내서(생성 AI 기반 서비스)는 AI 서비스의 품질 목표를 정량적으로 정의하고 평가할 것을 요구하며, 이를 위한 지표 설계가 선행되어야 한다.

▶ 편향·오류 탐지 근거 마련: 정밀도와 재현율을 분리하여 설계하면 특정 유형의 오류(예: 위험 상황 미탐지)가 사업 수행 중 조기에 드러나며, 이는 안전하고 신뢰할 수 있는 AI 시스템 구축의 기반이 된다.

▶ NIST AI RMF MEASURE 기능 이행: NIST AI RMF(MEASURE 2.3)는 "AI 시스템 성능 또는 보증 기준을 정량적·정성적으로 측정하고 문서화해야 한다"고 요구한다. 이에 따라 성능 지표 산식이 명시되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	성능 지표 항목의 명시 여부	{테스트계획서}·'AI 성능 요구사항 정의서'에서 정밀도·재현율·F1-Score·BLEU·ROUGE 등 AI 유형에 적합한 지표가 명시되었는지 대조 검토
2	지표 산식 및 계산 방법의 구체성	문서 검토와 담당자 인터뷰로 각 지표의 정의·산식이 기술되거나 국제 표준(ISO/IEC 등)을 인용해 명확히 정의되었는지 확인
3	목표값(임계값) 설정 여부	{요구사항추적표}·{테스트계획서} 교차 검토로 각 지표의 수치 목표(예: 재현율 90% 이상)가 설정되고 요구사항 분석 정의와 일치하는지 확인
4	테스트 데이터셋 설계 적정성	{테스트계획서}·'데이터 품질 계획서'에서 성능 측정용 테스트 데이터셋의 구성(크기·분포·대표성)이 명시되었는지 확인하고 필요 시 인터뷰
5	요구사항-지표 간 추적성	{요구사항추적표}에서 각 AI 성능 요구사항이 {테스트계획서}의 대응 지표 항목으로 연결·추적되는지 전수 검토
6	공정성·편향 관련 지표 포함 여부	문서 검토와 인터뷰로 집단·조건별 성능 편차 지표(예: 하위 그룹별 정밀도 비교)가 설계에 포함되었는지 확인(TTA 생성AI 가이드·NIST MEASURE 2.11 참조)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 유형에 적합한 성능 지표·산식·목표값이 {테스트계획서}에 모두 명시되고, 요구사항과의 추적 연결이 완전하며, 테스트 데이터셋 구성과 공정성·편향 지표까지 설계에 포함됨	현행 유지
보완권고 (△)	지표는 존재하나 산식·목표값이 일부 누락되었거나 요구사항 분석 단계의 목표값과 불일치하고, 테스트 데이터셋 구성 또는 공정성·편향 지표가 미흡함	해당 항목 보완 요청 후 재확인
부적합 (X)	성능 지표가 {테스트계획서}에 전혀 정의되지 않았거나, 요구사항과 연결되지 않은 지표명 나열만 존재하여 인수 기준으로 활용이 불가능함	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 요건은 반드시 수치화 요구: "AI는 적절한 정확도를 보여야 한다"와 같은 문구가 {테스트계획서}에 존재하면 향후 분쟁의 씨앗이 된다. 감리원은 "정밀도 92% 이상, 재현율 88% 이상 (테스트 데이터 5,000건 기준)"처럼 측정 가능한 형태로 구체화하도록 의견을 제시해야 한다.

2) 지표의 존재와 산식의 존재를 구분하여 확인: {테스트계획서}에 "F1-Score"가 언급되어 있더라도 산식이 기재되지 않았거나 어떤 분류 기준(임계값)에서 계산할 것인지가 명시되지 않으면 불완전한 설계이다. 지표명 나열과 산식·계산 조건의 완전한 정의는 다른 수준의 설계임을 구분하여 점검해야 한다.

3) 테스트 데이터셋의 대표성 확인: 데이터셋의 형식적 구비만 확인하지 말고 크기·분포·소수 클래스 포함 여부 등 구성 근거의 충실성을 함께 검토하여, 실제 운영 환경을 대표하는 데이터로 성능이 측정되는지 확인해야 한다.

4) 공정성·편향 지표의 설계 포함 여부 점검: 전체 평균 성능만 측정하면 특정 집단·조건에서의 성능 저하가 은폐될 수 있으므로, 하위 그룹별 성능 편차 측정 지표가 설계에 포함되었는지 확인해야 한다.

AG_SD42-1-03 | 성능 임계값 설계점검

시스템구조·보안 | 설계 | 운영·모니터링 | 모델평가

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 {테스트계획서}에 성능 평가 통과 임계값(모델 배포를 위한 최소 성능 임계값(%))이 구체적이고 측정 가능한 수준으로 정의되었는지 여부를 감리원이 점검하기 위한 검토 방향과 절차를 제공하는 것을 목적으로 한다.

감리원은 테스트 계획 단계에서 임계값이 정의되지 않거나 모호하게 기술될 경우, 이후 모델 배포 가부 결정 과정에서 발주기관·사업자 간 분쟁이 발생할 수 있다는 점을 인식하고, 산출물의 완결성과 명확성을 기준으로 검토하여야 한다.

2. 필요성

▶ AI 모델은 전통적 소프트웨어와 달리 학습 데이터와 모델 구조에 따라 성능이 가변적이며, 동일한 기능을 수행하더라도 업무 맥락에 따라 요구되는 정확도·신뢰도 수준이 다를 수 있다. 이로 인해 성능 기준을 "적절한 답변을 제공"처럼 정성적으로만 기술하는 경우, 테스트 단계 및 검수 단계에서 합격 판정 기준이 모호해지고 프로젝트 분쟁으로 이어질 수 있다.

▶ 또한 TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 기반 서비스 분야) 및 NIST AI RMF는 AI 시스템이 배포 전에 성능 기준(MEASURE 2.3: 배포 환경과 유사한 조건에서의 성능 기준 측정 및 문서화)을 정량적 또는 정성적으로 측정·문서화하고, 기준 미달 시 배포를 유보하거나 재학습을 수행하도록 요구한다.

▶ 따라서 모든 기능에 동일한 단일 임계값을 적용하는 것은 부적절할 수 있다. 서비스 내에서도 위험도가 높은 기능(예: 법적 판단, 의료 정보 제공)과 위험도가 낮은 기능(예: 단순 요약, 일상 대화)을 식별하고, 각 위험 수준(Risk Level)에 따라 차등화된 성능 임계값(예: 고위험 95% 이상, 저위험 85% 이상)을 적용하는 것이 바람직하다.

3. 검토방법

No	점검 항목	확인 방법
1	성능 평가 지표 정의 여부	배포 판단에 사용할 주요 성능 지표(정확도, F1-Score, BLEU, ROUGE 등)가 {테스트계획서}에 명시되어 있는지 문서 검토로 확인. 생성형 AI인 경우 응답 적절성 평가 방법(사람 평가 또는 자동 평가 지표) 기재 여부도 확인

No	점검 항목	확인 방법
2	최소 성능 임계값(%) 수치 정의 여부	각 성능 지표별 배포 가부를 결정하는 최솟값이 "정확도 90% 이상", "유해 콘텐츠 탐지율 95% 이상"처럼 측정 가능한 수치(%)로 명시되어 있는지 문서 검토로 확인
3	임계값과 요구사항의 연계성	{테스트계획서}의 임계값이 요구사항 정의서 또는 RFP에서 합의된 성능 수준과 일치하는지 교차 검토. 임계값이 임의 설정이 아닌 PoC 결과·발주기관 요건에서 도출된 근거가 있는지 확인
4	임계값 미충족 시 처리 절차 정의	임계값 미충족 시 재학습·재검증·배포 보류 등 후속 조치 절차가 {테스트계획서}에 포함되어 있는지 확인. 의사결정 주체(PM, 발주기관 등)와 처리 기한 기재 여부도 확인
5	테스트 데이터셋·환경 조건 명시	임계값을 측정할 테스트 데이터셋의 규모·구성·대표성과 성능 측정 환경(배포 환경과 유사한 조건 여부)이 {테스트계획서}에 기술되어 있는지 문서 검토로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	성능 평가 지표와 지표별 최소 임계값이 수치(%)로 {테스트계획서}에 명시되고, 요구사항·PoC 결과와의 연계 근거, 미충족 시 처리 절차, 테스트 데이터셋·측정 환경 조건까지 모두 정의됨	현행 유지
보완권고 (△)	성능 지표는 정의되었으나 일부 지표의 임계값 수치 미정의, 요구사항과의 연계 근거 부족, 미충족 시 처리 절차 또는 테스트 데이터셋 구성 기술이 미흡함	해당 항목 보완 요청 후 재확인
부적합 (X)	성능 임계값이 {테스트계획서}에 전혀 정의되지 않았거나 "적절한 답변을 제공해야 한다" 수준의 정성적 문구만 존재하여 배포 가부 판정 기준으로 활용이 불가능함	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 임계값 수치의 근거 확인: 임계값 수치 자체의 적정성(예: 정확도 90%가 맞는지)은 감리원의 기술 판단 영역이 아니다. 감리원은 해당 수치가 요구사항·PoC 결과와 연계된 근거가 있는지, 그리고 발주기관이 동의한 수치인지를 확인하는 데 집중해야 한다.

2) 생성형 AI 특성 고려: 생성형 AI 서비스는 분류 모델과 달리 단일 수치 임계값만으로 성능을 정의하기 어렵다. 사람 평가(Human Evaluation) 비율, 유해 콘텐츠 차단율, 응답 거부율 등 복수 지표의 조합이 임계값으로 정의되는 것이 바람직하다. 단일 지표만 기재된 경우 추가 지표 정의를 권고한다.

3) 모호한 기준의 수치화 유도: "AI가 정확한 답변을 해야 한다"와 같은 모호한 문구는 분쟁의 원인이 된다. 감리원은 "유해 콘텐츠 탐지율 95% 이상", "정확도 평가 통과 기준 F1-Score 0.85 이상" 등 측정 가능한 형태로 구체화하도록 유도한다.

4) 테스트 데이터셋 대표성 확인: 임계값이 있더라도 테스트셋이 실제 운영 환경을 반영하지 않으면 배포 후 성능 저하 위험이 있다. 테스트셋이 실제 서비스 예상 데이터 분포와 유사하게 구성되었는지를 {테스트계획서} 검토 시 확인한다.

5) 인터뷰 활용: 문서상 임계값이 기재되어 있더라도 이해 관계자(PM, QA) 인터뷰를 통해 해당 수치가 실제 달성 가능한 수준인지, 형식적 기재에 불과한 것은 아닌지를 확인한다. 인터뷰 결과는 반드시 감리 조서에 기록한다.

AG_SD42-1-04 | 안전모드 Fail-safe 트리거 설계 점검

시스템구조·보안 | 설계 | 시스템 구현 | 가용성/이중화

근거 | NIST AI RMF MANAGE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

AI 서비스 구축 사업에서 도출된 안전모드(Fail-safe) 동작 조건과 장애 탐지 체계에 관한 요구사항이 {시스템아키텍처설계서}의 안전모드 트리거 구조도와 이벤트 버스 및 장애 탐지 에이전트 구조도에 빠짐없이, 상세하고 구체적으로 반영되었는지를 검토하는 것이 목적이다.

특히, AI 추론 서비스의 특성상 모델 오작동, 추론 지연, 환각(Hallucination) 등 일반 IT 시스템과 구별되는 장애 유형이 존재하며, 이에 대한 트리거 조건·전환 흐름·폴백 응답이 설계 산출물에 명확히 규정되어 있는지를 검증한다.

또한, 이벤트 버스를 통한 장애 이벤트 전파 구조와 장애 탐지 에이전트의 역할·책임이 아키텍처 수준에서 정의되어 있는지를 확인한다.

2. 필요성

▶ AI 서비스는 일반 정보시스템과 달리 추론 결과의 불확실성, 모델 버전 충돌, GPU 자원 고갈, 프롬프트 인젝션에 의한 응답 오염 등 다양한 AI 고유 장애 유형에 노출된다. 이러한 장애가 발생했을 때 시스템이 스스로 안전한 상태로 전환(Fail-safe)하거나 운영자에게 신속히 알림을 전달하지 못하면, 공공 서비스에서 오정보 제공·서비스 중단 등 대국민 피해로 직결될 수 있다.

▶ 요구사항-설계 추적성 확보: 기획·분석 단계에서 도출된 안전모드 진입 조건, 복구 절차, 사용자 안내 메시지 요건이 아키텍처 설계서에 누락 없이 반영되었는지 검증해야 한다. 설계 단계에서의 누락은 구현·시험 단계에서 보완하기 어렵다.

▶ AI 고유 장애 유형 설계 반영 필요: 모델 응답 이상, 추론 타임아웃, 환각 탐지 임계값 초과 등 AI 특유의 장애 조건이 설계 산출물에 명시되지 않으면, 이후 시험 단계에서 검증 기준 자체가 부재하게 된다.

▶ 이벤트 버스 기반 장애 전파 구조의 복잡성 관리: MSA(마이크로서비스) 또는 에이전트 아키텍처에서 이벤트 버스를 통해 장애 이벤트가 전파될 때, 전파 경로·구독 컴포넌트·처리 순서가 명확히 설계되어 있지 않으면 장애가 연쇄적으로 확산될 위험이 있다.

▶ NIST AI RMF 및 TTA 지침 준수: NIST AI RMF MANAGE 기능은 AI 시스템이 안전하게 저하(degrade safely and gracefully)될 수 있도록 설계할 것을 명시하며, TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 기반 서비스 분야)는 실시간 모니터링 및 이상 감지 체계의 설계 반영을 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	안전모드 트리거 조건의 요구사항 반영 여부	기획/정의된 안전모드 진입 조건(예: 추론 응답시간 초과, 모델 오류율 임계값, 환각 탐지 점수 초과)이 {시스템아키텍처설계서}의 트리거 구조도에 구체적으로 기술되어 있는지 확인
2	안전모드 상태 전환 흐름의 완전성	안전모드 진입 후 폴백 응답(정적 안내 메시지, 대체 모델 전환, 서비스 일시 차단 등) 및 정상 복귀 조건·절차가 상태 전이 다이어그램 또는 흐름도로 명확히 표현되어 있는지 확인
3	이벤트 버스 구조 및 장애 이벤트 전파 경로 설계	이벤트 버스(또는 메시지 큐)를 통한 장애 이벤트의 발행(Publish)·구독(Subscribe) 구조가 구조도에 표현되어 있으며, 어느 컴포넌트가 어떤 이벤트를 발행하고 어느 컴포넌트가 구독·처리하는지 확인
4	장애 탐지 에이전트의 역할·책임 정의	장애 탐지 에이전트(또는 모니터링 컴포넌트)의 탐지 대상(모델 서버, API 게이트웨이, 추론 파이프라인 등), 탐지 방법(헬스체크, 임계값 비교, 로그 분석 등), 탐지 후 조치 절차가 설계에 정의되어 있는지 확인
5	AI 고유 장애 유형별 처리 설계 포함 여부	일반 시스템 장애(서버 다운, 네트워크 오류)뿐 아니라 AI 특유의 장애 유형(추론 결과 품질 저하·환각 임계값 초과, 모델 버전 충돌, 컨텍스트 윈도우 초과, GPU 자원 포화 등)별 처리 설계가 포함되어 있는지 확인
6	사용자 대면 안내 메시지 설계 반영	안전모드 진입 시 최종 사용자에게 제공할 안내 메시지(서비스 일시 중단 안내, 재시도 요청, 담당자 연락처 등)의 내용·형식·제공 시점이 설계 산출물에 명시되어 있는지 확인
7	요구사항-설계 추적 일관성 검증	'요구사항-설계 추적표' 또는 {시스템아키텍처설계서} 내 추적 매트릭스를 통해, 기획/식별된 Fail-safe 관련 요구사항 각각에 대해 대응하는 설계 항목이 1:1로 매핑되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{시스템아키텍처설계서}에 안전모드 트리거 조건이 정량적으로 명시되고, 폴백 처리·복귀 절차, 이벤트 버스의 발행자·이벤트 유형·구독자, AI 고유 장애 유형 처리 설계가 모두 포함되어 요구사항-설계 추적이 가능한 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	트리거 조건이 추상적으로 기술되어 정량 기준 보완이 필요하거나, 이벤트 버스 구조도의 구독자·이벤트 유형 일부 누락, AI 고유 장애 유형 일부 미포함, 사용자 안내 메시지 설계 미반영 등 1~2개 항목에 미흡이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{시스템아키텍처설계서}에 안전모드 트리거 구조도가 전혀 없거나, 이벤트 버스·장애 탐지 에이전트 관련 기술이 전무하거나, Fail-safe 요구사항이 설계에 반영되지 않아 추적이 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 트리거 조건의 "정량화" 요구: 설계 산출물에 "비정상 상태 발생 시 안전모드 전환"처럼 추상적으로만 기술된 경우, "추론 응답시간 N초 초과, 연속 M회 발생 시", "오류 응답률 N% 초과 시" 등으로 구체화하도록 유도한다. 정량적 기준이 없으면 검증 자체가 불가능해진다.

2) 이벤트 버스 구조 vs. 단순 API 호출 구분: 사업자가 제출한 구조도에서 이벤트 버스가 표현되어 있는지, 아니면 단순 동기식 API 호출 구조만 있는지를 구분하는 것이 중요하다. 안전모드 전환처럼 비동기·실시간성이 중요한 처리 흐름에서 동기식 호출만 설계되어 있다면, 실제 장애 상황에서 전파 지연이 발생할 수 있음을 지적 근거로 활용할 수 있다.

3) AI 고유 장애 유형의 포함 여부가 핵심: 서버 다운·네트워크 단절 등 일반 IT 장애 시나리오만 설계에 포함된 경우, AI 서비스의 특성을 반영하지 못한 설계로 판단한다. 환각(Hallucination) 탐지 임계값 초과, 모델 버전 롤백, 컨텍스트 길이 초과 시 처리 등 AI 특유의 항목이 최소 1개 이상 포함되어야 한다.

4) 인터뷰와 문서 교차 검증: 설계 산출물에 구조도가 존재하더라도, 인터뷰 시 담당자가 구조도를 제대로 설명하지 못한다면 형식적으로만 작성된 산출물일 가능성이 높다. "이 구조도에서 추론 서버가 응답을 반환하지 못할 때 어느 컴포넌트가 탐지하고 어떤 이벤트가 발행됩니까?"와 같은 구체적인 질문을 통해 교차 검증을 수행한다.

5) '요구사항-설계 추적표'가 없는 경우: 추적표가 별도 산출물로 제출되지 않았더라도, {시스템아키텍처설계서} 내에 요구사항 번호 또는 코드를 참조하는 기술이 존재하는지를 확인한다. 완전히 추적이 불가한 경우에만 부적합으로 판정하고, 부분 추적이 가능한 경우 보완권고 처리한다.

AG_SD42-1-05 | CI/CD 배포 파이프라인 보안 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | CI/CD파이프라인

근거 | NIST AI RMF GOVERN · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 정의된 CI/CD 배포 파이프라인 보안 요구사항 및 소스/모델 병합 시 자동 보안 검사 파이프라인 요구사항이 {시스템아키텍처 설계서}에 빠짐없이 상세하고 구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 서비스의 CI/CD 파이프라인은 소스코드뿐만 아니라 학습된 모델 가중치(Model Weight), 프롬프트 템플릿, 학습 데이터셋 등이 함께 병합·배포되는 구조적 특수성을 가지며, 이 과정에서 모델 변조(Model Tampering), 악의적 프롬프트 삽입, 오염된 학습 데이터 혼입 등 AI 고유의 보안 위협이 발생할 수 있으므로 설계 단계에서 이러한 위협에 대한 방어 설계가 구체적으로 반영되었는지 검증한다.

2. 필요성

▶ AI 서비스의 CI/CD 파이프라인은 소스코드 커밋부터 모델 서빙까지 자동화된 경로를 통해 운영 환경에 직접 반영되므로, 파이프라인 자체의 보안이 곧 서비스 보안의 핵심 방어선이 된다. 이를 검증해야 하는 구체적인 필요성은 다음과 같다.

▶ AI 모델 공급망 공격(Supply Chain Attack) 방지: AI 모델은 오픈소스 사전훈련 모델(Pre-trained Model)을 기반으로 파인튜닝하는 경우가 많으며, 외부에서 가져온 모델 가중치에 백도어(Backdoor)가 삽입되어 있을 수 있다. 따라서 모델 무결성 검증 절차가 파이프라인에 포함되었는지 확인해야 한다.

▶ 자동 배포 경로의 보안 게이트(Security Gate) 확보: 수동 검토 없이 자동으로 운영 환경에 배포되는 CI/CD 구조에서, 보안 검사가 파이프라인 단계로 설계에 명시되지 않으면 취약한 코드나 변조된 모델이 그대로 배포될 수 있다.

▶ 소스/모델 병합(Merge) 시점의 보안 검증 필요: 전통적 CI/CD에서는 소스코드 병합 시 코드 리뷰와 정적 분석(SAST)을 수행하지만, AI 파이프라인에서는 모델 파일, 학습 설정, 프롬프트 구성 파일 등도 병합 대상이므로 이에 대한 추가적인 보안 검사 설계가 필요하다.

▶ 규제 및 감사 추적성 확보: TTA 신뢰할 수 있는 인공지능 개발안내서에서는 AI 시스템의 보안성 확보를 위한 개발 프로세스 보안을 권고하고 있으며, NIST AI RMF GOVERN 기능에서도 보안 관리 프로세스 수립을 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	커밋 단계 시크릿 스캔 절차 명시 여부	API 키·인증 토큰·모델 접근 키 하드코딩 방지를 위한 커밋 단계 시크릿 스캔 도구·탐지 규칙·탐지 시 조치(빌드 차단 등)가 'CI/CD 파이프라인 설계서'에 명시되었는지 확인
2	SAST·AI 특화 취약점 스캔 포함 여부	소스코드 정적 분석(SAST)·AI 프레임워크 특화 취약점 스캔의 도구·규칙셋·심각도별 실패 기준이 파이프라인 단계에 정의되었는지 {시스템아키텍처설계서}와 교차 확인
3	컨테이너 이미지 보안 스캔 포함 여부	컨테이너 이미지 취약점 검사·Base Image 검증 도구와 Critical/High 검출 시 빌드 차단 조치가 빌드 단계에 명시되었는지 확인
4	모델 아티팩트 서명·무결성 검증 설계 여부	AI 모델 아티팩트 서명 방식과 배포 전 무결성 검증 절차·검증 실패 시 배포 중단 조치가 설계되었는지 확인하고 서명 키 관리 방식을 면담으로 확인
5	배포 승인 게이트 설계 여부	자동 보안 검사 통과 후 승인 절차 등 배포 승인 게이트의 승인 주체·조건과 미통과 시 배포 차단이 설계에 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	시크릿 스캔, SAST 및 AI 특화 취약점 스캔, 컨테이너 이미지 스캔, 모델 아티팩트 서명·무결성 검증, 배포 승인 게이트가 파이프라인 단계별로 도구·규칙·실패 시 조치까지 구체적으로 설계에 명시된 경우	현행 유지
보완권고 (△)	보안 검사 단계가 파이프라인 설계에 포함되어 있으나 일부 단계(예: 모델 무결성 검증, 승인 게이트)가 누락되었거나, 검사 도구·기준·승인 주체가 추상적으로만 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	CI/CD 파이프라인에 자동 보안 검사 설계가 전혀 없거나, 소스/모델 병합 및 배포 경로에 보안 게이트 없이 운영 환경으로 직접 배포되도록 설계되어 요구사항이 반영되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 소스코드 파이프라인과 모델 파이프라인의 구분 확인: AI 서비스의 CI/CD는 소스코드 배포 경로와 모델 배포 경로가 분리되어 운영되는 경우가 많다. 설계서 검토 시 두 경로 모두에 보안 게이트가 설계되었는지 반드시 확인해야 한다. 한쪽에만 보안 검사가 설계된 경우 보완을 권고한다.

2) '보안 검사 수행'의 구체성 검증: 설계서에 '배포 전 보안 검사를 수행한다'라는 선언적 문구만 있고, 구체적인 검사 항목(어떤 도구로, 어떤 규칙으로, 실패 시 어떤 조치를)이 명시되지 않은 경우 부적합으로 판정할 수 있다. 측정 가능한 기준(예: 'Critical/High 취약점 0건 시 배포 승인')이 설계에 포함되었는지 확인한다.

3) AI 모델 특화 보안 위협에 대한 설계 반영 확인: 전통적인 SAST/DAST만으로는 AI 고유 위협(모델 변조, 적대적 입력, 프롬프트 인젝션 등)을 방어할 수 없다. 설계서에 AI 특화 보안 검사(모델 무결성 검증, 적대적 테스트, 프롬프트 검증 등)가 파이프라인 단계에 포함되었는지 별도로 확인해야 한다.

4) 요구사항 추적성 확보 여부 확인: {보안요구사항}에 정의된 CI/CD 보안 관련 요구사항이 {요구사항추적표}를 통해 {시스템아키텍처설계서}의 어느 설계 항목에 반영되었는지 추적 가능해야 한다. 추적이 불가능한 경우 요구사항 누락의 가능성이 높으므로 보완을 권고한다.

5) 파이프라인 설정의 코드화(Infrastructure as Code) 확인: 파이프라인 구성이 코드로 관리(예: Jenkinsfile, .gitlab-ci.yml, GitHub Actions YAML)되도록 설계되었는지 확인한다. 파이프라인 설정이 코드로 관리되어야 변경 이력 추적과 보안 검토가 가능하다.

AG_SD42-1-06 | 모델 공격-방어 설계점검

시스템구조·보안 | 설계 | 인공지능 모델 개발 | AI특화보안

근거 | OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 정의된 보안 요구사항이 {보안설계서}에 모델 공격(Evasion Attack, 적대적 사례 등) 방어 설계 및 적대적 입력 유입 시 입력 정제·거부 로직으로 빠짐없이 상세하고 구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 모델은 의도적으로 조작된 입력(Adversarial Example)에 의해 오분류·오동작을 유발할 수 있고, 생성형 AI의 경우 프롬프트 인젝션(Prompt Injection)이나 탈옥(Jailbreak) 공격을 통해 시스템 정책을 우회할 수 있다.

이러한 AI 고유 위협에 대한 방어 설계가 {보안설계서}에 구체적인 방어 아키텍처·로직·절차로 반영되었는지, 요구사항 추적이 완전한지를 확인한다.

2. 필요성

▶ AI 모델은 기존 소프트웨어와 달리 입력 데이터의 미세한 변형만으로도 출력 결과가 근본적으로 달라질 수 있어, 전통적인 입력 검증(Input Validation)만으로는 방어가 충분하지 않다. 설계 단계에서 모델 공격 방어 전략과 적대적 입력 처리 로직이 설계 산출물에 구체적으로 반영되지 않으면, 개발 과정에서 방어 로직이 누락되더라도 이를 검증할 기준이 없어 보안 공백이 그대로 운영 환경까지 이어진다.

▶ Evasion Attack 방어 설계 필수: 공격자가 입력 데이터를 교묘하게 변조하여 모델의 판단을 오도하는 Evasion Attack은 AI 고유 위협으로, 기존 웹 방화벽이나 네트워크 보안 장비로는 탐지가 불가능하다. {보안설계서}에 이에 대한 방어 계층(입력 검증 → 이상 탐지 → 모델 강건성 확보)이 설계되어야 한다.

▶ 프롬프트 인젝션·탈옥 방어 설계: 생성형 AI 서비스에서는 사용자 입력(프롬프트)을 통해 시스템 프롬프트를 우회하거나 금칙어 정책을 무력화하는 공격이 빈번하다. 이를 방어하기 위한 입력 정제(Sanitization)·거부(Rejection) 로직의 구체적인 설계가 필요하다.

▶ 요구사항 추적 완전성 확보: 요구정의 단계에서 식별된 모델 보안 요구사항이 설계 산출물에 1:1로 추적·반영되었는지 확인하여, 설계 과정에서의 누락 또는 변질을 방지해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	Evasion Attack 방어 전략 설계 여부	Adversarial Training·입력 변환·앙상블 방어 등 Evasion Attack 방어 기법명·적용 대상 모델·적용 임계값이 {보안설계서}에 구체적으로 반영되었는지 확인
2	적대적 사례 이상치 탐지 로직 설계 여부	적대적 사례 탐지를 위한 이상치 탐지(Anomaly Detection) 로직의 기법명·적용 지점(입력 전처리 단계)·탐지 임계값이 'AI 모델 보안 아키텍처 설계서'에 설계되었는지 확인
3	Model Extraction 방어 설계 여부	API 호출 제한 기준(호출 빈도·사용자 단위)과 출력 정보 최소화 범위(확률값·로짓 노출 제한) 등 Model Extraction 방어가 {보안설계서}에 수치로 정의되었는지 확인하고 사업자 면담으로 검증
4	Data Poisoning 방어 설계 포함 여부	학습 데이터 무결성 검증 방법(해시·서명)과 이상 데이터 탐지 기법·임계값 등 Data Poisoning 방어가 {보안설계서}에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	Evasion Attack 방어 전략, 적대적 입력 이상치 탐지 로직, Model Extraction 대응(API 호출 제한·출력 최소화), Data Poisoning 대응(학습 데이터 무결성 검증)이 {보안설계서}에 방어 기법명·적용 대상·임계값 등 구체적 설계로 반영된 경우	현행 유지
보완권고 (△)	모델 공격 방어 설계가 일부 반영되어 있으나 '적대적 입력을 방어한다' 수준의 선언적 기술에 그치거나, 이상 탐지 임계값·호출 제한 기준 등 정량적 기준과 일부 공격 유형에 대한 대응 설계가 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안설계서}에 AI 모델 공격 방어 설계가 전혀 없거나 기존 웹 보안 설계만 존재하여, 요구정의 단계의 모델 보안 요구사항이 설계 산출물에 반영되지 않아 추적이 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 고유 공격과 기존 공격의 구분: 기존 웹 보안 설계(SQL 인젝션, XSS 방어 등)와 AI 고유 공격(Evasion, 프롬프트 인젝션, 모델 추출 등) 방어 설계를 혼동하지 않도록 주의한다. {보안설계서}에

'AI 모델 보안'이라는 별도 섹션이 존재하는지 우선 확인하고, 기존 웹 보안 설계와 분리되어 기술되었는지 점검하는 것이 효과적이다.

2) 모호한 설계의 구체화 유도: '적대적 입력을 방어한다'와 같은 선언적 문구는 검증이 불가능하다. 'Evasion Attack 대응으로 Adversarial Training 적용, 입력 이미지에 대한 JPEG 압축 전처리 적용', '프롬프트 인젝션 방어를 위해 시스템 프롬프트 Delimiter 격리 적용, 금칙어 500건 이상 필터링' 등 측정 가능한 수준으로 구체화하도록 권고해야 한다.

3) 요구사항 추적표 대조 필수: {보안요구사항}에서 모델 공격 방어 관련 요구사항이 2건이었다면, {보안설계서}에도 최소 2건 이상의 대응 설계 항목이 있어야 한다. {요구사항추적표}에서 '설계 반영' 컬럼이 공란이거나 '추후 반영'으로 표기된 항목은 실질적 미반영으로 판단하고 보완을 권고한다.

4) 생성형 AI 특화 점검 강화: 생성형 AI 서비스의 경우 프롬프트 인젝션 방어가 핵심이다. 사업자 면담 시 '시스템 프롬프트는 어떻게 보호하는가?', '금칙어 목록은 얼마나 확보되었고 갱신 절차는?', '간접 프롬프트 인젝션(외부 문서를 통한 공격)에 대한 대응 설계는?'과 같은 구체적 질문을 통해 설계의 실질성을 검증한다.

5) 다계층 방어(Defense in Depth) 관점 확인: 단일 방어 기법만으로는 AI 모델 공격을 완전히 차단할 수 없으므로, {보안설계서}에 입력 검증 → 이상 탐지 → 모델 강건성 → 출력 검증으로 이어지는 다계층 방어 아키텍처가 설계되었는지 확인한다. 이는 NIST AI RMF의 'Secure and Resilient' 특성과도 부합한다.

AG_SD42-1-07 | AI서비스 관리자페이지 접근통제 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 인증/인가/접근통제

근거 | NIST AI RMF GOVERN·MANAGE · 「전자정부법」 · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 정의된 관리자 페이지 접근 통제 및 2FA/MFA 다중 인증 요구사항이 {보안설계서}에 구체적이고 완전하게 반영되었는지를 제3자적 관점에서 검토하는 것을 목적으로 한다.

특히, 관리자 권한으로 AI 추론 서버·학습 파이프라인·프롬프트 설정 등 핵심 기능에 접근하는 경우, 단순 ID/패스워드 인증만으로는 계정 탈취 시 시스템 전체가 위험에 노출될 수 있으므로, 다중 인증 수단(2FA/MFA)과 접근 IP 통제의 설계 반영 여부를 집중 점검한다.

2. 필요성

▶ AI 서비스 관리자 페이지는 모델 파라미터 변경, 프롬프트 정책 수정, 사용자 데이터 조회 등 일반 사용자 인터페이스와 비교할 수 없는 높은 권한을 보유한다. 설계 단계에서 다중 인증 및 접근 IP 통제가 설계서에 명시되지 않으면, 이후 구현 단계에서 구현 누락이 발생하더라도 설계 위반 여부를 판별하는 기준이 없어 감리 지적 및 시정조치의 근거를 확보하기 어렵다.

▶ 관리자 계정 탈취 위험: AI 관리자 계정 1개 탈취 시 모델 전체를 무력화하거나 백도어 삽입이 가능하므로, 단일 패스워드만으로는 방어가 불충분하며 2FA/MFA 적용이 필수적이다.

▶ 설계→구현 추적성 확보: 설계서에 2FA/MFA 적용 범위, OTP 방식, 인증 수단 등이 구체적으로 기술되어야 구현 및 요건 충족 여부를 객관적으로 검증할 수 있다.

▶ IP 통제 미반영 위험: 관리자 페이지에 대한 접근 IP 허용 목록(Whitelist) 또는 차단 기준이 설계서에 없으면, 개발 완료 후 퇴직자·외부인의 원격 접근이 무방비로 허용될 수 있다.

▶ 공공기관 보안 의무 이행: 행정안전부 전자정부 보안 지침 및 NIST AI RMF(GOVERN·MANAGE)는 AI 시스템 관리자 접근에 대한 강화 인증 조치를 요구하며, 설계 단계 문서화를 의무화한다.

3. 검토방법

No	점검 항목	확인 방법
1	관리자 페이지 접근 통제 정책 반영 여부	{보안설계서}에서 관리자 페이지의 접근 주체·권한 등급·인증 절차를 포함한 접근 통제 정책이 명시되었는지 확인

No	점검 항목	확인 방법
2	2FA/MFA 적용 대상 및 방식의 구체성	{보안설계서}에서 다중 인증 적용 대상(계정 유형·기능 범위), 인증 수단(OTP·인증서·생체 등), 예외 처리 절차가 구체적으로 기술되었는지 확인
3	관리자 접근 IP 통제 기준	설계서에 접근 허용 IP 목록(Whitelist) 또는 차단 기준이 포함되었는지 확인하고, 원격근무·유지보수 시 예외 승인 절차를 사업자 면담으로 확인
4	관리자 계정 보안 정책(세션·잠금)	설계서에서 세션 타임아웃 시간, 로그인 실패 잠금 임계값, 비밀번호 정책이 수치로 명시되었는지 확인
5	요구사항의 설계 반영 추적성	{시스템요구사항}의 인증 관련 요건 목록과 {보안설계서}를 대조하여 누락 없이 추적·반영되었는지 확인
6	관리자·일반 사용자 페이지 논리적 분리 설계	아키텍처 설계서에서 관리자 페이지가 일반 사용자 페이지와 URL·네트워크 구간·인증 체계 측면에서 분리 설계되었는지 확인
7	인증 실패 로그 및 관리자 접근 감사 추적	설계서에 인증 실패 로그 기록 항목과 관리자 접근 감사 추적(Audit Trail) 요건이 반영되었는지 확인하고, 로그 보존 기간·열람 권한을 면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{보안설계서}에 관리자 페이지 접근 통제, 2FA/MFA 인증 수단·적용 범위, IP Whitelist 통제 기준, 세션 타임아웃·계정 잠금 정책이 구체적으로 기술되고, 관리자 페이지가 일반 사용자 페이지와 URL·네트워크 구간·인증 체계 측면에서 분리 설계되며, 인증 실패 로그 기록과 관리자 접근 감사 추적(Audit Trail) 요건이 반영되고, {시스템요구사항}의 인증 요건이 누락 없이 추적·반영된 경우	현행 유지
보완권고 (△)	접근 통제와 다중 인증 설계가 반영되어 있으나 인증 수단 유형(OTP 방식 등), 허용 IP 범위, 세션·잠금 정책 등 일부 항목이 추상적으로 기술되었거나 요구사항 추적표에 인증 요건이 일부 누락된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안설계서}에 2FA/MFA 및 관리자 페이지 접근 통제 설계가 전혀 없거나 일반 사용자와 동일한 인증 방식만 기술되어, 요구정의 단계의 인증 요구사항이 설계서에 반영되지 않아 추적이 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 인증 설계의 수치화·구체화 유도: "2FA를 적용한다"와 같은 선언적 문구는 분쟁의 소지가 있다. 감리원은 '인증 수단 유형(TOTP/SMS/FIDO2 중 선택)', '1차 인증 실패 N회 시 계정 잠금', '세션 만료 시간 M분' 등 측정 가능한 형태로 설계서에 기술하도록 사업자에게 권고해야 한다.

2) 요구사항-설계 추적표({과업대비표}) 활용: {시스템요구사항}의 인증 관련 요건이 {보안설계서}에 1:1 매핑되어 있는지를 {과업대비표}를 통해 확인한다. 추적표에 인증 요건이 누락된 경우 이를 보완권고 또는 부적합 지적 근거로 활용할 수 있다.

3) IP 통제 설계의 우회 경로 확인: IP Whitelist 정책이 설계서에 있더라도 'VPN 경유 후 접속' 또는 '관리용 Jump Server' 등 우회 접근 경로에 대한 별도 통제 방안이 없으면 실효성이 낮다. 아키텍처 설계서와 함께 교차 검토하도록 한다.

4) AI 특화 관리자 기능 범위 확인: 일반 IT 시스템 관리자와 달리 AI 시스템 관리자는 모델 가중치 교체, 프롬프트 시스템 변경, RAG 데이터 삭제 등 AI 고유 기능에 대한 접근 권한을 가진다. 이러한 AI 특화 기능이 접근 통제 설계 대상에 포함되어 있는지 별도로 확인해야 한다.

5) 설계 단계 점검의 한계 인식: 이 단계에서는 실제 구현 여부가 아닌 '설계서에의 반영 여부'를 점검한다. 2FA 라이브러리 선정, 실제 OTP 발급 서버 구성 등은 구현 단계 감리 시 점검 사항이며, 설계 단계에서는 이러한 사항들이 설계서에 요건으로 명시되어 있는지를 확인하는 것이 적절한 범위이다.

AG_SD42-1-08 | API 접근 인증-인가 통제설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 인증/인가/접근통제

근거 | OWASP LLM Top 10 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI 보안 가이드북」(국가정보원, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 식별된 AI 추론 API 접근 인증·인가 통제 보안요구사항이 {보안설계서}에 빠짐없이, 상세하고 구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 추론 API는 생성형 AI 모델의 핵심 접점(Endpoint)으로서, 비인가 호출 시 모델 탈취(Model Extraction), 프롬프트 인젝션(Prompt Injection), 과도한 토큰 소모 등 AI 고유의 보안 위협에 직접 노출된다.

이에 따라 OAuth 2.0, JWT(JSON Web Token), API Gateway 등 인증·인가 메커니즘이 보안설계서에 구체적인 설계 수준으로 명시되었는지, 요구사항 대비 누락·모호함이 없는지를 검증한다.

2. 필요성

▶ AI 추론 API는 학습된 모델에 직접 접근할 수 있는 경로이므로, 인증·인가 통제가 미흡할 경우 AI 고유의 심각한 보안 위험이 발생한다. 설계 단계에서 이를 구체화하지 않으면 구현 단계에서의 보안 공백이 불가피하며, 다음과 같은 구체적 필요성이 있다.

▶ 모델 탈취 방지: 비인가 사용자가 AI 추론 API를 대량 호출하여 모델 동작을 복제(Model Extraction Attack)하는 공격을 차단하기 위해, 호출자 식별·인가·호출량 제한이 설계 수준에서 정의되어야 한다.

▶ 권한 남용 및 횡적 이동 차단: AI 서비스의 역할별 API 접근 범위(예: 관리자 vs 일반 사용자 vs 외부 연동 시스템)를 세분화하고, 최소 권한 원칙(Least Privilege)에 따라 인가 정책을 설계해야 한다.

▶ 토큰 탈취·재사용 공격 방어: JWT 서명 알고리즘, 만료 시간(exp), 갱신 정책(Refresh Token Rotation) 등이 설계서에 구체적으로 명시되지 않으면, 토큰 탈취 시 무제한 API 접근이 가능해질 수 있다.

▶ 감사 추적성 확보: API Gateway를 통한 중앙 집중 인증·로깅 설계가 누락되면, AI 추론 호출에 대한 감사 추적(Audit Trail)이 불가능하여 침해 사고 발생 시 원인 분석이 어렵다.

3. 검토방법

No	점검 항목	확인 방법
1	OAuth 2.0 시나리오별 인가 흐름 설계 여부	사용자 직접 호출·서버 간 호출·외부 시스템 연동 등 AI API 호출 시나리오별로 OAuth 2.0 인가 흐름(Grant Type)이 구분되어 {보안설계서}에 설계되었는지 확인
2	JWT 토큰 설계 명시 여부	서명 알고리즘·만료 시간·Claim 구조·Refresh Token 정책 등 JWT 토큰 설계가 {보안설계서}에 구체적으로 명시되었는지 확인
3	API Gateway 중앙 인증·인가 아키텍처 설계 여부	API Gateway를 통한 중앙 집중 인증·인가 아키텍처가 {아키텍처설계서}·{보안설계서}에 설계되었는지 교차 검토
4	엔드포인트별 접근 권한 매트릭스 설계 여부	AI 추론 API 엔드포인트별 접근 권한(Role/Scope) 매트릭스가 {보안설계서}·'API 명세서'에 설계되었는지 확인
5	토큰 수명주기 프로세스 설계 여부	토큰 발급·검증·갱신·폐기(Revocation) 프로세스가 {보안설계서}에 상세 설계되었는지 확인하고 면담으로 보완
6	비인가 접근 에러 응답·로깅 정책 설계 여부	비인가 접근 시 에러 응답(HTTP 401/403)·로깅 정책이 {보안설계서}·'인터페이스설계서'에 설계되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	시나리오별 OAuth 2.0 Grant Type, JWT 토큰 구조·만료·갱신 정책, API Gateway 중앙 인증 아키텍처, 엔드포인트별 Role/Scope 매트릭스, 토큰 생명주기 및 비인가 접근 로깅 정책이 {보안설계서}에 구현 가능한 수준으로 구체적으로 설계된 경우	현행 유지
보완권고 (△)	인증·인가 메커니즘이 설계에 반영되어 있으나 'OAuth 2.0 적용' 등 선언적 기술에 그치거나, 토큰 폐기 절차·접근 권한 매트릭스·로깅 정책 등 일부 항목이 누락되거나 추상적으로 기술된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 추론 API에 대한 인증·인가 설계가 {보안설계서}에 없거나, 보안요구사항 대비 핵심 통제(인증 방식, 인가 정책, Gateway 통제)가 반영되지 않아 비인가 호출 차단이 불가능한 설계인 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선언적 기술과 설계 수준의 구분: {보안설계서}에 'OAuth 2.0을 적용한다', 'JWT 토큰을 사용한다'는 선언만 있는 경우가 빈번하다. 시나리오별 흐름도(Sequence Diagram), 토큰 구조 정의서,

인가 정책 매트릭스 등 구현 가능한 수준의 상세 설계가 필요하므로, 선언적 기술만으로는 '보완권고' 이상을 판정해야 한다.

2) AI 특화 인증 요건 식별: 일반 웹 서비스의 OAuth/JWT 설계와 AI 추론 API의 설계는 차이가 있다. AI API는 단일 호출로 대량의 연산 자원(GPU)을 소모하고, 반복 호출로 모델 동작을 추론할 수 있으므로, 호출량 제한(Rate Limiting), 모델별 접근 제한(Scope), 이상 호출 탐지 등 AI 특화 보안 통제가 설계에 포함되었는지 확인해야 한다.

3) 요구사항 추적의 양방향 확인: {요구사항추적표}를 검토할 때, 보안요구사항 → 설계 산출물 방향(정방향 추적)뿐만 아니라, 설계 산출물 → 보안요구사항 방향(역방향 추적)도 확인하여, 요구사항에 없는 불필요한 설계나 요구사항 대비 과소/과잉 설계를 식별해야 한다.

4) API Gateway 우회 경로 확인: API Gateway를 설계했더라도, AI 추론 서비스에 직접 접근 가능한 우회 경로(예: 내부 네트워크 직접 호출, 디버그 포트 노출 등)가 차단되도록 설계되었는지 확인해야 한다. 아키텍처설계서의 네트워크 구성도에서 우회 가능 경로를 확인하고, 방화벽·ACL 정책과의 일관성을 검토한다.

5) 토큰 보안 강도 수치화 유도: JWT 서명 알고리즘으로 HS256(대칭키)만 적용한 경우 키 유출 시 모든 토큰이 위조 가능하므로, RS256/ES256 이상의 비대칭 알고리즘 적용을 권고해야 한다. Access Token 만료 시간이 24시간 이상으로 설계된 경우, AI API의 특성을 고려하여 30분~1시간 이내로 단축할 것을 권고한다.

AG_SD42-1-09 | 모델저장소 접근통제점검

시스템구조·보안 | 설계 | 시스템 구현 | 인증/인가/접근통제

근거 | NIST AI RMF MANAGE · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에 정의된 보안요구사항이 {보안설계서}를 비롯한 설계 산출물에 빠짐없이 상세하고 구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 모델 저장소(Model Registry)에 대한 접근 통제 정책이 역할 기반 접근 제어(RBAC), 최소 권한 원칙 등 구체적 기술 설계로 전환되었는지, 인가된 CI/CD 파이프라인 이외의 경로를 통한 모델 파일 직접 접근을 차단하는 설계가 명확히 포함되었는지를 확인한다.

모델 저장소 접근 통제가 설계 단계에서 구체화되지 않으면 비인가 사용자의 모델 탈취, 위·변조된 모델의 배포, 학습 데이터 유출 등 AI 고유의 보안 위협에 노출될 수 있으므로, 설계 완전성 확인이 본 가이드라인의 핵심이다.

2. 필요성

▶ AI 모델 저장소(Model Registry)는 학습이 완료된 모델 가중치(Weight), 하이퍼파라미터, 메타데이터, 버전 정보 등을 중앙에서 관리하는 핵심 자산 저장소이다. 이 저장소에 대한 접근 통제가 설계 단계에서 구체화되지 않으면 다음과 같은 AI 고유의 보안 위협이 현실화될 수 있다.

▶ 모델 탈취(Model Extraction) 위험: 비인가 사용자가 학습된 모델 파일에 직접 접근하여 모델 가중치를 복사·유출할 경우, 조직의 핵심 지적 재산이 외부로 유출되며 적대적 공격(Adversarial Attack)의 기초 자료로 악용될 수 있다.

▶ 모델 위·변조(Model Tampering) 위험: 인가되지 않은 경로로 모델 파일을 교체하거나 수정할 경우, 백도어(Backdoor)가 삽입된 모델이 운영 환경에 배포되어 의도하지 않은 편향된 결과나 보안 사고를 유발할 수 있다.

▶ 공급망 공격(Supply Chain Attack) 위험: CI/CD 파이프라인 외부에서 모델을 직접 업로드·배포할 수 있는 경로가 존재하면, 검증되지 않은 모델이 운영 환경에 투입되어 서비스 안정성과 신뢰성을 훼손할 수 있다.

▶ 규정 준수 위험: NIST AI RMF는 AI 시스템의 보안과 복원력(Secure and Resilient)을 핵심 신뢰 특성으로 요구하며, 모델 자산의 기밀성·무결성·가용성 보호를 위한 구체적 설계가 필요하다.

3. 검토방법

No	점검 항목	확인 방법
1	모델 저장소 RBAC 접근 통제 설계 여부	모델 저장소 접근 통제 정책이 {보안설계서}에 RBAC 기반으로 상세 설계되었는지 확인
2	모델 저장소 인증·인가 메커니즘 설계 여부	모델 저장소 인증·인가 메커니즘이 {보안설계서}·'AI 인프라 아키텍처 설계서'에 구체적으로 반영되었는지 확인
3	인가된 CI/CD만 쓰기 가능 설계 여부	인가된 CI/CD 파이프라인만 모델 저장소에 쓰기(Write) 가능하도록 'CI/CD 파이프라인 설계서'에 설계되었는지 확인하고 사업자 면담으로 보완
4	모델 파일 무결성 검증 설계 여부	모델 파일 무결성 검증(서명·해시) 설계가 {보안설계서}에 반영되었는지 확인
5	모델 상태 전이 승인 워크플로우 설계 여부	모델 상태 전이(Stage Transition) 승인 워크플로우가 '모델 관리 정책서'에 설계되었는지 확인
6	모델 저장소 접근 감사 로그 설계 여부	모델 저장소 접근 감사 로그(Audit Log) 설계가 {보안설계서}에 반영되었는지 확인
7	모델 저장소 보안요구사항 매핑 여부	모델 저장소 관련 보안요구사항이 {요구사항추적표}에서 설계 항목으로 완전히 매핑되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	역할별 RBAC 매트릭스, 인증·인가 메커니즘, CI/CD 파이프라인 전용 쓰기 권한, 모델 서명·해시 무결성 검증, 상태 전이 승인 워크플로우, 접근 감사 로그가 설계 산출물에 구체적으로 반영되고 보안요구사항이 추적표에 완전히 매핑된 경우	현행 유지
보완권고 (△)	접근 통제 설계가 반영되어 있으나 RBAC 역할 구분·무결성 검증·감사 로그 등 일부 항목이 선언적 수준에 그치거나, 파이프라인 외 직접 접근 경로에 대한 차단·예외 통제 설계가 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	모델 저장소 접근 통제 설계가 {보안설계서}에 없거나, 개발자 누구나 모델을 직접 업로드·수정할 수 있는 구조로 설계되어 보안요구사항이 반영되지 않고 추적도 불가능한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 설계 구체성 검증: {보안설계서}에 "모델 저장소에 접근 통제를 적용한다"와 같은 선언적 문구만 있고 RBAC 매트릭스, 인증 방식, 네트워크 정책 등 구체적 설계가 없으면 미흡으로 판단해야 한다. "어떤 역할이, 어떤 조건에서, 어떤 작업을 할 수 있는가"가 명확히 정의되어야 한다.

2) 요구사항 추적성 중점 확인: 도출된 모델 저장소 관련 보안요구사항이 설계서의 어느 절에 매핑되었는지를 {요구사항추적표}로 반드시 확인해야 한다. 추적이 불가능한 요구사항이 있으면 설계 누락 가능성이 높다.

3) AI 고유 위협과 일반 IT 위협의 구분: 일반적인 파일 서버 접근 통제나 네트워크 보안은 일반 IT 감리 영역에서 다루므로, 본 가이드라인에서는 모델 저장소 특유의 위협(모델 탈취, 모델 위변조, 파이프라인 우회 배포 등)에 대한 설계 반영 여부에 집중해야 한다.

4) 예외 경로 존재 여부 반드시 확인: 긴급 배포(핫픽스) 등을 위한 예외 경로가 설계에 포함된 경우, 해당 예외의 승인 절차, 이중 인증, 사후 감사 기록 설계가 함께 포함되었는지 확인해야 한다. 예외 경로가 통제 없이 허용되면 접근 차단 설계의 실효성이 크게 감소한다.

5) 면담 시 확인 포인트: 사업자 면담 시 "모델 저장소에 개발자가 직접 모델을 업로드할 수 있는 경로가 존재하는가?", "모델 배포 시 서명 검증을 생략할 수 있는 예외 조건이 있는가?" 등 구체적 질문을 통해 설계의 실효성을 확인하는 것이 중요하다.

AG_SD42-1-10 | HSM 연동 키관리 아키텍처 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 암호화

근거 | KCMVP · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 시 도출된 보안요구사항이 {보안설계서}에 빠짐없이 상세하고 구체적으로 반영되었는지를 하드웨어 보안 모듈(HSM) 연동 설계 및 API 키·암호화 키 관리 아키텍처 설계 관점에서 제3자적 입장으로 점검하는 것을 목적으로 한다.

AI 시스템에서는 모델 가중치 서명 키, 추론(Inference) API 키, 외부 LLM 연동 API 키, 학습 데이터 암호화 키 등 기존 정보시스템 대비 다수의 이종 암호화 키가 존재하며, 이러한 키가 유출될 경우 모델 탈취, 무단 추론 API 호출, 학습 데이터 복원 등 AI 고유의 보안 위협으로 직결되므로 설계 단계에서 HSM/KMS 기반의 체계적인 키 관리 아키텍처가 수립되었는지 확인할 필요가 있다.

2. 필요성

▶ AI 시스템은 전통적 웹/WAS 시스템과 달리 모델 자산 자체가 핵심 지적재산이며, 이를 보호하는 암호화 키의 관리 수준이 전체 시스템 보안의 근간이 된다. 특히 생성형 AI 서비스에서는 AI 특유의 키 관리 위험이 존재한다.

▶ 모델 가중치(Weight) 탈취 위험: AI 모델 파일에 대한 서명·검증 키가 부재하면, 공급망 공격(Supply-Chain Attack)을 통해 변조된 모델이 배포될 수 있다.

▶ 추론 API 키 유출 위험: 추론 API 키가 소스코드에 하드코딩되거나 평문으로 저장되면, 대규모 무단 추론 호출(모델 추출 공격)에 노출된다.

▶ 외부 LLM API 키 노출: GPT, Claude 등 외부 LLM API 키가 환경변수 분리 없이 관리되면, Git 저장소 유출 등으로 과도한 과금 및 정보 유출이 발생할 수 있다.

▶ 학습 데이터 암호화 키 단일 관리 위험: DEK(Data Encryption Key)와 KEK(Key Encryption Key)가 분리되지 않으면, 단일 키 유출 시 전체 학습 데이터가 복호화되는 위험이 있다.

3. 검토방법

No	점검 항목	확인 방법
1	HSM 연동 인터페이스 설계	{보안설계서}에서 HSM 연동 방식(PKCS#11·KMIP 등), 연동 대상 키, 장애 시 대체 경로가 구체적으로 설계되었는지 확인

No	점검 항목	확인 방법
2	AI 모델 서명·검증 키 관리 설계	모델 파일의 서명 생성·검증 키의 보관 위치와 배포 시 무결성 검증 절차가 설계되어 변조 모델 배포를 차단하는지 확인
3	AI 추론 API 키 발급·순환 설계	추론 API 키의 발급·순환 주기·폐기 절차가 설계되었는지 확인하고, 소스코드 하드코딩 금지와 비밀 저장소 연동 여부를 면담으로 확인
4	LLM API 키 환경변수 분리 설계	외부 LLM API 키가 환경변수 또는 비밀 저장소로 분리되어 저장소 커밋 시 노출되지 않도록 설계되었는지 확인
5	학습 데이터 암호화 키 분리 보관 설계	DEK와 KEK가 분리 설계되어 단일 키 유출 시 전체 학습 데이터가 복호화되지 않도록 계층화되었는지 확인
6	키 접근 권한 및 감사 로그 설계	키 사용·조회 권한이 역할별로 분리되고, 키 접근 이력이 감사 로그로 기록되도록 설계되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	HSM 연동 인터페이스, API 키 수명주기, 모델 서명 키, 학습데이터 암호화 키 관리 아키텍처가 {보안설계서}에 AI 특화 요건을 포함하여 빠짐없이 상세히 설계되어 있으며, 키 접근 권한 통제 및 감사 로그 설계까지 구체적으로 반영된 경우	현행 유지
보완권고 (△)	HSM 연동 설계 또는 API 키 관리 설계는 존재하나, AI 모델 서명 키 관리, LLM API 키 비밀 저장소 연동, 키 순환 주기 등 일부 AI 특화 항목이 누락되었거나 구체성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	HSM 연동 설계가 부재하거나, API 키·암호화 키 관리 아키텍처가 {보안설계서}에 전혀 반영되지 않았으며, 키 수명주기(생성·순환·폐기) 설계가 확인되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 키 유형의 식별과 분류: 감리 시 {보안설계서}에 기재된 키의 유형을 'AI 모델 서명 키', '추론 API 키', '외부 LLM API 키', '학습 데이터 암호화 키'로 분류하여 누락 여부를 체크리스트 방식으로 확인하면 효과적이다. 특히 외부 LLM API 키는 기존 보안설계서에서 누락되기 쉬운 항목이므로 별도 확인이 필요하다.

2) 하드코딩 여부의 설계적 검증: 소스코드를 직접 검토할 수는 없으나, {보안설계서}에 '키의 런타임 주입 방식(환경변수/Vault/KMS)'이 명시되어 있는지, '소스코드 내 비밀정보 포함 금지' 원칙이 기술되어 있는지를 통해 간접 확인할 수 있다.

3) HSM vs KMS 구분 확인: 공공기관의 경우 국가정보원 검증필 암호모듈(KCMVP) 적용이 요구될 수 있으며, 이 경우 클라우드 KMS만으로는 요건 충족이 불가능할 수 있다. {보안설계서}에 KCMVP 적용 대상 키와 클라우드 KMS 사용 가능 키가 구분되어 있는지 확인한다.

4) 요구사항 추적성(RTM) 확인: {요구사항정의서}의 보안요구사항 중 HSM·키 관리 관련 항목이 {보안설계서}로 1:1 추적 가능한지 확인한다. 설계 단계 감리에서는 요구사항 대비 설계 반영의 완전성(Completeness)이 핵심 점검 포인트이다.

5) 모호한 설계 문구의 구체화 유도: '적절한 키 관리를 수행한다', 'HSM을 연동한다'와 같은 모호한 문구는 구현·검사 단계에서 분쟁의 원인이 된다. '키 순환 주기 90일 이내', 'PKCS#11 v2.40 이상 인터페이스 적용', 'DEK/KEK 봉투 암호화 구조 적용' 등 측정·검증 가능한 설계 기준으로 구체화하도록 유도해야 한다.

AG_SD42-1-11 | 구간암호화 TLS 인증서 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 암호화

근거 | 「개인정보 보호법」 · KCMVP · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 수립된 보안요구사항(구간 암호화, HTTPS 적용, 인증서 관리 등)이 {보안설계서}에 빠짐없이 상세하고 구체적으로 반영되었는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 서비스는 사용자 프롬프트, 추론 결과, 학습 데이터, 모델 파라미터 등 민감한 정보가 다수의 구간(클라이언트↔API Gateway↔추론 서버↔학습 서버↔외부 연계 API)을 통해 전송되며, 각 구간의 통신 보안이 설계 단계에서 구체적으로 명세되지 않으면 운영 시 데이터 도청, 중간자 공격(MITM), 인증서 위변조 등의 위험에 노출될 수 있다.

따라서 본 가이드라인은 이러한 AI 고유의 다중 구간 통신 보안 설계 적정성에 초점을 맞춘다.

2. 필요성

▶ AI 서비스는 일반 웹 서비스와 달리 추론 API 호출, 모델 학습 파이프라인 통신, 벡터 DB 연계, 외부 LLM API 호출 등 복수의 내부·외부 통신 경로가 존재한다. 이러한 통신 구간별로 TLS 적용 여부와 인증서 관리 방안이 설계서에 누락될 경우, 보안 사각지대가 발생하여 AI 모델 탈취, 프롬프트/응답 데이터 유출 등의 침해 위험이 커진다.

▶ AI 추론 API 통신 보호: 사용자 입력(프롬프트)과 AI 응답은 개인정보, 업무 기밀을 포함할 수 있으므로, 클라이언트-서버 간 TLS 1.2 이상 적용이 설계서에 명시되어야 한다.

▶ 내부 마이크로서비스 간 통신 보호: AI 서비스의 추론 엔진, 전·후처리 서비스, 벡터 DB, 캐시 서버 등 내부 구간에서도 mTLS(상호 TLS 인증) 또는 서비스 메시(Service Mesh) 기반 암호화가 설계되어야 AI 모델 파라미터와 중간 결과의 도청·변조를 방지할 수 있다.

▶ 외부 LLM/API 연계 구간 보호: 외부 생성형 AI API(예: GPT, Claude 등) 호출 시 API 키와 프롬프트 데이터가 전송되므로, 해당 구간의 TLS 적용과 인증서 검증(Certificate Pinning 등) 설계가 필수적이다.

▶ 인증서 수명주기 관리: AI 서비스의 24/7 무중단 운영 특성상, 인증서 만료로 인한 서비스 중단 방지를 위해 자동 갱신(ACME 프로토콜 등) 및 만료 사전 경고 설계가 보안설계서에 반영되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	통신 구간별 TLS 적용 대상 목록화	{보안설계서}에서 클라이언트↔API GW, API GW↔추론서버, 추론서버↔벡터 DB, 외부 LLM API 등 통신 구간이 목록으로 식별되고 각 구간의 TLS 적용 여부가 명시되었는지 확인
2	TLS 최소 버전 및 암호 스위트 명시	설계서에서 TLS 프로토콜 최소 버전(1.2 이상)과 권장 암호 스위트(Cipher Suite)가 구체적으로 지정되었는지 확인
3	내부 마이크로서비스 간 암호화 통신 설계	추론 엔진·전후처리·벡터 DB·캐시 등 내부 구간에 mTLS 또는 서비스 메시 기반 암호화가 설계되었는지 확인하고, 미적용 구간의 사유를 면담으로 확인
4	외부 생성형 AI API 인증서 검증 방안	외부 LLM API 연계 구간에 대해 CA 검증 또는 Certificate Pinning 등 인증서 검증 방안이 설계에 반영되었는지 확인
5	HTTPS 전 구간 강제 적용 설계	설계서에 HTTP→HTTPS 리다이렉션과 HSTS 헤더 적용이 포함되었는지 확인
6	인증서 수명주기 관리 절차	설계서에서 인증서 발급·갱신·폐기 절차와 자동 갱신(ACME 등), 만료 사전 알림 기준이 구체적으로 명시되었는지 확인
7	학습 파이프라인 구간 TLS 적용 설계	학습 데이터 전송 및 모델 배포 구간의 TLS 적용 설계가 반영되었는지 확인하고, 내부 폐쇄망 구간의 예외 여부를 면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	통신 구간 목록별 TLS 적용 대상, 프로토콜 최소 버전·암호 스위트, 내부 구간 mTLS, 외부 API 인증서 검증, HSTS 적용, 인증서 수명주기 관리 절차가 {보안설계서}에 구체적으로 명시된 경우	현행 유지
보완권고 (△)	주요 외부 구간의 TLS 적용은 설계되어 있으나 내부 마이크로서비스 구간, 학습 파이프라인, 외부 LLM API 연계 구간 등 일부 구간이 누락되었거나, '전 구간 TLS 적용' 등 선언적 문구만 있고 버전·암호 스위트·인증서 관리 방안이 구체화되지 않은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	구간 암호화 및 인증서 관리 설계가 {보안설계서}에 반영되지 않았거나, 민감 데이터가 전송되는 주요 구간이 평문 통신으로 설계되어 보안요구사항을 충족하지 못하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 고유 통신 구간의 누락 여부에 집중: 일반 웹 서비스의 HTTPS 적용은 대부분 설계에 반영되지만, AI 서비스 고유의 통신 구간(추론 엔진↔벡터 DB, 학습 파이프라인, 외부 LLM API 연계 등)은 빈번하게 누락된다. 감리원은 {아키텍처설계서}의 통신 흐름도에서 모든 구간을 식별한 후, {보안설계서}와 대조하는 방식으로 누락 구간을 찾아야 한다.

2) '전 구간 TLS 적용'이라는 선언적 문구 경계: {보안설계서}에 '모든 통신 구간에 TLS를 적용한다'는 포괄적 문구만 있고, 구간별 구체적 설계(TLS 버전, 암호 스위트, 인증서 유형 등)가 없으면 실제 구현 시 누락 또는 오류가 발생할 수 있으므로 보완 권고 대상이다.

3) 인증서 자동 갱신 설계 확인: AI 서비스의 24시간 무중단 운영 특성상, 인증서 수동 갱신은 만료 사고의 주된 원인이 된다. Let's Encrypt + ACME 프로토콜, 또는 기관 내 사설 CA 자동 갱신 프로세스 등 구체적인 자동화 방안이 설계서에 포함되었는지 확인해야 한다.

4) 내부 통신 암호화의 'AI 특수성' 확인: 일반 IT시스템은 내부 통신 암호화가 선택적일 수 있으나, AI 시스템은 모델 파라미터, 임베딩 벡터, 프롬프트 원문 등 고민감 데이터가 내부 구간에서 전송되므로, 내부 구간 암호화(mTLS, 서비스 메시 등)의 설계 반영이 필수적이다.

5) 외부 LLM API 연계 시 인증서 핀닝(Certificate Pinning) 확인: 외부 생성형 AI API를 호출하는 경우, 단순 TLS 적용 외에 인증서 핀닝 또는 CA 고정을 통해 중간자 공격을 방지하는 설계가 반영되었는지 확인한다. 이는 AI 서비스에서 프롬프트와 API 키가 외부로 전송되는 만큼 특히 중요하다.

AG_SD42-1-12 | DDoS 공격방어 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 시큐어코딩

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 도출된 서비스 거부(DDoS) 공격 방어 및 비정상 트래픽 차단 관련 보안요구사항이 {보안설계서}에 구체적이고 상세하게 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 AI 추론 API 엔드포인트는 일반 웹 서비스 대비 GPU 연산 자원을 대량으로 소모하는 특성이 있어, 소량의 악의적 요청만으로도 서비스 가용성이 심각하게 저하될 수 있다.

이에 따라 방화벽·WAF·L7 스위치 등을 활용한 DDoS 방어 설계가 AI 서비스의 특성을 고려하여 구체적으로 수립되었는지, 해당 설계가 요구사항 분석 내용과 정합성을 갖추고 있는지를 검토한다.

2. 필요성

▶ 생성형 AI 서비스는 추론 요청 1건당 일반 웹 트랜잭션 대비 수십~수백 배의 연산 자원(GPU, 메모리)을 소모하므로, 전통적인 대용량 트래픽 기반 DDoS 공격뿐 아니라 소량의 정교한 요청으로도 AI 추론 파이프라인의 가용성을 마비시킬 수 있는 새로운 유형의 서비스 거부 위협이 존재한다.

▶ AI 추론 자원의 고비용 특성: GPU 기반 추론은 단일 요청의 처리 비용이 높으므로, 공격자가 대량의 복잡한 프롬프트를 반복 전송하면 정상 사용자의 서비스 응답 지연 또는 불가 상태가 발생할 수 있다. 이를 방어하는 구체적인 메커니즘이 보안설계서에 명시되어야 한다.

▶ AI 특화 공격 벡터의 다양성: AI 서비스에 대한 DDoS 공격은 단순 트래픽 증폭 외에도 프롬프트 인젝션을 활용한 연산 자원 고갈(Compute-Exhaustion Attack), API 엔드포인트 대상 L7 계층 공격 등 AI 고유의 공격 벡터를 포함하므로, 이를 고려한 설계가 필요하다.

▶ 요구사항 추적성 확보: 요구정의 단계에서 도출된 {보안요구사항}의 DDoS 방어 관련 항목이 설계 산출물에 누락 없이 반영되었는지 확인함으로써, 설계-구현 단계 간 추적성을 확보하고 향후 검사 단계의 기준을 명확히 할 필요가 있다.

▶ 일반적인 HTTP 요청 횟수(Request Rate) 기반의 방어뿐만 아니라, 프롬프트 입력 길이에 따른 토큰(Token) 수 기반의 자원 고갈 공격(Resource Exhaustion)을 방어하기 위한 설계가 필수적이다. TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI)에서 강조하듯, 악의적으로 조작된 긴 컨텍스트 입력에 의한 연산 자원 고갈을 차단하는 설계가 요구된다.

3. 검토방법

No	점검 항목	확인 방법
1	DDoS 방어 요구사항 설계 반영 여부	{보안요구사항}의 DDoS 방어 관련 항목이 {보안설계서}에 빠짐없이 반영되었는지 {요구사항추적표}와 대조
2	AI 추론 API L7 DDoS 방어 설계 명시 여부	AI 추론 API 엔드포인트 대상 L7 DDoS 방어 설계가 {보안설계서}에 구체적으로 명시되었는지 확인
3	WAF AI 특화 공격 패턴 규칙 설계 여부	프롬프트 인젝션 기반 연산 고갈 등 AI 특화 공격 패턴 탐지·차단 규칙이 WAF 정책에 설계되었는지 {보안설계서}·'WAF 정책 설계서'에서 확인
4	L7 스위치 트래픽 분산·차단의 AI 특성 반영 여부	L7 스위치의 트래픽 분산·비정상 트래픽 차단 설계가 AI 추론 서비스 특성을 반영했는지 {보안설계서}·{아키텍처설계서}에서 확인
5	DDoS 자동 대응 절차·임계값 포함 여부	DDoS 공격 탐지 시 자동 대응(Auto-Mitigation) 절차·임계값이 설계에 포함되었는지 {보안설계서} 검토·면담으로 확인
6	DDoS 방어 장비 배치 위치 설계 여부	방화벽·WAF·L7 스위치 등 DDoS 방어 장비의 배치 위치가 AI 서비스 구간을 적절히 보호하도록 네트워크 토폴로지에 설계되었는지 {아키텍처설계서}에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{보안요구사항}의 DDoS 방어 관련 항목이 {보안설계서}에 빠짐없이 반영되고, AI 추론 API 대상 L7 방어 설계, AI 특화 공격 패턴 탐지·차단 규칙, 자동 대응(Auto-Mitigation) 절차와 임계값, 방어 장비 배치가 구체적으로 설계되어 있는 경우	현행 유지
보완권고 (△)	DDoS 방어 설계는 존재하나 AI 특화 공격 패턴 차단 규칙, 자동 대응 임계값 등 일부 항목이 누락되었거나, '적절한 DDoS 방어를 적용한다' 수준의 선언적 문구로 기술되어 정량적 기준이 부족한 경우	설계 요건 보완 및 허용 기준 명시 요구 후 재확인
부적합 (X)	DDoS 방어 및 비정상 트래픽 차단 설계가 {보안설계서}에 반영되지 않았거나, {요구사항추적표}상 DDoS 관련 요구사항의 설계 매핑이 확인되지 않아 추적성이 확보되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 특화 DDoS와 일반 DDoS 설계의 구분: 본 가이드라인은 AI 서비스 고유의 DDoS 방어 설계(추론 API 보호, 연산 고갈 공격 방어 등)를 점검 대상으로 한다. 일반적인 네트워크 L3/L4 DDoS 방어(SYN Flood, UDP Amplification 등)는 기존 IT 감리 영역에서 점검하므로, 이를 구분하여 점검 범위를 명확히 해야 한다.

2) 설계 수치의 현실성 검증: Rate Limiting 임계값이나 페이로드 크기 제한 등 수치가 설계에 포함되어 있더라도, 해당 수치가 AI 서비스의 실제 사용 패턴(예: 최대 입력 토큰 수, 평균 추론 시간)을 고려한 합리적인 값인지 사업자 면담을 통해 확인해야 한다. 예를 들어 Rate Limiting을 초당 1,000건으로 설정했지만 AI 추론 서버의 실제 처리 용량이 초당 50건이라면, 해당 임계값은 방어 효과가 없다.

3) 클라우드 환경의 DDoS 방어 설계 특성: 클라우드 기반 AI 서비스의 경우, CSP(Cloud Service Provider)가 제공하는 DDoS 방어 서비스(AWS Shield, Azure DDoS Protection, GCP Cloud Armor 등)의 활용 계획이 설계에 반영되었는지 확인하되, CSP 기본 제공 기능만으로 AI 특화 L7 공격까지 방어 가능한지 추가 검토가 필요하다.

4) 선언적 문구의 구체화 유도: '적절한 DDoS 방어를 적용한다', 'WAF를 통해 보안을 강화한다'와 같은 모호한 설계 문구는 구현 및 분쟁의 원인이 된다. 'AI 추론 API에 초당 100건 Rate Limiting 적용', 'WAF에 프롬프트 인젝션 탐지 규칙 30건 이상 등록' 등 측정 가능한 설계 기준으로 구체화하도록 유도해야 한다.

5) 요구사항 추적성의 완전성 우선 확인: 설계 내용의 기술적 적정성 검토에 앞서, {요구사항추적표}를 통한 추적성 확인을 선행해야 한다. 추적표에 매핑이 누락된 요구사항이 존재하면, 해당 항목은 설계 반영 여부와 관계없이 '추적성 미확보'로 별도 지적이 필요하다.

AG_SD42-1-13 | 보안약점 예방 아키텍처 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 시큐어코딩

근거 | 「전자정부법」 · OWASP LLM Top 10 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)

1. 목적

설계(상세설계) 단계에서 요구사항 단계에서 수립된 보안 요구사항이 {보안설계서}와 관련 설계 산출물에 구체적으로 반영되었는지 검토한다.

특히 AI 서비스의 구조적 취약점인 SQL Injection 및 프롬프트 인젝션 공격에 대응하기 위한 방어 필터 레이어가 시스템 아키텍처 수준에서 설계되었는지 여부와, 소프트웨어 보안약점 예방 원칙이 구성요소 설계에 내재화되었는지를 확인하는 것이 목적이다.

2. 필요성

▶ 생성형 AI 서비스는 사용자 입력(Prompt)을 자연어 형태로 처리하는 구조를 가지므로, 기존 웹 서비스에서 발생하는 SQL Injection과 함께 AI 고유의 보안 위협인 프롬프트 인젝션에 동시에 노출된다. 이러한 공격이 설계 단계에서 방어 레이어 없이 구현될 경우 다음과 같은 문제가 발생할 수 있다.

▶ SQL Injection 공격 시 데이터베이스의 비인가 조회·삭제·변조가 발생하여 공공데이터 및 개인정보가 유출될 수 있다.

▶ 프롬프트 인젝션 공격 시 AI 모델의 System Prompt를 우회하거나 악의적인 명령을 삽입하여 AI가 금지된 콘텐츠를 생성하거나, 내부 시스템 정보를 노출할 수 있다.

▶ 설계 단계에서 방어 레이어가 누락되면 이후 구현·시험 단계에서의 보완이 구조적으로 어려워지고, 재설계 비용이 급증하는 Risk가 있다.

▶ 「전자정부 SW 개발·운영자를 위한 소프트웨어 개발보안 가이드」에서는 보안 요구사항 반영을 의무화하고 있으며, 공공 AI 서비스는 이를 준수해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	SQL Injection 방어 설계 반영 여부	{보안설계서}에 SQL Injection 방어 요구사항이 아키텍처 설계 수준으로 구체화되어 있는지 확인
2	프롬프트 인젝션 방어 레이어 설계 여부	{보안설계서} 및 'AI 서비스 아키텍처 설계서'에 프롬프트 인젝션 공격 대응을 위한 필터 레이어가 명시되었는지 확인

No	점검 항목	확인 방법
3	입력값 유효성 검증 및 출력 인코딩 설계	'인터페이스 설계서'를 검토하여 모든 외부 입력 경로(API, 파일 업로드, 채팅 입력 등)에 대해 입력 검증 정책이 정의되었는지 확인
4	보안 요구사항 추적성 확인	'보안요구사항 추적표'를 검토하여, 기획/분석 단계에서 도출된 SQL Injection 및 프롬프트 인젝션 관련 보안 요구사항이 설계 산출물({보안설계서}, 'AI 서비스 아키텍처 설계서' 등)에 누락 없이 매핑되어 있는지 확인
5	System Prompt 보안 관리 설계	'Prompt 설계서' 또는 {보안설계서}에 System Prompt의 보안 관리 방침이 명시되었는지 확인
6	보안약점 제거 설계 원칙 적용 여부 (OWASP Top 10 LLM / SW 보안약점 기준)	{보안설계서}가 행정안전부·KISA의 「소프트웨어 개발보안 가이드」에서 규정한 설계 단계 보안약점 제거 원칙을 반영하고 있는지 점검

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{보안설계서}에 SQL Injection 및 프롬프트 인젝션 방어 레이어가 아키텍처 수준에서 명시적으로 설계되고, 입력 유효성 검증·출력 인코딩 정책과 System Prompt 보안 관리 방침이 반영되어 있으며, 보안요구사항 추적표로 분석~설계 간 추적이 가능한 경우	현행 유지
보완권고 (△)	SQL Injection 방어 방침은 존재하나 프롬프트 인젝션 방어 레이어 설계가 명시되지 않았거나, 일부 입력 경로의 검증 정책 누락, 추적표 부재(인터뷰로 반영 확인), System Prompt 보안 방침의 미흡 기술 등 부분적 미비가 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안설계서}에 SQL Injection 및 프롬프트 인젝션 방어 설계가 전혀 없고, 아키텍처 설계서에 보안 처리 레이어가 존재하지 않으며, 분석 단계의 보안 요구사항이 설계 산출물에 반영되지 않아 추적이 불가능한 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 보안 설계와 '솔루션 도입 계획'을 혼동하지 말 것: 일부 사업자는 {보안설계서}에 "WAF 도입", "AI 보안 솔루션 적용" 등 솔루션 명칭만 기술하고 실제 필터 레이어 설계 내용을 생략한다. 감리원은 솔루션 도입 계획이 아닌, "어느 컴포넌트에서, 어떤 방식으로, 어떤 입력에 대해" 방어가 수행되는지를 아키텍처 수준에서 확인해야 한다.

2) 프롬프트 인젝션 점검은 생성형 AI 전용 관점으로 접근할 것: 두 공격은 공통적으로 "입력 조작"이라는 특성을 가지지만, 프롬프트 인젝션은 AI 모델의 언어 이해 메커니즘을 악용하는 점에서 SQL Injection과 다르다. {보안설계서}가 프롬프트 인젝션을 단순 키워드 필터링으로만 대응하도록 설계한 경우, 우회 공격에 취약하므로 다계층 방어(입력 정제, Delimiter 격리, 출력 검증 등) 설계를 권고해야 한다.

3) 정량적 기준 수치화 유도: 점검 시 형식 충족 여부에 그치지 말고, 실질적 이행 근거와 책임자 서명 여부까지 확인하라.

AG_SD42-1-14 | 웹셸 악성파일 업로드 차단 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 시큐어코딩

근거 | 「전자정부법」 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · OWASP LLM Top 10

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구사항 분석 시 도출된 웹셸(Webshell) 차단, 악성파일 업로드 방지, 파일 확장자 제한 및 업로드 디렉터리 실행 통제 요건이 {보안설계서}에 누락 없이 구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 AI 서비스는 문서·이미지·데이터셋 등 다양한 형태의 파일 업로드 기능을 제공하는 경우가 많아, 웹셸 업로드를 통한 서버 침탈 및 AI 모델·학습데이터 탈취 위험이 일반 정보시스템보다 높으므로 파일 업로드 관련 보안 통제 요건이 설계 단계부터 명확히 반영되었는지 확인한다.

아울러 AI 환경(특히 RAG 기반)에서는 단순 실행파일(.php 등) 업로드 차단뿐만 아니라, PDF나 Word 문서 내부에 은닉된 악성 매크로나 페이로드가 파싱(Parsing) 과정에서 서버를 공격하는 것을 막아야 하므로, 악성 문서 파일 파싱 과정에서의 취약점 방어(예: 샌드박스 격리) 관점까지 포함하여 점검한다.

2. 필요성

▶ AI 서비스는 사용자로부터 학습 데이터, 참조 문서, 이미지 등 다양한 파일을 업로드받는 기능을 포함하는 경우가 빈번하며, 이러한 파일 업로드 경로는 공격자가 웹셸(.php, .asp, .jsp 등 실행 가능한 스크립트 파일)을 서버에 심어 원격에서 시스템 명령을 실행하는 통로로 악용될 수 있다. 웹셸이 서버에 설치되면 시스템 전체가 공격자에게 장악되어 AI 모델과 학습데이터가 탈취될 수 있다.

▶ AI 서비스 고유의 파일 업로드 위협: 생성형 AI 서비스는 RAG(검색 증강 생성) 연계를 위한 문서 업로드, 멀티모달 입력 처리 등 다양한 파일 업로드 경로가 존재한다. 이는 일반 웹 서비스 대비 공격 표면이 넓어 웹셸 업로드 차단 설계의 중요도가 높다.

▶ 설계 단계 반영의 중요성: 파일 업로드 보안 통제는 개발 완료 후 사후 적용이 매우 어렵다. 업로드 디렉터리 실행 권한 분리, 파일 확장자 화이트리스트, MIME 타입 검증 등은 시스템 아키텍처 설계 단계에서 결정되어야 한다.

▶ 공공 AI 사업의 보안 의무 준수: 행정안전부 소프트웨어 개발 보안 가이드, NIST SSDF(시큐어 소프트웨어 개발 프레임워크)에서는 입력값 검증 및 파일 업로드 통제를 필수 보안 요건으로 규정하고 있으며, 공공 AI 사업에서 이를 설계에 반영하지 않을 경우 검사 단계에서 중대 결함으로 지적될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	파일 확장자 화이트리스트 통제 설계 여부	{보안설계서}에서 허용 확장자 목록이 서버 측 검증으로 설계되었는지 확인. 클라이언트 측 검증만 기술된 경우 부적합으로 판정
2	MIME 타입 및 파일 헤더(매직 바이트) 이중 검증 설계 여부	확장자 위장 업로드를 차단하기 위해 MIME 타입과 매직 바이트를 함께 검증하도록 설계되었는지 확인하고, 검증 실패 시 처리 방식을 면담으로 확인
3	업로드 디렉터리 실행 권한 분리 및 저장소 설계 여부	업로드 디렉터리가 웹 루트 외부에 위치하고 실행 권한이 제거되도록 설계되었는지 확인하고, 오브젝트 스토리지 사용 시 접근 정책을 면담으로 확인
4	AI 서비스 특화 파일 입력 경로(RAG 등) 보안 통제 설계 여부	RAG 문서 업로드·멀티모달 입력 등 AI 특화 파일 입력 경로에도 동일한 업로드 보안 통제가 적용되도록 설계되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{보안설계서}에 ① 파일 확장자 화이트리스트(서버 측 검증), ② MIME 타입·매직 바이트 이중 검증, ③ 업로드 디렉터리 실행 권한 분리, ④ AI 입력 경로 보안 통제가 모두 구체적으로 명시된 경우	현행 유지
보완권고 (△)	일부 항목(예: MIME 타입 이중 검증 또는 AI 특화 경로 보안)이 누락되었으나, 파일 확장자 제한 및 디렉터리 권한 분리는 설계에 포함된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{보안설계서}에 파일 업로드 관련 보안 통제 항목이 전혀 기술되지 않았거나 클라이언트 측 검증만 기술된 경우, 또는 업로드 디렉터리가 웹 루트 내에 설계된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 보안 요건의 구체성 확인: '파일 업로드 보안을 적용한다'와 같은 선언적 문구는 향후 이행 여부를 확인하기 어렵다. '허용 확장자: .pdf, .docx, .xlsx, .png, .jpg 5종으로 제한', '업로드 디렉터리는 웹 루트 외부에 배치하고 스크립트 실행 권한 제거' 등 방식이 명확한 설계 내용으로 구체화하도록 유도해야 한다.

2) 파일 업로드 기능 유무 먼저 확인: AI 서비스라도 파일 업로드 기능이 없는 경우 해당 상세 검토 항목은 '해당 없음(N/A)'으로 처리한다. 그러나 RAG 연계 문서 수집, 멀티모달 입력, API를 통한 배치 업로드 등 간접적인 파일 입력 경로가 있는지 반드시 사업자 면담을 통해 확인해야 한다.

3) AI 특화 간접 프롬프트 인젝션 점검: 웹셸 차단 외에도, 업로드 문서 내 악의적인 지시문이 AI 모델에 전달되는 '간접 프롬프트 인젝션' 위협은 일반 IT 감리에서 다루지 않는 AI 특화 위협이다. {보안설계서}에 RAG 파이프라인 입력 문서의 콘텐츠 검사 절차가 별도로 반영되어 있는지 확인한다.

4) 요구사항과 설계 간 추적성 확인: '요구사항 추적표' 또는 '{보안요구사항추적표}'와 {보안설계서}를 대조하여, 도출된 파일 업로드 보안 요건이 설계에 빠짐없이 반영되었는지 추적성(Traceability) 관점에서도 점검한다.

5) 문서 파싱 라이브러리 연계 보안 설계 점검: 사용자가 업로드한 문서(PDF, PPT, DOCX)를 AI가 읽기 위해 텍스트로 추출(Parsing)하는 모듈이 설계상 웹 루트나 메인 API 서버와 논리적/물리적으로 격리(예: 컨테이너 분리)되어 설계되었는지 {아키텍처설계서}를 통해 점검한다.

AG_SD42-1-15 | 클라우드 인프라 보안그룹 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 망분리/폐쇄망

근거 | CSAP · 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 수립된 클라우드 인프라 보안 요구사항이 {보안설계서}의 VPC(Virtual Private Cloud) 아키텍처 및 Security Group 규칙으로 빠짐없이 상세하게 설계되었는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 서비스 특성상 모델 학습용 GPU 인스턴스, 추론 API 서버, 벡터 DB, 외부 LLM API 연계 구간 등 다양한 컴포넌트 간 네트워크 흐름이 복잡하게 구성되므로, Security Group의 포트 오픈 범위가 최소 권한 원칙(Least Privilege)에 부합하도록 설계되었는지, VPC 서브넷 간 격리가 AI 워크로드의 보안 등급에 맞게 구획되었는지를 확인한다.

2. 필요성

▶ AI 서비스 인프라는 모델 학습, 추론 API, 데이터 파이프라인, 외부 LLM 연계 등 다수의 네트워크 구간이 상호 연결되어 있으며, 각 구간의 보안 수준이 상이하다. 설계 단계에서 VPC와 Security Group이 부적절하게 구성되면 다음과 같은 AI 특유의 보안 위험이 발생할 수 있다.

▶ 모델 탈취(Model Extraction) 방지: AI 추론 API 서버의 포트가 불필요하게 외부에 개방되면, 반복 질의를 통한 모델 복제·탈취 공격에 노출된다. Security Group에서 추론 API 포트를 내부 서브넷 또는 허가된 IP 대역으로 한정하도록 설계되어야 한다.

▶ 학습 데이터 유출 차단: 모델 학습용 GPU 인스턴스가 위치한 서브넷이 외부 인터넷에 직접 연결되면, 학습 데이터 또는 모델 가중치(Weights)가 외부로 유출될 위험이 있다. Private Subnet 배치 및 NAT Gateway 경유 설계가 필요하다.

▶ 프롬프트 인젝션 경유 내부 침투 방지: 생성형 AI 서비스에서 프롬프트 인젝션을 통해 내부 시스템 정보를 탈취하거나 내부 API를 호출하는 공격이 발생할 수 있으므로, AI 서비스 계층과 내부 관리 계층 간 VPC 서브넷 격리 및 Security Group 분리 설계가 요구된다.

▶ 외부 LLM API 연계 구간 보안: 외부 클라우드 LLM API(예: OpenAI, Claude API)와 통신하는 구간의 아웃바운드 규칙이 무제한 개방(0.0.0.0/0, All Ports)으로 설계되면, 데이터 유출 경로가 되므로 목적지 IP·포트를 한정하는 설계가 필수적이다.

3. 검토방법

No	점검 항목	확인 방법
1	학습 인스턴스 Private Subnet 배치 설계 여부	AI 모델 학습 인스턴스가 Private Subnet에 배치되도록 {보안설계서} 서브넷 구성도에 설계되었는지 확인
2	추론·학습 서버 서브넷 분리 설계 여부	추론 API 서버와 학습 서버가 별도 서브넷으로 분리 설계되었는지 {보안설계서} VPC 구성도에서 확인
3	데이터 저장소 전용 서브넷 격리 설계 여부	벡터 DB·학습 데이터 저장소가 데이터 전용 서브넷에 격리 설계되었는지 '클라우드 인프라 설계서'에서 확인
4	외부 접근 NAT Gateway 경유 설계 여부	외부 인터넷 접근이 필요한 구간(외부 LLM API 연계)만 NAT Gateway 경유로 설계되었는지 {보안설계서} 라우팅 설계에서 확인
5	VPC 피어링·Private Link 보안 통제 설계 여부	VPC 간 피어링(Peering) 또는 Private Link 구간의 보안 통제 설계가 포함되었는지 {아키텍처설계서} 연계 구간에서 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	학습 인스턴스의 Private Subnet 배치, 추론·학습 서버 서브넷 분리, 데이터 계층 격리, NAT Gateway 경유 아웃바운드 통제, 피어링 구간 접근제어가 {보안설계서}에 포트 매트릭스 수준으로 구체적으로 설계된 경우	현행 유지
보완권고 (△)	VPC·Security Group 설계가 반영되어 있으나 일부 구간(데이터 서브넷 격리, 피어링 통제 등)이 누락되었거나, 규칙이 텍스트 서술에 그쳐 포트·IP 대역 등 세부 기준이 구체화되지 않은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	서브넷 격리 설계 없이 AI 워크로드가 Public Subnet에 배치되거나, 0.0.0.0/0 전면 개방 규칙 등 최소 권한 원칙에 반하는 설계로 보안 요구사항이 반영되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 포트 매트릭스 요구: {보안설계서}에 Security Group 규칙이 텍스트로만 서술되어 있는 경우, 반드시 포트 매트릭스(소스 IP/대상 IP/프로토콜/포트/허용·차단) 형태의 표로 작성되었는지 확인해야 한다. 텍스트 서술만으로는 설계 누락 여부를 판별하기 어렵다.

2) AI 특화 포트 확인: 일반 웹 서비스와 달리 AI 서비스에서는 GPU 통신(NCCL: 29400), 모델 서빙(gRPC: 8500-8501, TensorFlow Serving: 8501), 벡터 DB(Milvus: 19530, Pinecone: 443) 등 AI

전용 포트가 사용된다. 이러한 포트가 설계서에 구체적으로 명시되었는지, 불필요한 포트가 추가로 개방되지 않았는지 확인해야 한다.

3) 0.0.0.0/0 규칙 전수 조사: Security Group 설계에서 소스 또는 목적지가 0.0.0.0/0(Any)으로 설정된 규칙이 존재하면, 해당 규칙의 필요성과 대안(특정 IP 대역 제한, Security Group 참조 방식 등)을 사업자에게 확인해야 한다. 이를 발견하지 못하면 구축 후 수정에 큰 비용이 소요된다.

4) CSP별 용어 차이 주의: AWS의 'Security Group'은 Azure의 'NSG(Network Security Group)', GCP의 '방화벽 규칙(Firewall Rules)'과 대응된다. 사업자가 사용하는 CSP의 정확한 용어를 사전에 확인하고, 설계서의 용어와 실제 CSP 기능 간 불일치가 없는지 점검해야 한다.

5) 설계-구현 Gap 방지를 위한 산출물 확인: 설계 단계 감리에서는 실제 클라우드 환경을 직접 확인할 수 없으므로, {보안설계서}에 Security Group 규칙이 IaC(Infrastructure as Code) 형태(예: Terraform, CloudFormation 템플릿)로 코드화될 계획이 포함되어 있는지 확인한다. IaC 기반 배포 설계가 있으면 설계와 구현 간 불일치를 줄이고 구현 단계 감리 시 검증도 용이해진다.

AG_SD42-1-16 | 테스트환경 구축 계획

시스템구조·보안 | 설계 | 시스템 구현 | 테스트시나리오

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업에서, 사업자가 AI 시스템 검증에 필요한 테스트 환경을 {테스트계획서}에 구체적으로 정의하였는지, 특히 가상 테스트 환경과 실제 운영 환경 간의 차이를 요구사항 수준에서 인식하고 반영하였는지를 제3자 관점에서 점검하는 것을 목적으로 한다.

테스트 환경 정의 단계에서의 누락이나 모호성은 이후 테스트 수행 단계의 품질 분쟁과 검수 지연의 직접적 원인이 된다.

이에 따라 감리원은 {테스트계획서} 상에 AI 특유의 테스트 환경 요건(데이터 유사성, 모델 버전 관리, LLM 연계 환경 등)이 측정 가능한 수준으로 기술되었는지를 중점적으로 확인한다.

2. 필요성

▶ AI 시스템은 학습 데이터의 분포, 모델 버전, 외부 API 응답 특성 등이 환경에 따라 상이하여 일반 IT시스템과 달리 테스트 환경과 운영 환경 간의 차이가 성능 및 품질에 직접적인 영향을 미치며, 요구정의 단계에서 테스트 환경의 정의가 불명확하거나 운영 환경과의 유사성 요건이 누락되면 다음과 같은 위험이 발생한다.

▶ 테스트 결과의 재현 불가: 테스트 환경과 운영 환경의 데이터·모델 조건이 다를 경우, 테스트에서 통과한 기능이 운영 시 동일하게 동작함을 보장할 수 없어 배포 이후 예기치 않은 오동작이 발생할 수 있다.

▶ 검수 기준 분쟁: {테스트계획서}에 테스트 환경의 구성 조건이 명시되지 않으면 발주기관과 사업자 간에 "어떤 환경에서 어떤 결과를 확인했는가"에 대한 해석 차이가 발생하고, 이는 검수 단계에서의 극심한 갈등으로 이어진다.

▶ 생성형 AI 특유 리스크: 외부 LLM API(예: 클라우드 기반 언어모델)를 활용하는 AI 서비스의 경우, 테스트 환경에서 사용한 모델 버전과 운영 환경의 버전이 상이하면 동일한 프롬프트에 대한 응답 결과가 달라질 수 있다.

▶ 품질보증 체계의 공백: TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 분야) 및 NIST AI RMF는 AI 시스템의 검증은 배포 조건에 가까운 환경에서 수행되어야 함을 명시하고 있으며, 이를 요구사항으로 반영하지 않으면 품질보증 체계의 공백이 발생한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 테스트 환경 구성 요소 정의 여부	{테스트계획서}에 AI 테스트에 필요한 환경 구성 요소(데이터셋, 모델 버전, 외부 API 연계 정보, 인프라 사양 등)가 항목별로 명시되었는지 확인
2	운영 환경 유사성 요건 명시 여부	AI 시스템의 테스트 결과가 운영에서도 재현 가능하도록 운영 환경과의 유사도(동일 모델 버전, 동일 API 엔드포인트, 동등한 데이터 분포 등)를 요구사항으로 명시하였는지 확인
3	가상 테스트 환경 정의 및 활용 계획 여부	운영 환경 직접 구축이 어려운 경우 가상 테스트 환경(Mock API, 합성 데이터 등)의 활용 계획과 한계에 대한 명시가 {테스트계획서}에 포함되었는지 확인
4	AI 특화 테스트 항목 정의 여부	일반 기능 테스트 외에 AI 특유의 검증 항목(환각(Hallucination) 발생률, 편향 검출, 프롬프트 인젝션 저항성, 응답 일관성 등)이 {테스트계획서}에 포함되었는지 확인
5	테스트 데이터 대표성 및 품질 요건 명시 여부	테스트에 사용할 데이터셋이 운영 시 예상되는 실제 사용자 입력 분포를 반영하는지에 대한 요건과 테스트 데이터 품질 기준이 명시되었는지 확인
6	테스트 환경 버전 관리 계획 반영 여부	LLM 모델 버전, 라이브러리 버전 등 AI 시스템의 구성 요소 버전 관리 계획이 {테스트계획서} 또는 {사업수행계획서}에 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	테스트 환경 구성 요소, 운영 환경 유사성 요건, 가상 테스트 환경 활용 계획과 한계, AI 특화 테스트 항목, 테스트 데이터 대표성, 버전 관리 계획이 모두 {테스트계획서}에 측정 가능한 수준으로 구체적으로 명시된 경우	현행 유지
보완권고 (△)	일부 항목이 반영되었으나 운영 환경 유사도 요건의 선언적 기술, 가상 환경 한계 미명시, AI 특화 검증 항목 일부 누락, 편향 허용 기준 미정의 등 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 테스트 환경 요건이 {테스트계획서}에 전혀 반영되지 않았거나, 테스트 데이터·모델 버전 관리 등 핵심 요건이 완전히 누락되어 테스트 결과의 재현성과 검수 기준을 확보할 수 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 기준의 수치화 유도: "AI가 정확한 답변을 제공해야 한다"는 선언적 문구는 검수 분쟁의 씨앗이 된다. 감리원은 "응답 정확도 90% 이상", "환각 발생률 5% 이하", "테스트 데이터 커버리지 80% 이상" 등 측정 가능한 기준을 {테스트계획서}에 명시하도록 보완 권고하여야 한다.

2) 가상 테스트 환경의 위험 인식 여부 확인: 사업자가 Mock API 또는 합성 데이터를 활용하는 경우, 이를 단순히 허용할 것이 아니라 실제 운영 환경과의 차이로 인해 발생할 수 있는 위험에 대한 인식과 보완 방안이 {테스트계획서}에 문서화되어 있는지 반드시 확인해야 한다.

3) LLM 버전 고정(Model Freeze) 계획 점검: 외부 클라우드 LLM API를 활용하는 AI 서비스는 테스트 시점과 운영 시점의 모델 버전이 달라질 수 있다. 테스트 환경에서 사용한 모델 버전을 고정하거나 버전 변경 시 재검증 절차를 {테스트계획서}에 포함하도록 권고한다.

4) 일반 IT 테스트와 AI 테스트 항목의 구분 확인: {테스트계획서}가 기존 일반 IT 시스템 기준으로만 작성되어 AI 특화 항목(편향, 환각, 공정성, 프롬프트 인젝션 등)이 누락되는 사례가 많다. 감리원은 AI 특화 항목의 존재 여부를 별도로 체크리스트화하여 확인할 필요가 있다.

AG_SD42-1-17 | 자동화 테스트 설계점검

시스템구조·보안 | 설계 | 시스템 구현 | 테스트시나리오

근거 | NIST AI RMF MEASURE · OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

요구사항 분석 단계에서 정의된 AI 서비스의 기능 요구사항 및 품질 요구사항이 상세설계 단계에서 {테스트계획서}, "AI 테스트 자동화 설계서", "성능/보안 자동화 테스트 스크립트 작성 계획"에 빠짐없이 구체적으로 반영되었는지를 검토하는 것이 목적이다.

AI 서비스는 출력 결과의 비결정성(Non-deterministic)으로 인해 일반적인 소프트웨어와 동일한 테스트 설계 방식만으로는 품질을 보장할 수 없다.

따라서 자동화 도구 활용 계획의 적정성, AI 특화 테스트 항목(환각 탐지, 프롬프트 인젝션 대응 등)의 반영 여부, 수작업 테스트와 자동화 테스트의 역할 분리 명확성을 중점 점검하여 사업 후반부의 품질 분쟁을 사전에 예방한다.

2. 필요성

▶ 상세설계 단계는 요구사항이 구현 가능한 산출물로 구체화되는 시점이므로, 이 단계에서 테스트 설계가 충분히 이루어지지 않으면 검사 단계에서 수용기준 불명확으로 인한 분쟁이 발생하거나 AI 특화 결함이 사업 종료 직전에 발견되는 리스크가 높아진다.

▶ AI 서비스 구축 사업에서 자동화 테스트 설계가 특히 중요한 이유는 다음과 같다.

▶ 첫째, AI 모델의 응답 품질(정확도, 응답 일관성 등)은 대규모 반복 테스트 없이는 통계적으로 의미 있는 평가가 어렵다. 자동화 도구 없이 수작업만으로 충분한 테스트 커버리지를 확보하기 어렵다.

▶ 둘째, 성능 테스트(동시 접속, 응답 지연 등)와 보안 취약점 테스트(프롬프트 인젝션 등)는 자동화 도구 없이는 현실적으로 실행이 불가능하다.

▶ 셋째, 요구사항과 테스트케이스 간 추적이 가능한 {요구사항추적표}가 부재할 경우, 감리원이 요구사항의 테스트 반영 여부를 객관적으로 확인하기 어렵다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 기능 요구사항의 테스트 설계 반영 여부	{테스트계획서} 검토

No	점검 항목	확인 방법
2	AI 품질 요구사항(정확도, 응답시간, 안전성 등)의 테스트 항목화 여부	{테스트계획서}, "AI 품질 지표 정의서" 검토
3	자동화 테스트 대상 선정 기준의 적절성	"AI 테스트 자동화 설계서" 검토
4	성능 자동화 테스트 스크립트 작성 계획의 구체성	"성능/보안 자동화 테스트 스크립트 작성 계획" 검토
5	보안 자동화 테스트 항목 및 AI 특화 취약점 점검 계획 포함 여부	"성능/보안 자동화 테스트 스크립트 작성 계획" 검토
6	자동화 도구의 AI 서비스 적합성 검토 여부	"AI 테스트 자동화 설계서", 인터뷰
7	{요구사항추적표}를 통한 테스트 케이스 추적 가능성	{요구사항추적표}, {테스트계획서} 교차 검토
8	수작업 테스트와 자동화 테스트의 역할 분리 명확성	{테스트계획서} 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 기능·품질 요구사항이 측정 가능한 테스트 항목으로 반영되고, 자동화 대상 선정 기준·스크립트 작성 계획·AI 특화 보안 점검 계획이 구체적으로 설계되며, 자동화 도구의 AI 서비스 적합성이 검토되고 수작업 테스트와 자동화 테스트의 역할이 명확히 구분되어 {요구사항추적표}로 추적이 가능한 경우	현행 유지
보완권고 (△)	테스트 설계는 존재하나 품질 기준의 수치화 미흡, AI 특화 취약점 항목 일부 누락, 수작업·자동화 역할 구분 불명확 등 보완이 필요한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	자동화 테스트 설계 산출물이 부재하거나 AI 특화 테스트 항목이 반영되지 않아 요구사항의 테스트 가능성이 확보되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 테스트 항목의 수치화 유도: 설계 산출물에 "AI는 정확한 답변을 해야 한다"와 같이 모호하게 기술된 경우, 구현 완료 후 검사 단계에서 수용 여부를 놓고 분쟁이 발생할 가능성이 높다. 감리 지적 시 "환각 탐지율 5% 이하", "정답 일치율 90% 이상" 등으로 측정 가능한 기준으로 구체화하도록 권고한다.

2) 자동화 범위의 현실성 검토: 자동화 테스트 계획이 과도하게 광범위하게 설정된 경우(예: AI 응답 품질 전부를 자동화), 실제 구현 불가로 계획과 실행이 불일치하는 사태가 발생할 수 있다. 자동화 적합 항목(성능, 보안, 회귀테스트)과 수작업 필수 항목(맥락 적절성, 윤리성 평가)이 현실적으로 구분되어 있는지를 확인한다.

3) 도구 명칭 확인 시 AI 특화 여부 검토: JMeter, Selenium 등 일반 도구만 나열된 계획은 AI 응답 품질·프롬프트 인젝션 테스트를 수행할 수 없는 경우가 많다. 선정된 도구가 AI 특화 테스트를 지원하는지, 미지원 시 별도 스크립트 개발 계획이 있는지를 인터뷰로 확인한다.

AG_SD42-1-18 | SLM 경량 모델 활용 요건

시스템구조·보안 | 설계 | 인공지능 모델 개발 | 모델경량화

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · 「저작권법」

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 온프레미스 또는 엣지 환경에서 SLM(Small Language Model, 소형 언어 모델)을 활용할 때 모델 선정·성능 검증·보안·운영 요건이 적절히 수립되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 'SLM 모델 선정 평가서', '파인튜닝 계획서', 'SLM 운영 인프라 설계서', 라이선스 의견서를 근거로 6대 평가 차원의 정량 비교, 파인튜닝 설계의 구체성, 추론 인프라 산정의 타당성, 라이선스 준수와 보안 통제를 확인한다.

감리원의 점검은 모델 성능의 절대적 우수성이 아니라 선정 근거의 정량성과 운영 요건의 완결성에 한정한다.

2. 필요성

▶ 외부 LLM API 의존도 감소, 데이터 보안 강화, 비용 절감을 목적으로 온프레미스 SLM 활용이 증가하고 있으며, 공공기관 망분리 환경에서는 외부 API 호출이 불가하여 SLM이 유일한 대안이 되는 경우가 많다.

▶ SLM은 대규모 모델 대비 범용 성능이 낮으므로, 도메인 특화 평가 없이 공개 벤치마크 점수만으로 선정하면 실제 업무에서 요구 정확도를 충족하지 못한다.

▶ 파인튜닝 방식(LoRA·QLoRA·Full)에 따라 필요한 GPU 자원과 학습 데이터 규모가 크게 달라진다. 방식 선택의 근거와 데이터 라벨링 품질 기준이 사전에 정의되어야 한다.

▶ 오픈 가중치 모델의 라이선스는 상업적 이용과 재배포, 파인튜닝 결과물의 배포 조건이 모델별로 상이하다. 라이선스 검토 없이 도입하면 서비스 개시 후 법적 분쟁의 원인이 된다.

▶ 모델 가중치 파일이 자체 인프라에 저장되므로 무결성 검증과 유출 방지 통제가 필요하며, TTA 신뢰할 수 있는 인공지능 개발안내서는 운영 환경에 적합한 모델 선정을 권고한다.

3. 검토방법

No	점검 항목	확인 방법
1	6대 평가 차원 검증	'SLM 모델 선정 평가서'에서 한국어 성능·도메인 특화 성능·추론 속도·메모리 효율·라이선스·커뮤니티 유지보수의 6개 차원이 정량 점수와 함께 평가되었는지 확인하고, 공개 벤치마크 결과와 대조

No	점검 항목	확인 방법
2	파인튜닝 설계 (LoRA·QLoRA·Full, 학습 데이터 규모·라벨링 품질)	'파인튜닝 계획서'에서 파인튜닝 방식(LoRA·QLoRA·Full) 선택 근거와 학습 데이터 규모·라벨링 품질 기준이 제시되었는지 확인
3	추론 인프라 적정성	'SLM 운영 인프라 설계서'에서 GPU 자원 산정 근거, 양자화 적용 여부, 추론 엔진 선정이 기술되었는지 확인하고, 부하 시험 결과로 목표 처리량 달성을 검증
4	라이선스 준수 (상업·재배포 허용 조건)	SLM 모델의 상업적 이용·재배포 허용 조건을 라이선스 원문과 대조하고, 법무 검토 의견서에 사업 이용 형태의 적법성이 명시되었는지 확인
5	보안 통제 (모델 무결성·가중치 유출 방지·민감정보 보호)	모델 파일의 무결성 검증(서명·해시), 가중치 유출 방지 통제, 추론 입력의 민감정보 보호 방안이 'SLM 운영 인프라 설계서'에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	6대 평가 차원·파인튜닝·인프라·보안 요건이 정량 근거와 함께 설계되어 후속 단계 추적이 가능한 경우	현행 유지
보완권고 (△)	모델 선정은 수행했으나 파인튜닝·추론 엔진 또는 보안 요건이 일부 누락되어 구체성·완전성이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	단일 모델을 임의 선정하고 양자화·추론 엔진 검토가 부재하는 등 요건이 미반영되었거나 단순 선언적 문구만 존재하는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 운영 복잡성 인식: SLM은 '작은' 만큼 '쉽게 운영'되지 않는다. GPU 자원 관리·양자화·추론 엔진 등 운영 노하우가 외부 API 사용 대비 더 복잡할 수 있다.

2) 라이선스 사전 확인: Llama 라이선스(월 활성 사용자 7억 명 조건 등)를 공공 활용 시 반드시 확인하고, 법무 담당자의 모델 라이선스 검토 의견을 확보한다.

3) 한국어 특화 모델 우선 검토: 한국어 도메인 SLM은 'EXAONE', 'Solar', 'Polyglot-Ko' 등 한국어 특화 모델을 우선 검토하고, KoBEST 등 한국어 평가와 발주기관 도메인 골든셋 평가를 병행한다.

4) 추론 엔진 선택 기준: 추론 엔진은 처리량 우선이면 vLLM, 지연시간 우선이면 TensorRT-LLM 등 우선순위에 따라 선택하며, GPU 산정 시 KV Cache·동시성 영향을 반영하고 50% 안전마진을 확보한다.

5) 파인튜닝·양자화 검증: 파인튜닝 후 평가 없이 운영에 투입하거나 양자화 후 정확도 저하 측정을 누락하지 않도록 점검하고, 다중 모델 서빙 시 GPU 자원 경합 여부를 확인한다.

AG_SD42-1-19 | 모델추론서버 스케일링(Auto-scaling) 점검

시스템구조·보안 | 설계 | 시스템 구현 | 추론인프라

근거 | NIST AI RMF MEASURE · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 단계에서 도출된 AI 모델 추론 서버의 성능·확장성 요구사항이 {시스템아키텍처설계서}에 빠짐없이 상세·구체적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 생성형 AI 서비스는 사용자 요청 건당 추론(Inference) 연산 부하가 일반 웹 서비스 대비 수십~수백 배 높고 동시 접속 트래픽 변동 폭이 크기 때문에, 추론 서버의 수평·수직 스케일링(Scaling) 전략과 Auto-scaling 정책이 설계에 구체적으로 반영되지 않으면 서비스 가용성 저하 또는 과잉 비용 발생 위험이 높다.

본 가이드라인은 이러한 AI 고유의 인프라 설계 리스크를 감리원이 설계 산출물 검토를 통해 사전에 확인할 수 있도록 상세 검토 항목과 판정 기준을 제시한다.

2. 필요성

▶ 생성형 AI 추론 서버는 대규모 언어 모델(LLM) 또는 이미지 생성 모델 등을 서비스하기 위해 GPU/NPU 가속기 기반의 전용 인프라를 필요로 하며, 일반 WAS(Web Application Server)와는 본질적으로 다른 자원 소비 패턴을 보이므로, AI 추론 특화 스케일링 전략이 구체적으로 수립되어야 한다.

▶ 추론 지연시간(Latency) 민감성: 생성형 AI는 토큰 단위 순차 생성 방식으로 동작하여 응답 지연이 사용자 체감 품질(QoS)에 직접 영향을 미친다. 동시 요청 증가 시 GPU 메모리(VRAM) 포화로 인한 OOM(Out of Memory) 오류가 발생하면 서비스 전면 장애로 이어질 수 있으므로, 설계 시점에 구체적인 대응 방안이 마련되어야 한다.

▶ 비용 효율성과 자원 낭비 방지: GPU 자원은 일반 CPU 대비 시간당 비용이 10~50배 높다. 고정 할당 방식으로 설계하면 유휴 시간에도 과도한 비용이 발생하고, 반대로 과소 할당 시 서비스 중단 위험이 있다. Auto-scaling 정책을 통해 수요 기반 탄력적 자원 배분이 설계에 반영되어야 한다.

▶ AI 모델 교체·업데이트 대응: 모델 버전 업그레이드 또는 모델 크기 변경(예: 7B → 70B 파라미터) 시 필요 GPU 수량, 메모리, 네트워크 대역폭이 크게 달라지므로, 스케일링 설계가 모델 변경에 유연하게 대응 가능한 구조인지 확인이 필요하다.

▶ NIST AI RMF 정합성: NIST AI RMF의 MEASURE 2.6에서는 AI 시스템의 안전성을 평가할 때 시스템 신뢰성, 견고성, 실시간 모니터링 및 장애 시 응답 시간 등을 안전성 지표로 반영하도록 권고하고 있으며, 이는 추론 서버 스케일링 설계의 적정성과 직접적으로 연관된다.

3. 검토방법

No	점검 항목	확인 방법
1	추론 서버 스케일링 전략 기술 여부	Scale-out/Scale-up 방식·대상 컴포넌트·확장 단위 등 추론 서버 수평/수직 스케일링 전략이 {시스템아키텍처설계서}에 구체적으로 기술되었는지 확인
2	AI 특화 메트릭 기반 Auto-scaling 설계 여부	GPU 사용률·VRAM 사용률·요청 큐 깊이·토큰 처리 지연 등 AI 추론 특화 메트릭 기반으로 Auto-scaling 정책이 설계되었는지 'Auto-scaling 정책 설계서'·{시스템아키텍처설계서}에서 대조
3	인스턴스 수·임계값·쿨다운 수치 정의 여부	최소/최대 인스턴스 수·스케일링 임계값(Threshold)·쿨다운(Cooldown) 기간이 'Auto-scaling 정책 설계서'에 수치적으로 정의되었는지 확인
4	GPU/NPU 용량 산정 근거 여부	모델 파라미터 크기·배치 사이즈·동시 사용자 수 등 AI 워크로드 기반으로 GPU/NPU 용량 산정 근거가 '인프라 용량 산정서'에 산출되었는지 확인
5	모델 교체 시 스케일링 유연성 고려 여부	모델 크기 변경 시 대응 방안 등 모델 교체·업그레이드 시 스케일링 설계의 유연성이 {시스템아키텍처설계서} 확장성 섹션에 고려되었는지 확인
6	스케일링 실패 시 폴백 전략 포함 여부	스케일링 실패 시 폴백/Graceful Degradation 전략이 {시스템아키텍처설계서} 장애 대응 섹션에 포함되었는지 확인
7	성능·확장성 요구사항 추적 매핑 여부	성능·확장성 요구사항이 {요구사항추적표}를 통해 설계 항목에 빠짐없이 매핑되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	수평/수직 스케일링 전략, AI 추론 특화 메트릭 기반 Auto-scaling 정책, 최소/최대 인스턴스 수·임계값·쿨다운 기간의 수치 정의, AI 워크로드 기반 용량 산정 근거, 모델 변경 대응 방안 및 폴백 전략이 설계서에 구체적으로 반영되고 요구사항 추적성이 확보된 경우	현행 유지

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	스케일링 설계는 존재하나 CPU·메모리 등 일반 메트릭만 반영되어 AI 특화 메트릭이 누락되었거나, 임계값·인스턴스 수 등 정량적 기준이 미정의되었거나, 용량 산정 근거·폴백 전략의 구체성이 부족한 경우	해당 항목 보완 및 정량 기준 명시 요구 후 재확인
부적합 (X)	추론 서버 스케일링 전략 및 Auto-scaling 정책이 설계에 반영되지 않았거나, 'Auto-scaling을 적용한다' 수준의 선언적 문구만 존재하고 요구사항-설계 간 추적성이 확인되지 않는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 선언적 문구 vs. 구체적 설계 구분: 'Auto-scaling을 적용한다', '필요 시 서버를 증설한다'와 같은 선언적 문구만으로는 설계가 완료되었다고 볼 수 없다. 반드시 스케일링 트리거 메트릭, 임계값, 최소/최대 인스턴스 수, 쿨다운 기간 등이 수치적으로 정의되어 있어야 적정 판정이 가능하다.

2) AI 추론 메트릭과 일반 웹 메트릭의 차이 인식: CPU 사용률, 메모리 사용률만으로 Auto-scaling을 설계한 경우 AI 추론 환경에 부적합하다. GPU 사용률, VRAM 점유율, 추론 큐 깊이 등 AI 고유 메트릭 반영 여부를 반드시 확인해야 한다. 이는 일반 IT 감리 점검사항과 구별되는 AI 감리 고유의 핵심 포인트이다.

3) 용량 산정의 AI 워크로드 기반 여부 확인: '서버 N대 도입'이라는 결론만 있고 산정 근거가 없거나, 일반 웹 TPS(Transactions Per Second) 기준으로만 산정된 경우 부적합하다. 모델 파라미터 크기(B), 동시 사용자 수, 토큰 생성 속도, 배치 사이즈 등 AI 특화 변수를 기반으로 용량이 산정되어야 한다.

4) 클라우드 vs. 온프레미스 환경 차이: 클라우드 환경(AWS, Azure, GCP 등)은 Auto-scaling이 플랫폼 네이티브 기능으로 제공되지만, 온프레미스(물리 서버) 환경에서는 GPU 추가 확보에 수일~수주가 소요될 수 있다. 온프레미스 환경이라면 '예비 자원(Buffer) 확보 계획'이 설계에 포함되었는지 추가 확인이 필요하다.

5) 폴백 전략의 실효성 점검: '장애 시 대응한다'는 문구만 있고 구체적 방안이 없는 경우 보완권고 대상이다. 특히 GPU 웜업 시간(신규 인스턴스 기동에 3~5분 소요)을 감안한 트래픽 버퍼링 전략, 경량 모델 자동 전환(Graceful Degradation) 시나리오가 설계되었는지 확인한다.

AG_SD43-1-01 | 적대적 사례 공격방어 점검

시스템구조·보안 | 구현 | 인공지능 모델 개발 | AI특화보안

근거 | NIST AI RMF MEASURE · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조

1. 목적

본 가이드라인은 종료(시험/전개) 단계에서 AI 서비스에 대한 적대적 사례(Evasion) 공격 방어 기능이 실제로 구현되어 시험되었는지를 확인하는 것을 목적으로 한다.

구체적으로는 {취약점점검결과}와 'AI 보안시험 시나리오' 등 산출물을 통해, 변형된 이미지·텍스트·프롬프트를 이용한 회피 공격 시도에 대한 탐지 및 차단 기능이 측정 가능한 수치(방어 성공률 등)로 검증되었는지를 점검한다.

2. 필요성

▶ 생성형 AI 시스템은 사용자의 입력(프롬프트, 이미지 등)을 모델이 직접 처리하는 구조이므로, 공격자는 미세하게 변형된 입력을 통해 AI 모델의 판단을 의도적으로 오도하는 회피(Evasion) 공격을 시도할 수 있다. 이 공격은 겉으로 보기에 정상적인 입력처럼 보이지만, 모델 내부에서는 잘못된 응답이나 보안 우회를 유발한다.

▶ 특히 종료 단계는 설계 단계에서 정의된 방어 요건이 실제 시스템에서 구현·시험되었는지를 검증하는 마지막 관문으로, 이 단계에서 방어 기능이 확인되지 않으면 운영 중 심각한 보안 위협에 노출될 수 있다. NIST AI RMF는 AI 시스템의 보안·복원력에 대한 평가와 문서화를 명시적으로 요구(MEASURE 2.7)하고 있으므로, 방어 시험 결과의 정량적 문서화 여부까지 확인이 필요하다.

3. 검토방법

No	점검 항목	확인 방법
1	적대적 입력 방어 요건의 시험 계획 반영 여부	'AI 보안시험 시나리오'에 정의된 적대적 입력 방어 요건이 시험 항목으로 반영되었는지 확인
2	프롬프트 인젝션 공격 시험 수행 여부	직접·간접 프롬프트 인젝션 시나리오가 TC 단위로 기술되고 {취약점점검결과}에 수행 결과가 기재되었는지 확인하며, 시험 환경이 운영 환경과 동일한지 인터뷰로 확인
3	변형된 입력 이미지에 대한 방어 성공률 측정 여부	회피(Evasion) 공격 시험의 방어 성공률이 목표치와 함께 정량적으로 기재되었는지 {취약점점검결과}에서 확인
4	입력 필터링·정규화 기능 구현 및 시험 여부	입력 정규화·필터링 기능의 구현 여부와 시험 결과를 확인하고, 정상 입력에 대한 오탐률 측정 여부를 인터뷰로 확인

No	점검 항목	확인 방법
5	탐지 실패 시 폴백 처리 메커니즘 존재 여부	탐지 실패 건에 대한 2차 방어 또는 안전 응답 전환(폴백) 설계와 그 시험 결과를 확인
6	취약점 조치 결과 및 재점검 여부	발견된 취약점의 조치 내역과 재점검 결과가 {취약점점검결과}에 기재되고, 미조치 항목의 사유가 설명되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{취약점점검결과}에 AI 회피 공격(Evasion Attack) 항목이 명시적으로 포함되고, 프롬프트 인젝션(직접·간접) 시험이 'AI 보안시험 시나리오'에 TC 단위로 기술되었으며, 방어 성공률이 목표치 이상으로 달성되었음이 정량적으로 문서화되고, 입력 정규화·필터링 기능이 구현되어 정상 입력에 대한 오탐률이 측정되며, 탐지 실패 시 2차 방어 또는 안전 응답 전환(폴백) 메커니즘이 시험으로 확인되고 발견된 취약점의 조치 완료가 확인되는 경우	현행 유지
보완권고 (△)	일부 Evasion 시험 항목(예: 이미지 변형 시나리오)이 누락되었거나, 방어 성공률이 목표치 대비 미달인데 사유가 미기재되었거나, 탐지 실패 건에 대한 2차 방어 또는 보완 계획이 없는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	{취약점점검결과}가 존재하지 않거나 AI 특화 취약점 항목이 전혀 없고, 프롬프트 인젝션 시험이 수행되지 않았으며, 방어 성공률 등 정량적 지표가 전무하고 시험 자체가 미실시 상태인 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 회피 공격 시험의 수작업 검토 범위 명확화: {취약점점검결과}와 'AI 보안시험 시나리오'를 대조하여 Evasion 공격 항목이 계획-수행-결과까지 일관되게 문서화되었는지 확인하고, 자동화 도구 결과만으로 판단하지 말고 시험 시나리오 원본과 결과 로그를 직접 검토한다.

2) 방어 성공률 수치의 의미 해석: 사업자가 제시하는 방어 성공률 수치는 어떤 공격 세트(Test Set)로 측정했는가에 따라 의미가 달라진다. 단순히 수치의 크기만 확인하지 말고, 시험에 사용된 공격 시나리오의 다양성(유형 수, 난이도)과 목표치 설정의 합리성을 함께 확인해야 한다. 단일 공격 유형으로만 99% 방어 성공률을 달성한 경우 충분한 검증으로 보기 어렵다.

3) 프롬프트 인젝션 시험 항목 누락 여부 확인: 직접 인젝션뿐 아니라 외부 문서·검색 결과를 경유하는 간접 프롬프트 인젝션 시나리오가 시험 계획에 포함되었는지 확인한다. 시험 TC 목록에서 공격 유형별 건수를 집계하면 누락 여부를 빠르게 식별할 수 있다.

4) 탐지 실패 대비 폴백 확인: 필터가 공격을 탐지하지 못한 경우에 대비한 2차 방어(출력 검증, 안전모드 전환, 운영자 알림 등)가 구현·시험되었는지 확인한다. 방어 성공률이 100%가 아닌 이상 탐지 실패 시 처리 설계가 반드시 필요하다.

AG_SD43-1-02 | 모델저장소 비인가 접근차단 점검

시스템구조·보안 | 구현 | 시스템 구현 | 인증/인가/접근통제

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · OWASP LLM Top 10 · ISO/IEC 27001

1. 목적

AI 모델 저장소(Model Registry/Repository)는 학습된 모델 가중치(웨이트), 하이퍼파라미터, 토크나이저 등 AI 시스템의 핵심 지적재산이 집중되는 저장 공간이다.

본 가이드라인은 종료단계(시험/전개)에서 설계된 모델 저장소 접근통제 정책이 실제 시스템에 구현되었는지, 인가된 CI/CD 파이프라인과 관리자 IP만 접근이 허용되는지를 {보안점검결과} 및 관련 산출물을 통해 확인하는 것을 목적으로 한다.

특히, 모델 저장소에 비인가 접근이 발생할 경우 모델 가중치 탈취(모델 추출 공격), 모델 위/변조(백도어 삽입), 서비스 장애 유발 등 AI 시스템 고유의 중대한 보안 사고로 이어질 수 있으므로, 시험/전개 단계에서 이에 대한 구현 적정성을 점검하는 것이 핵심이다.

2. 필요성

▶ AI 모델 저장소는 기존 소프트웨어의 소스코드 저장소와 달리, 모델 가중치 파일이 유출되면 공격자가 동일한 AI 모델을 복제하거나, 취약점을 분석하여 적대적 공격(Adversarial Attack)을 정밀하게 수행할 수 있다. 또한 CI/CD 파이프라인은 모델의 빌드/테스트/배포를 자동화하는 핵심 경로이므로, 이 경로가 침해되면 위·변조된 모델이 자동으로 운영 환경까지 배포될 수 있다.

▶ 따라서 종료단계에서는 요구정의 및 설계 단계에서 수립한 모델 저장소 접근통제 정책이 실제 시스템에 적용되었는지, 보안점검을 통해 비인가 접근 경로가 실제로 차단되었는지를 검증해야 한다. 이는 일반적인 IT 보안점검을 넘어, AI 모델 자산의 기밀성과 무결성을 보장하기 위한 AI 특화 보안 점검이다.

3. 검토방법

No	점검 항목	확인 방법
1	모델 저장소 접근통제 구현	설계서상 RBAC/ACL 정책이 실제 시스템에 적용되었는지 보안점검결과에서 확인
2	CI/CD 파이프라인 접근 제한	모델 배포 파이프라인이 인가된 서비스 계정/에이전트만 실행 가능한지 확인
3	관리자 IP 화이트리스트	모델 저장소 관리 콘솔/API에 접근 가능한 IP가 화이트리스트로 제한되었는지 확인

No	점검 항목	확인 방법
4	보안점검 수행 적정성	모델 저장소 보안점검이 수행되었고, 발견된 취약점 조치가 완료되었는지 확인
5	모델 접근 감사로그 설정	모델 저장소 접근/수정/삭제 행위가 감사로그로 기록되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	모델 저장소에 RBAC/IP 제한이 설계대로 구현되고, CI/CD 파이프라인만 접근 가능하며, 모델 저장소 보안점검이 수행되어 발견된 취약점 조치가 완료되고, 감사로그가 정상 기록됨	현행 유지
보완권고 (△)	접근통제가 대부분 구현되었으나 일부 미흡이 있는 경우 (예: 감사로그에 모델명 누락)	해당 항목 보완 요청 후 재확인
부적합 (X)	모델 저장소에 비인가 접근이 가능하거나, 보안점검이 수행되지 않은 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 설계와 구현의 Gap 확인: 요구정의 및 설계 단계에서 수립된 접근통제 정책이 실제 시스템에 그대로 적용되었는지를 반드시 대조 확인해야 한다. 문서상 정책과 실제 설정이 다른 경우가 빈번하므로, 스크린샷 등 객관적 증빙을 확보한다.

2) 수동 배포 경로 차단 확인: CI/CD 파이프라인을 우회하여 개발자가 개인 계정으로 직접 모델을 Push할 수 있는 경로가 열려 있는지 확인한다. 이 경로가 존재하면 비인가된 모델 변경이 발생할 수 있어 실질적인 보안 위협이 된다.

3) 인증 토큰/키 관리 방식 확인: CI/CD 파이프라인에서 모델 저장소 접근에 사용하는 인증 토큰이 파이프라인 설정 파일에 평문으로 노출되어 있는지 확인한다. Vault, AWS Secrets Manager 등 비밀관리 도구를 통해 관리되는지가 핵심이다.

4) 모델 저장소 유형별 차이 인지: MLflow, Hugging Face Hub, S3/GCS 등 모델 저장소 유형에 따라 접근통제 구현 방식이 다르다. 저장소 유형별 접근통제 구현 방법을 사전에 파악하고, 해당 방식에 맞는 점검을 수행한다.

5) 감사로그 변조 방지 여부: 모델 저장소 접근 감사로그가 변조될 수 있는 환경(같은 서버에 저장, 관리자 수정 가능)인지 확인한다. 별도 로그 서버로 전송하거나 WORM(Write Once Read Many) 저장소에 보관하는 방식이 권장된다.

AG_SD43-1-03 | 모델·파일스토리지 암호화 점검

시스템구조·보안 | 구현 | 시스템 구현 | 암호화

근거 | 「개인정보 보호법」 · KCMVP · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · ISO/IEC 27001

1. 목적

AI 시스템의 모델 가중치 파일(.pt, .bin, .safetensors, .onnx 등) 및 학습/추론 관련 데이터가 저장된 파일 스토리지와 데이터베이스에 대해 물리적 암호화가 설계대로 구현되었는지 검증하고, 보안점검 결과에 그 이행 상황이 적절히 반영되어 있는지 확인하는 것이 목적이다.

종료단계에서는 요구정의/설계 단계에서 수립한 암호화 요구사항이 실제 시스템에 정확히 반영되었는지, 나아가 암호화 적용 이후 시스템이 정상적으로 동작하는지를 시험/검증하는 데 초점을 둔다.

2. 필요성

▶ AI 모델의 가중치 파일은 대규모 학습 데이터와 컴퓨팅 자원을 투입하여 생성된 핵심 지적재산이다. 이 파일이 암호화 없이 저장될 경우, 물리적 저장매체 탈취, 내부자 무단 복제, 백업 미디어 유출 등을 통해 모델 전체가 외부에 그대로 노출될 수 있다.

▶ 특히 생성형 AI 모델의 가중치 파일은 수십억~수천억 개의 파라미터를 포함하며, 이를 탈취한 공격자는 모델을 그대로 복제하거나 변조(Tampering)하여 악의적으로 활용할 수 있다. 또한 학습 데이터에 포함된 개인정보나 민감정보가 모델 파일을 통해 간접적으로 유출될 위험도 존재한다.

▶ 따라서 모델 가중치 파일 및 DB의 물리적 암호화 구현 상황을 검증하고, 보안점검결과에 해당 항목이 적절히 기록되어 있는지 확인하는 것이 필요하다.

3. 검토방법

No	점검 항목	확인 방법
1	모델 가중치 파일 암호화 구현	모델 가중치 파일(.pt, .bin, .safetensors 등)이 저장된 파일시스템/오브젝트 스토리지에 암호화가 적용되었는지 확인
2	DB 암호화 구현	학습 데이터, 파인튜닝 데이터, 임베딩 벡터 DB 등에 대한 컬럼/테이블스페이스 암호화 적용 확인
3	암호화 키 관리 체계	암호화 키 생성, 보관, 교체, 폐기 절차가 수립되고 이행되고 있는지 확인

No	점검 항목	확인 방법
4	암호화 적용 후 기능 검증	암호화 적용 후 모델 로딩/추론 성능에 영향이 없는지 시험결과 확인
5	보안점검결과 반영	{보안점검결과}에 암호화 상세 검토 항목 및 결과가 구체적으로 기록되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	모델 가중치 파일 및 DB에 대해 국제 표준 암호화 알고리즘(AES-256 등)이 적용되어 있고, 키 관리 체계가 수립되어 있으며, 암호화 적용 후 시스템 정상 동작이 검증되고, {보안점검결과}에 해당 항목이 구체적으로 기록되어 있는 경우	현행 유지
보완권고 (△)	암호화가 적용되어 있으나, 일부 파일 또는 DB에 미적용되었거나, 키 관리 절차가 미흡하거나, 보안점검결과 기록이 불충분한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	모델 가중치 파일 또는 DB에 암호화가 전혀 적용되지 않았거나, 키 관리 체계가 없거나, 보안점검결과에 해당 항목이 누락된 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 암호화 범위의 누락 확인: 모델 가중치 파일 외에도 체크포인트 파일, 옵티마이저 상태 파일, 토크나이저 파일, 설정 파일(config.json) 등도 민감한 정보를 포함할 수 있으므로 암호화 범위에 포함되어 있는지 확인한다.

2) 성능 영향 분석 확인: 암호화 적용 후 모델 로딩 시간이 과도하게 증가하면 서비스 운영에 지장이 발생할 수 있다. {테스트결과서}에 암호화 전/후 성능 비교 데이터가 포함되어 있는지 확인한다.

3) 클라우드 환경별 차이 주의: AWS S3의 SSE-S3/SSE-KMS, Azure Blob의 SSE, GCP Cloud Storage의 CMEK 등 클라우드 환경별로 암호화 방식이 다르므로, '보안 아키텍처 설계서'에 해당 환경에 맞는 암호화 방식이 명시되어 있는지 확인한다.

4) 키 유출 시나리오 확인: 암호화 키가 유출된 경우의 대응 절차(키 교체, 재암호화, 사고 보고 등)가 '암호화 키 관리 절차서'에 포함되어 있는지 확인한다.

5) 모델 버전별 암호화 일관성 확인: AI 시스템은 모델이 주기적으로 업데이트되므로, 신규 버전의 모델 파일에도 동일한 암호화 정책이 적용되는지 확인한다. 모델 레지스트리에 암호화 상태가 관리되고 있는지도 함께 점검한다.

AG_SD43-1-04 | 시스템 보안약점 예방필터 계층점검

시스템구조·보안 | 구현 | 시스템 구현 | 시큐어코딩

근거 | OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

AI 서비스 구축 사업에서, 정의된 보안약점 예방 필터 계층이 실제 시스템에 적정하게 구현되었는지, 그리고 해당 보안 메커니즘에 대한 시험(검증)이 충분히 수행되었는지를 확인하는 것이 목적이다.

특히, 전통적인 웹 공격(SQLi, XSS 등)에 대한 방어뿐만 아니라, AI 서비스 특성상 발생 가능한 프롬프트 인젝션(Prompt Injection), 탈옥(Jailbreak) 공격 등 AI 고유의 입력값 조작 위협에 대한 필터링 계층이 AI 특화 게이트웨이(AI Gateway) 또는 애플리케이션 레벨의 입출력 가드레일(Input/Output Guardrails)에서 실제로 차단 동작을 하는지를 검증한다.

2. 필요성

▶ AI 서비스는 사용자 입력을 기반으로 동적 응답을 생성하는 구조적 특성으로 인해, 기존 웹 애플리케이션 대비 공격 표면(Attack Surface)이 확장된다. 전통적인 SQL Injection, Cross-Site Scripting(XSS) 공격은 WAF(웹 애플리케이션 방화벽) 및 시스템 레이어에서 차단할 수 있으나, 프롬프트 인젝션·탈옥 등 AI 고유 공격은 AI 특화 필터 계층이 없으면 차단할 수 없다.

▶ 종료 단계에서는 요구정의/설계 단계에서 정의한 보안약점 예방 필터 계층이 실제로 구현되었는지, 시험이 수행되어 차단 효과가 입증되었는지를 검증해야 한다. 이를 통해 배포/전개 전 보안 결함을 사전에 식별하고, 운영 환경에서 발생 가능한 보안 사고를 예방할 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	웹 공격 차단 구현 검증	SQLi, XSS 등 주요 웹 공격에 대한 WAF 및 시스템 레이어 차단 규칙이 설계대로 구현되었는지 확인
2	AI 전용 필터링 구현 검증	프롬프트 인젝션/탈옥 공격에 대한 입력값 필터링 모듈이 AI 서비스 계층에 실제로 적용되어 동작하는지 검증
3	보안 시험 수행 적정성	보안 시험 계획서 대비 실제 시험 수행 결과가 충분하고 필터 계층별 차단 시험이 포함되었는지 확인
4	시큐어코딩 점검 결과 확인	AI 서비스 관련 소스코드에 대한 보안약점 점검 도구 수행 결과 및 조치 완료 여부 확인

No	점검 항목	확인 방법
5	전개 전 보안 확인 절차	'전개 계획서'에 배포 전 보안 확인 절차(차단 테스트 통과 기준, 취약점 조치 완료 기준 등)가 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	보안약점 예방 필터 계층(WAF·시스템 레이어의 웹 공격 차단, AI 서비스 계층의 프롬프트 인젝션·탈옥 입력 필터링)이 설계대로 구현되고, 보안 시험이 계층별 차단 시험을 포함하여 충분히 수행되어 차단 효과가 입증되며, AI 서비스 관련 소스코드의 시큐어코딩 점검 결과와 조치가 완료되고, 전개 계획서에 배포 전 보안 확인 절차가 명시된 경우	현행 유지
보완권고 (△)	구현은 되었으나, 일부 계층의 시험이 불충분하거나 경미한 미비 사항이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	필터 계층이 미구현되거나, 보안 시험이 전혀 수행되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 전통적 웹 공격과 AI 고유 공격의 구분 필요: SQLi/XSS와 같은 전통적 웹 공격에 대한 차단 여부와, 프롬프트 인젝션/탈옥 등 AI 고유 공격에 대한 필터링 여부는 별개의 상세 검토 항목으로 구분하여 확인해야 한다. WAF에서 SQLi/XSS가 차단된다고 해서 프롬프트 인젝션까지 차단되는 것은 아니다.

2) 시험 충분성 판단 기준: 보안 시험 수행 여부만 확인하지 말고, 테스트 케이스의 다양성과 건수를 함께 판단해야 한다. 예를 들어, 프롬프트 인젝션 시험이 단 1건의 시나리오로만 수행된 경우는 충분한 검증으로 보기 어렵다.

3) 수작업과 도구 활용 점검의 구분: 시큐어코딩 점검 도구의 결과는 도구 리포트 원본과 조치 이력으로 확인하고, AI 필터 계층의 차단 동작은 실제 공격 문자열 입력 시연 등 수작업 검증으로 교차 확인한다. 형식적 문서 확인에 그치지 말고 실질적 이행 근거와 책임자 확인 여부까지 점검해야 한다.

AG_SD43-1-05 | 파일업로드 악성코드 방지점검

시스템구조·보안 | 구현 | 시스템 구현 | 시큐어코딩

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · OWASP LLM Top 10

1. 목적

생성형 AI 서비스가 사용자로부터 파일(이미지, 문서 등)을 업로드 받아 AI 모델이 분석하는 과정에서, 업로드된 파일에 포함된 악성코드가 실행되거나 이미지/문서 분석 과정에서 악성 스크립트가 트리거되는 보안 위험을 방지하기 위한 보안 통제가 설계대로 구현되고 시험되었는지를 점검하는 것이 목적이다.

특히, 기존 웹 애플리케이션과 달리 AI 모델이 파일 콘텐츠를 직접 해석하는 구조에서는, 악성 파일이 모델 추론 파이프라인 또는 서버 환경 전체를 위협할 수 있으므로, 이에 대한 방어 및 통제 메커니즘의 구현 적정성을 확인한다.

2. 필요성

▶ 생성형 AI 서비스는 사용자가 업로드한 이미지나 문서를 AI 모델이 직접 파싱하여 분석하는 기능을 제공하는 경우가 많다. 이 과정에서 다음과 같은 AI 고유의 보안 위험이 발생한다.

▶ 첫째, 업로드된 파일에 삽입된 악성코드(Macro, JavaScript, Shellcode 등)가 AI 모델의 전처리 (preprocessing) 단계에서 실행될 수 있다.

▶ 둘째, 이미지 파일(SVG, TIFF 등)에 포함된 악성 스크립트가 이미지 디코딩 라이브러리의 취약점을 통해 서버 측에서 실행될 수 있다.

▶ 셋째, PDF/DOCX 등 문서 파일에 내장된 스크립트나 외부 링크가 AI 서버 환경의 보안을 우회할 수 있다.

▶ 이러한 위험은 기존 웹 애플리케이션의 파일 업로드 보안과는 차원이 다른 문제로, AI 모델이 파일 콘텐츠를 직접 해석하는 특성상 공격 표면이 확장되므로, 설계된 보안 통제가 실제로 구현되어 작동하는지를 시험/전개 단계에서 확인할 필요가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	파일 업로드 보안 통제 구현	허용 파일 유형 제한(화이트리스트), 파일 용량 제한, 확장자 검증, MIME 타입 검증 등이 실제 구현되었는지 확인
2	악성코드 스캔 구현	업로드 파일에 대한 악성코드/바이러스 스캔 기능이 실제 작동하는지 확인 (AV 엔진 연동 또는 정적분석 구현)

No	점검 항목	확인 방법
3	이미지 파일 악성스크립트 차단	SVG 내 JavaScript, TIFF/EXIF 메타데이터 내 악성코드, 이미지 디코딩 라이브러리 취약점 방어 구현 확인
4	문서 파일 악성스크립트 차단	PDF/DOCX/스프레드시트 내 매크로/스크립트/외부링크 제거 또는 비활성화 구현 확인
5	샌드박스/격리 환경 구현	파일 분석 프로세스가 격리된 환경(샌드박스, 컨테이너 등)에서 실행되는지 구현 확인
6	보안시험 수행 적정성	악성파일 업로드 시나리오에 대한 보안시험 케이스가 수행되고 결과가 문서화되었는지 확인
7	배포/전개 시 보안통제 적용	배포 환경에 파일 업로드 보안통제가 적용된 상태로 전개되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	파일 업로드 보안통제(확장자/용량/MIME 검증, 악성코드 스캔, 이미지/문서 악성스크립트 차단, 샌드박스 격리)가 설계대로 구현되어 시험이 완료되고, 배포 환경에 해당 보안통제가 적용된 상태로 전개됨이 확인된 경우	현행 유지
보완권고 (△)	주요 보안통제는 구현되었으나, 일부 항목(예: 이중확장자 차단, EXIF 스트리핑 등)이 미흡하거나 시험 커버리지가 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	파일 업로드 보안통제가 구현되지 않았거나, 악성파일이 차단되지 않고 AI 모델에 전달되는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 고유 공격 벡터에 주목: 파일 업로드 보안을 점검할 때, 기존 웹 애플리케이션의 파일 업로드 보안(확장자 차단, 용량 제한 등)만으로 충분하지 않다. AI 모델이 파일 콘텐츠를 직접 해석하는 특성상, 이미지 내 스크립트(SVG), 문서 내 매크로(DOCX/PDF), 메타데이터 내 페이로드(EXIF/TIFF) 등 AI 고유의 공격 벡터에 대한 방어가 구현되어 있는지 확인해야 한다.

2) 실제 시험 결과 확인 필수: 설계서에 '악성코드 스캔 적용'이라고 기술되어 있더라도, {통합시험결과서}에서 EICAR 테스트 파일, 악성 SVG, 악성 매크로 포함 DOCX 등을 활용한 실제 보안시험 결과가 문서화되어 있는지 반드시 확인해야 한다. '구현만 되고 시험되지 않은 보안통제'는 실효성을 보장할 수 없다.

3) 도구 활용 시 주의점: 파일 업로드 차단 기능을 직접 테스트할 때는, 사전에 사업자와 협의하여 테스트 환경에서 진행해야 한다. 운영 환경에 직접 악성파일을 업로드하는 행위는 보안 사고를 유발할 수 있으므로, 반드시 개발/스테이징 환경에서 수행해야 한다.

4) 배포 전 보안통제 적용 확인: 개발/스테이징 환경에서 적용된 보안통제가 운영 환경 배포 시에도 동일하게 적용되는지 '배포/전개 계획서'의 체크리스트를 통해 확인한다. CI/CD 파이프라인에서 보안 설정이 코드화(IaC)되어 있으면 환경 간 일관성을 보다 신뢰할 수 있다.

5) 인터뷰 시 핵심 확인 포인트: 사업자에게 '업로드된 파일이 AI 모델에 전달되기 전에 어떤 검증을 거치는가?'라는 질문으로 시작하면, 파일 업로드 보안 통제의 전체적인 구현 수준을 파악할 수 있다. 단순히 '화이트리스트 적용했다'는 답변보다는, 구체적인 검증 흐름(확장자 → MIME → 매직바이트 → 악성코드스캔 → 콘텐츠정화 → 샌드박스 전달)을 설명할 수 있는지 확인해야 한다.

AG_SD44-1-01 | 민감정보 유출 차단점검

시스템구조·보안 | 시험 | 시스템 구현 | 개인정보보호

근거 | 「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)·「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)·OWASP LLM Top 10

1. 목적

생성형 AI 시스템에서 사용자가 프롬프트를 통해 주민등록번호, 여권번호, 신용카드번호 등 민감정보(개인식별정보, PII)를 입력하는 경우, 해당 정보가 외부 AI 모델 서버로 전송되거나 응답 데이터에 포함되어 유출되는 것을 방지하기 위한 DLP(데이터 유출 방지) 솔루션 연동이 설계대로 구현되었는지 확인하고, 실제 시험을 통해 차단/경고 기능이 정상 작동하는지 검증하는 것을 목적으로 한다.

특히, 생성형 AI 서비스는 사용자 입력값이 외부 LLM API로 직접 전송되는 구조적 특성이 있어, 기존 정보시스템 대비 민감정보 유출 경로가 다양하고 위험이 크므로, 종료단계에서 이에 대한 구현 적정성과 시험 완료성을 점검한다.

2. 필요성

▶ 생성형 AI 시스템에서는 사용자가 프롬프트 입력창에 자유롭게 텍스트를 입력할 수 있으며, 입력된 내용이 외부 AI 모델(LLM) 서버로 전송되는 구조이므로, 사용자가 부주의하게 주민등록번호, 여권번호, 신용카드번호 등을 포함하여 전송할 경우 심각한 개인정보 유출 사고로 이어질 수 있다.

▶ 또한, AI 모델의 응답에도 학습 데이터에 포함된 민감정보가 복원되어 출력될 가능성이 있으며, 이는 개인정보 보호법 위반 및 공공기관 신뢰 훼손으로 직결된다. 따라서 프롬프트 송수신 구간 전반에 걸친 DLP 연동과 민감정보 필터링 기능이 설계대로 구현되었는지, 그리고 실제 주민번호 등을 입력했을 때 차단 또는 경고가 정상 작동하는지를 시험 결과로 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	DLP 연동 구현 확인	설계서상 DLP 솔루션 연동 아키텍처가 실제 시스템에 구현되었는지 확인
2	민감정보 탐지 패턴 적용 확인	주민등록번호, 여권번호, 신용카드번호 등 개인식별정보 탐지 정규식/패턴 적용 여부 확인
3	프롬프트 입력 차단/경고 시험	주민번호 포함 프롬프트 전송 시 차단 또는 경고 메시지가 정상 작동하는지 시험 결과 확인

No	점검 항목	확인 방법
4	AI 응답 출력 필터링 확인	AI 모델 응답에 포함된 민감정보가 마스킹/제거되어 사용자에게 노출되지 않는지 확인
5	차단 로그 및 이벤트 기록 확인	DLP 차단/경고 이벤트가 로그로 기록되고, 관리자 알림 체계가 구현되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	DLP 솔루션이 설계대로 연동되어 있고, 주민번호 포함 프롬프트 전송 시 차단/경고가 정상 작동하며, 입출력 양방향 필터링과 로그 기록이 모두 확인된 경우	현행 유지
보완권고 (△)	DLP 연동은 구현되었으나, 일부 변형 패턴(공백 변형 등) 탐지 누락, 출력측 필터링 미적용, 또는 로그 기록 불충분 등 부분적 미비사항이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	DLP 연동이 구현되지 않았거나, 주민번호 포함 프롬프트가 차단 없이 외부 AI 모델로 전송되는 경우, 또는 시험 결과가 존재하지 않는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 입력측과 출력측을 분리하여 점검할 것: DLP 연동은 프롬프트 입력(사용자 → AI 모델) 구간과 AI 응답(모델 → 사용자) 구간을 모두 포함해야 한다. 입력측만 차단하고 출력측을 간과하는 경우, 학습 데이터에 포함된 민감정보가 응답을 통해 유출될 수 있다.

2) 테스트 시나리오의 변형 패턴 포함 여부를 확인할 것: 주민번호의 공백 삽입, 하이픈 삭제, 특수문자 치환 등 변형 패턴을 통해 DLP 탐지를 우회할 수 있다. 테스트 시나리오에 이러한 변형 사례가 포함되어 있는지 확인하는 것이 중요하다.

3) DLP 차단 로그 자체의 개인식별정보 보호 여부 확인: 차단 이벤트 로그에 탐지된 민감정보 원문이 평문으로 기록되면 로그가 2차 유출 경로가 될 수 있으므로, 로그 저장 시 마스킹 처리와 로그 접근 통제가 적용되어 있는지 확인한다.

AG_SD44-1-02 | 성능 평가지표 달성여부

시스템구조·보안 | 시험 | 인공지능 모델 개발 | 모델평가

근거 | NIST AI RMF MEASURE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업에서 설계 시 정의한 성능평가 지표(F1-score, 정밀도, 재현율 등)가 {테스트결과서}에 실제 측정값으로 기록되었는지, 목표 임계값을 충족하였는지, 그리고 미달 시 조치 여부를 확인하는 것이 목적이다.

감리원은 AI 모델의 성능을 직접 평가하는 것이 아니라, 성능목표가 관련 산출물에 일관되고 측정 가능하게 문서화되었는지와 목표 달성 여부 및 후속 조치의 적정성을 제3자 관점에서 검토한다.

목표 임계값은 자의적 설정이 불가하며, 반드시 발주기관의 사전 승인을 받은 '성능평가 기준서' 또는 요구사항 명세서를 근거로 해야 함을 점검 기준에 포함해야 한다.

2. 필요성

▶ AI 모델은 일반 소프트웨어와 달리 "예측 오류"가 내재적으로 존재한다. F1-score 등 성능지표가 목표값을 충족하였는지를 검증하지 않으면, 배포 이후 서비스 품질 분쟁이 발생할 경우 사업자·발주기관 모두 기준점이 없어 분쟁 해결이 어렵다.

▶ 특히 다음과 같은 사유로 성능평가 지표 점검이 필수적이다.

▶ 목표 지표 미설정 또는 모호한 목표: "적절한 정확도"처럼 측정 불가능한 기준은 검사(인수) 단계 분쟁의 원인이 된다.

▶ 테스트 데이터 편향 위험: 목표를 충족하기 쉬운 데이터만으로 테스트하여 실제 운영 환경과 괴리가 발생할 수 있다.

▶ 성능지표 선택의 적정성: 분류 문제에서 단순 Accuracy만 사용할 경우 불균형 데이터 환경에서 왜곡된 성능 수치가 나타날 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	성능 목표 지표 임계값 명시 여부	'F1-score ≥ 0.85' 등 측정 가능한 성능 목표 지표가 수치화된 임계값과 함께 {테스트결과서}에 명시되었는지 확인

No	점검 항목	확인 방법
2	테스트 데이터셋 분리·대표성 문서화 여부	성능평가 테스트 데이터셋이 학습 데이터와 분리되어 운영 환경을 대표하는지 'AI 성능평가 기준서' 또는 {테스트결과서} 데이터셋 명세에서 학습/테스트 분리 방식 확인
3	목표 대비 실제 측정값 항목별 기재 여부	F1-score 등 목표 지표 대비 실제 측정값이 항목별로 {테스트결과서} 성능측정 결과표에 기재되었는지 확인
4	임계값 미달 항목 조치 계획 문서화 여부	성능 임계값 미달 항목에 대한 조치 계획(재튜닝·임계값 재협의·발주기관 승인 등)이 {테스트결과서} 또는 'AI 성능평가 보고서' 미달 처리 섹션에 문서화되었는지 확인
5	AI 서비스 유형별 적합 지표 선택 여부	분류·생성·추천 등 해당 AI 서비스 유형에 적합한 성능 지표가 선택되었는지 'AI 성능평가 기준서' 지표 선정 근거에서 확인(생성형 AI의 경우 BLEU/ROUGE/사람 평가 등)
6	성능평가 수행 주체 독립성 명시 여부	성능평가 수행 주체(개발팀/QA팀/제3자 등)가 명시되어 독립성이 확인되는지 'AI 성능평가 보고서' 또는 {테스트결과서} 수행 주체란에서 확인
7	성능평가 재현성 정보 기록 여부	성능평가 결과의 재현성 확보를 위해 테스트 환경·버전·파라미터 등이 {테스트결과서} 테스트 환경·버전 명세 섹션에 기록되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	발주기관이 승인한 수치화된 성능 목표가 명시되고, 해당 AI 서비스 유형(분류·생성·추천 등)에 적합한 지표가 선정 근거와 함께 채택되었으며, 목표값 대비 실제 측정값이 항목별로 기재되고, 미달 항목의 조치 계획, 성능평가 수행 주체의 독립성, 테스트 환경·버전 등 재현성 정보가 모두 문서화되어 있는 경우	현행 유지
보완권고 (△)	성능 측정은 수행되었으나 목표 임계값의 근거 불명확, 데이터셋 분리 명세 누락, 미달 항목 조치 계획 미흡 등 일부 기록이 불완전한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	성능 목표가 수치로 정의되지 않았거나 측정 결과가 {테스트결과서}에 기록되지 않아 목표 달성 여부를 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 성능 목표의 수치화 확인: "AI는 사용자의 질문에 잘 답변해야 한다"와 같은 정성적 기준은 인수검사 단계에서 분쟁의 씨앗이 된다. {테스트결과서}에 "F1-score 0.85 이상", "응답 적절성 사람

평가 4.0/5.0 이상" 등 측정 가능한 수치 목표가 명시되어 있는지 반드시 확인하고, 없거나 정성적으로만 기술된 경우 감리 지적사항으로 기록하고 수치화를 요구한다.

2) 생성형 AI의 경우 자동 지표만으로 불충분: BLEU/ROUGE 등 자동 지표만으로는 실제 답변 품질을 충분히 반영하지 못하므로, 'AI 성능평가 보고서'에 도메인 전문가 또는 사용자가 참여한 사람 평가(Human Evaluation) 결과가 포함되어 있는지 확인한다. 사람 평가의 경우 평가자 수, 평가 기준, 점수 분포 등이 명시되어야 한다. 성능 지표는 전체 평균뿐 아니라 소수 클래스·엣지 케이스 기준도 포함하여 점검한다.

3) 미달 항목의 사후 처리 이력 확인: 성능 미달 항목에 대해 재튜닝 후 재측정 결과, 또는 발주기관과의 임계값 재협의·승인 문서가 존재하는지 확인한다. 조치 근거 없이 미달 상태로 종결된 경우 감리 지적사항으로 기록한다.

AG_SD44-1-03 | 시스템 목표용량 점검

시스템구조·보안 | 시험 | 시스템 구현 | 모델평가

근거 | NIST AI RMF MEASURE ·「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.) ·「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업의 종료단계(시험/전개)에서, 설계 시 사전 정의된 BMT(Benchmark Test) 목표치—TPS(Transactions Per Second), 동시 접속 사용자 수(Concurrent Users)—를 실제 시스템이 100% 충족하는지를 검증하는 것이 목적이다.

AI 시스템은 일반 웹 애플리케이션과 달리, LLM 추론(Inference) 과정에서 GPU 메모리 점유, 모델 로딩 지연, 토큰 생성 시간 등 고유한 성능 병목이 존재하므로, 기존 웹/WAS 중심의 부하테스트만으로는 시스템 용량의 적정성을 판단할 수 없다.

본 가이드라인은 AI 추론 파이프라인의 성능 특성을 반영한 용량 증명 절차가 적정하게 수행되었는지를 확인하는 데 초점을 둔다.

2. 필요성

▶ AI 서비스는 추론 요청 시 대규모 GPU 연산을 수반하며, 모델 크기와 입력 데이터 특성에 따라 응답 시간이 크게 변동할 수 있다. 특히 생성형 AI는 토큰을 순차적으로 생성(Auto-regressive)하므로, 동시 접속자 수 증가 시 GPU 메모리 부족, 큐 대기(Queue Waiting), 응답 지연 등이 급격하게 발생할 수 있다.

▶ 따라서 요구정의 단계에서 설정한 목표 용량(Capacity)이 실제 AI 추론 환경에서 충족되는지를 시험/전개 단계에서 객관적으로 증명하지 않으면, 운영 개시 후 성능 장애, SLA(Service Level Agreement) 위반, 대규모 사용자 불만 등이 발생할 위험이 높다. 감리원은 BMT 목표치의 정의, 시험 수행 절차, 결과 증빙의 적정성을 종합적으로 점검해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	BMT 목표치 정의 적정성	요구사항 대비 BMT 목표(TPS, Concurrent Users, 응답시간)가 정량적으로 정의되어 있는지 확인
2	AI 고유 성능지표 포함 여부	추론 지연시간(Latency), 토큰 처리량(Tokens/sec), TTFT, GPU 사용률 등 AI 특화 지표가 BMT 항목에 포함되었는지 확인
3	테스트 환경 동등성 확인	BMT 수행 환경이 운영 환경(GPU 종류/수량, 모델 버전, 배치 크기)과 동일 또는 동등 조건인지 확인

No	점검 항목	확인 방법
4	부하 시나리오 현실성	실제 서비스 사용 패턴(프롬프트 길이 분포, 동시 세션 수, 피크 시간대 등)을 반영한 시나리오인지 확인
5	BMT 목표 달성 증빙	사전 정의된 TPS, Concurrent 목표 100% 달성 여부를 결과 데이터로 증빙하였는지 확인
6	스트레스/한계 테스트 수행	목표치 초과 부하(예: 150~200%) 시 시스템의 안정적 degradation 또는 장애 발생 패턴을 확인하였는지 검토
7	테스트 도구 및 방법론 적정성	사용된 부하테스트 도구(JMeter, Locust, k6 등)와 AI 추론 부하 생성 방법의 적합성 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	사전 정의된 BMT 목표(TPS, Concurrent) 100% 달성이 정량적으로 증빙되고, AI 고유 성능지표가 포함되어 있으며, 테스트 환경이 운영 환경과 동등한 조건에서 수행되고, 부하 시나리오가 실제 서비스 사용 패턴(프롬프트 길이 분포·동시 세션 수·피크 시간대)을 반영하며, 사용된 부하테스트 도구와 AI 추론 부하 생성 방법의 적합성이 확인된 경우	현행 유지
보완권고 (△)	BMT 목표는 달성하였으나, AI 고유 성능지표가 일부 누락되었거나, 테스트 환경과 운영 환경 간 차이가 존재하거나, 스트레스 테스트가 미수행된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	BMT 목표(TPS, Concurrent) 100% 미달이거나, 성능테스트 자체가 미수행되었거나, 결과 증빙이 없는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 성능지표 누락 확인: BMT 결과에 기존 웹 시스템 지표(TPS, 응답시간)만 있고 AI 고유 지표(추론 지연시간, 토큰 처리량, GPU 사용률 등)가 누락된 경우가 빈번하다. '토큰 처리량', 'TTFT', 'GPU 메모리 사용률' 등의 지표가 BMT 항목에 포함되어 있는지 반드시 확인한다.

2) 환경 차이 함정 주의: BMT 수행 환경과 운영 환경이 다른 경우(GPU 종류 변경, 모델 버전 업데이트 등), 사업자가 '비슷한 환경이므로 문제없다'고 주장할 수 있다. 이 경우 환경 차이에 대한 영향 분석 보고서 및 보정 계수 적용 근거를 요청한다.

3) 모호한 목표치의 수치화 유도: '충분한 성능을 확보한다' 등 정성적 표현만 있는 경우, TPS·동시 사용자 수·응답시간 상한 등 측정 가능한 수치로 목표를 구체화하도록 유도하고, 목표 미달 시의 조치 계획(증설, 튜닝 등)이 수립되어 있는지 확인한다.

AG_SD44-1-04 | 안전모드 이중화 전환점검

시스템구조·보안 | 시험 | 시스템 구현 | 가용성/이중화

근거 | NIST AI RMF MEASURE · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계된 안전모드(Fail-safe) 이중화 전환 기능이 시스템으로 적정하게 구현되었는지, 장애 발생 시 이중화 서버로의 정상 전환이 설계 시 정의된 목표 복구 시간(RTO) 내에 완료되는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, 생성형 AI 시스템은 추론 서버의 GPU 장애, 모델 로딩 실패, 이상 추론 응답 등 일반 웹 서비스에서 발생하지 않는 AI 고유의 장애 유형이 존재하며, 이에 대한 폴백 모델/서버 전환 및 안전모드 운영 메커니즘이 시험을 통해 검증되었는지를 확인한다.

2. 필요성

▶ 생성형 AI 시스템은 일반 웹 서비스와 달리 추론 서버의 GPU 메모리 부족(OOM), 모델 가중치 손상, 환각(Hallucination) 폭발 등 AI 고유의 장애 유형이 존재하며, 장애 발생 시 전체 서비스 중단으로 이어질 위험이 높다. 이에 대한 이중화 전환 및 안전모드 운영이 실제로 동작하는지 검증할 필요가 있다.

▶ AI 추론 서버 장애의 특수성: GPU OOM, CUDA 오류, 모델 체크포인트 손상 등 일반 서버 장애와 다른 특수한 장애 유형에 대해 Failover가 정상 작동하는지 시험으로 검증해야 한다.

▶ 폴백 모델 응답 품질 보장: 장애 시 전환되는 폴백 모델이 최소한의 서비스 품질(응답 정확도, 응답 시간)을 보장하는지 시험 결과로 확인해야 한다.

▶ 전환 시간 목표 충족: 설계에서 명시한 이중화 전환 목표 시간(예: 3초 이내)이 실제 시험에서 달성되었는지 결과 데이터로 확인해야 한다.

▶ 서비스 연속성 보장: NIST AI RMF MEASURE 2.6에서 AI 시스템이 안전하게 실패(fail safely)할 수 있어야 하며, 장애 응답 시간이 측정되어야 함을 명시하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	안전모드 구현·Failover 시험 수행 여부	안전모드(Fail-safe)가 설계대로 구현되어 Failover 시험이 수행되었는지 {테스트결과서}에서 확인
2	이중화 서버 목표 시간 내 전환 검증 여부	장애 발생 시 이중화 서버로 목표 시간(예: 3초) 이내 전환됨을 {테스트결과서} 전환 시간 측정값으로 검증

No	점검 항목	확인 방법
3	폴백 모델·서버 최소 응답 품질 시험 여부	폴백 모델·서버의 최소 응답 품질(정확도·응답시간)이 'AI 추론 서버 이중화 시험 결과서'에서 시험되었는지 확인
4	AI 장애 유형별 전환 절차 시험 여부	GPU OOM·모델 실패·환각 폭발 등 AI 장애 유형별 전환 절차가 구분되어 '장애 대응 절차서'·{테스트결과서}에서 시험되었는지 대조
5	전환 시 사용자 안내·서비스 제한 시험 여부	안전모드 전환 시 사용자 안내 메시지·서비스 제한 정책 적용이 {테스트결과서}·화면 캡처에서 시험되었는지 확인
6	정상 복구(Failback) 절차 시험 여부	정상 복구(Failback) 절차가 시험되어 원래 모델·서버로 정상 복귀됨이 {테스트결과서} Failback 시험 항목에서 확인되는지 확인
7	안전모드 운영 절차·전환 기준 명시 여부	안전모드 운영 절차·전환 기준(임계값)이 {전개계획서}에 명시되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	안전모드 이중화 전환이 설계대로 구현되어 Failover/Failback 시험이 완료되었고, 전환 시간이 목표(RTO) 내 달성되었으며, AI 고유 장애 유형별 전환 절차, 전환 시 사용자 안내 메시지·서비스 제한 정책 적용, 폴백 모델 품질이 시험 결과로 증빙된 경우	현행 유지
보완권고 (△)	이중화 전환 기능은 구현되었으나 일부 AI 장애 유형 시험 누락, 폴백 모델 품질 기준 미정의, 전개계획서 내 안전모드 운영 절차·임계값 미비 등 부분적 미흡이 있는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	안전모드 이중화 전환 기능이 미구현되었거나 Failover 시험이 수행되지 않은 경우, 또는 전환 시간이 목표를 충족하지 못한 채 방치된 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 전환 시간 목표치의 구체성 확인: '이중화 전환을 지원한다'와 같은 모호한 문구가 아닌, '추론 서버 장애 발생 시 3초 이내 예비 추론 서버로 자동 전환'과 같이 측정 가능한 기준으로 명시되어 있는지 확인해야 한다. 모호한 기준은 분쟁의 원인이 된다.

2) AI 고유 장애와 일반 장애 구분: 일반 서버 Failover(프로세스 다운, OS 행, 네트워크 단절 등)는 기존 IT 감리에서 다루므로, 본 가이드라인에서는 GPU OOM, 모델 로딩 실패, 환각 폭발 감지, 추론 Timeout 등 AI 고유 장애에 대한 전환 시험 여부에 중점을 두어야 한다.

3) 시험 재현 요청 활용: {테스트결과서} 문서 검토만으로는 전환 시간 목표 달성 여부를 판단하기 어려운 경우, 사업자에게 시험 재현을 요청하여 실제 전환 과정을 확인할 수 있다. 이는 감리 기준에서 허용하는 현장 확인 방법이다.

4) 폴백 모델 품질 기준의 사전 합의: 폴백 모델은 주 모델 대비 응답 품질이 낮을 수 있으므로, 허용 가능한 최소 품질 기준(정확도, 응답 시간 등)이 합의되어 있는지 확인해야 한다. 기준 없이 폴백 모델을 운영하면 서비스 품질 분쟁이 발생할 수 있다.

5) 전개계획서와 시험결과서의 일관성: {전개계획서}의 안전모드 운영 절차와 {테스트결과서}의 시험 항목이 서로 일치하는지 대조 검토해야 한다. 전개계획서에 명시된 전환 임계값이 실제 시험에서 적용되었는지 확인하는 것이 핵심이다.

AG_SD44-1-05 | API 게이트웨이 인증_인가 통제

시스템구조·보안 | 시험 | 시스템 구현 | 인증/인가/접근통제

근거 | OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계된 API 게이트웨이의 인증(Authentication) 및 인가(Authorization) 통제 메커니즘이 구현된 시스템에 정상적으로 반영되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 추론(Inference) API는 일반 웹 API와 달리 모델 추출(Model Extraction) 공격, 프롬프트 인젝션(Prompt Injection) 경유 토큰 탈취, 대량 추론 요청을 통한 비용 폭증(Cost Explosion) 등 AI 고유의 보안 위협에 노출되어 있다.

따라서, 유효하지 않은 JWT(JSON Web Token) 또는 만료된 토큰으로 AI 추론 API를 호출했을 때 HTTP 401(Unauthorized) 응답으로 정확히 차단되는지, 인가되지 않은 역할(Role)이 고위험 AI 기능(모델 재학습 트리거, 프롬프트 템플릿 수정 등)에 접근하지 못하도록 통제되는지를 {테스트결과서} 기반으로 확인한다.

2. 필요성

▶ AI 추론 API는 모델 가중치(Weight)와 학습 데이터에 간접적으로 접근할 수 있는 통로이므로, 인증/인가 통제가 미흡할 경우 아래와 같은 AI 특화 위험이 현실화될 수 있다.

▶ 모델 추출 공격(Model Extraction): 인증 없이 또는 탈취된 토큰으로 대량의 추론 요청을 반복 전송하면, 응답(Output)을 수집하여 원본 AI 모델을 복제할 수 있다. 이는 지적재산 유출과 동일한 결과를 초래한다.

▶ 비용 폭증(Cost Explosion): GPU 기반 추론은 호출당 비용이 높으므로, 무인가 대량 호출이 차단되지 않으면 예산 초과 및 서비스 가용성 저하가 발생한다.

▶ AI 기능 무단 조작: 모델 재학습(Fine-tuning) 트리거, 프롬프트 템플릿 변경, RAG 검색 인덱스 수정 등 고위험 관리 기능에 대한 역할 기반 접근통제(RBAC)가 없으면, 내부자 또는 외부 공격자가 AI 서비스의 동작을 변조할 수 있다.

▶ 토큰 탈취 후 측면 이동(Lateral Movement): AI 서비스는 내부 마이크로서비스 간 통신에도 토큰을 사용하는 경우가 많으며, 만료되지 않은 장기 토큰이 유출될 경우 내부망 전체로 피해가 확산될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 API 인증 메커니즘 구현 확인	{보안설계서}와 API 게이트웨이 설정을 대조하여 AI 추론 API에 JWT 등 토큰 기반 인증이 구현되었는지 확인
2	토큰 만료·갱신 정책 적정성 검증	토큰 만료 시간과 갱신(Refresh) 정책이 설계서에 정의된 값대로 설정되었는지 확인하고, 만료 시간 설정 근거를 사업자 면담으로 확인
3	역할 기반 접근통제(RBAC) 구현 확인	역할별 접근 권한(RBAC) 매트릭스와 실제 API 응답을 대조하여 권한 없는 역할의 호출이 차단되는지 API 테스트 도구로 검증
4	AI 추론 API 인가 실패 시 HTTP 403 처리	인가되지 않은 역할로 AI 추론 API를 호출했을 때 HTTP 403이 반환되고 오류 응답에 내부 정보가 노출되지 않는지 API 테스트 도구로 확인
5	서비스 간 통신(Service-to-Service) 토큰 통제	서비스 간 통신에 사용되는 토큰의 발급 주체·유효 범위·저장 위치가 설계서에 정의되고 구현되었는지 확인하고, 키 관리 방식을 면담으로 확인
6	토큰 탈취 대응 체계(Revocation) 구현 확인	토큰 탈취 시 즉시 폐기(Revocation)할 수 있는 절차와 폐기 목록 관리 방식이 구현되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	API 게이트웨이의 인증/인가 통제(JWT 검증, 401/403 처리, RBAC, 토큰 만료·폐기 정책)가 설계대로 구현되고 테스트 결과로 확인된 경우	현행 유지
보완권고 (△)	주요 인증 테스트는 통과했으나 일부 항목(토큰 갱신 정책, 서비스 간 통신 토큰 통제, AI 관리 기능 RBAC 등)이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	핵심 인증/인가 통제가 미구현되었거나, 무인가 접근이 차단되지 않는 등 테스트에 실패한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 추론 API와 일반 API의 구분: 감리 대상 시스템에서 AI 추론 전용 API 엔드포인트를 먼저 식별해야 한다. 일반 CRUD API와 AI 추론 API는 위협 수준과 비용 구조가 다르므로, 'API 게이트웨이 정책 설정 문서'에서 AI 추론 API에 대한 별도의 인증/인가 정책이 적용되어 있는지를 우선적으로 확인한다.

2) 테스트 환경과 운영 환경의 정책 동일성 확인: {테스트결과서}의 테스트가 수행된 환경(Staging)과 실제 운영 환경(Production)의 API 게이트웨이 인증 정책이 동일한지 면담을 통해 확인한다. 테스트 환경에서는 정상 작동했으나 운영 환경에서 인증 검증이 비활성화(Bypass)된 사례가 빈번하다.

3) 모호한 보안 요건의 수치화 유도: 'API 인증을 적용한다'와 같은 선언적 문구만 있는 경우, '만료 토큰 차단율 100%', '변조 토큰 탐지율 100%', 'Access Token 만료 시간 1시간 이내' 등 측정 가능한 기준으로 구체화되었는지 확인하고, 미흡 시 구체화를 권고한다.

4) AI 고유 인가 항목 확인: 일반 웹 서비스의 RBAC 점검과 달리, AI 서비스에서는 '모델 재학습(Fine-tuning) 트리거 API', '프롬프트 템플릿 수정 API', 'RAG 검색 인덱스 갱신 API' 등 AI 특화 관리 기능에 대한 역할 분리가 설계·구현되었는지를 중점적으로 확인해야 한다.

5) 토큰 관련 로그 확인 요청: 인증 실패(401) 및 인가 실패(403) 로그가 중앙 집중 로깅 시스템(SIEM 등)에 정상적으로 수집되고 있는지를 면담 또는 로그 샘플 확인을 통해 점검한다. 향후 보안 사고 추적 가능성을 확보하기 위한 전제 조건이다.

AG_SD44-1-06 | 내부 서버간 제로트러스트 인증

시스템구조·보안 | 시험 | 시스템 구현 | 인증/인가/접근통제

근거 | 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · ISO/IEC 27001

1. 목적

AI 서비스 아키텍처에서 WAS(웹 애플리케이션 서버)와 AI 추론(Inference) 서버는 내부 네트워크로 연결되어 대량의 사용자 입력 데이터와 모델 추론 결과를 주고받는다.

이 구간에 인증 없이 평문 통신이 이루어질 경우, 내부 공격자나 측면 이동(Lateral Movement)을 통해 모델 추출, 프롬프트 인젝션, 데이터 탈취 등 AI 특유의 보안 위협에 노출된다.

본 가이드라인은 종료(시험/전개) 단계에서, 설계된 제로트러스트 인증 체계(mTLS, 토큰 인증 등)가 실제 시스템에 구현되었는지, 보안점검이 적정하게 수행되었는지를 검토하는 것을 목적으로 한다.

2. 필요성

▶ AI 추론 서버는 모델 가중치(Weight), 사용자 입력 데이터(프롬프트), 추론 결과 등 고가치 자산을 처리한다. 기존 방화벽 기반의 경계 보안만으로는 내부 네트워크에서 발생하는 측면 이동 공격을 차단할 수 없다. 특히 AI 서비스에서는 추론 API 엔드포인트를 통한 모델 추출(Model Extraction) 공격 등 내부 구간을 경유하는 AI 고유의 공격 경로가 존재하므로, 내부 서버 간에도 상호 인증과 암호화가 필요하다.

▶ 제로트러스트 내부 인증이 없으면 내부망 침해 발생 시 AI 추론 서버·모델 가중치·사용자 프롬프트 전체가 노출된다. 국가정보원의「제로트러스트 가이드라인」(2023)과 금융보안원 지침은 내부 서버 간 인증 강화를 명시적으로 요구하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	mTLS 적용 구현 확인	WAS↔AI 추론 서버 간 상호 인증서 기반 TLS 통신이 설계대로 구현되었는지 확인
2	토큰 인증 체계 구현 확인	AI 추론 API 호출 시 JWT 등 토큰 기반 인증/권한 검증이 적용되었는지 확인
3	비인가 접근 차단 시험	인증 없이 AI 추론 서버에 직접 접근 시 차단되는지 시험 결과 확인
4	인증서/토큰 만료/폐기 처리	인증서 만료 시 자동 갱신 또는 연결 차단 처리 절차가 구현되었는지 확인

No	점검 항목	확인 방법
5	통신 구간별 암호화 적용	모든 내부 서버 간 통신이 TLS 1.2 이상으로 암호화되었는지 확인
6	보안점검 결과 문서화	{보안점검결과}에 AI 추론 서버 관련 상세 검토 항목이 포함되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	WAS↔AI 추론 서버 간 mTLS 상호 인증과 토큰 인증 체계가 설계대로 구현되고, 비정상 시나리오(무토큰, 만료, 위변조) 시험 및 인증서/토큰 수명주기 관리 체계가 보안점검 결과로 확인된 경우	현행 유지
보완권고 (△)	mTLS는 구현되었으나 비정상 시나리오 시험이 수행되지 않았거나, 토큰 TTL이 과도하게 길게 설정(예: 24시간 이상)되거나 인증서 갱신 계획·AI 추론 서버 관련 점검 항목이 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	WAS↔AI 추론 서버 간 인증 없이 평문 통신이 가능하거나, mTLS/토큰 인증이 전혀 구현되지 않았거나, 보안점검결과서가 존재하지 않는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 추론 서버는 민감 자산(모델 가중치, 사용자 프롬프트)을 처리하므로, 내부 서버라고 해서 인증을 생략해서는 안 된다. '내부 네트워크라서 안전하다'는 전제는 제로트러스트 관점에서 부적절하며, 이러한 구간이 공격 경로로 악용된 사례가 다수 보고되고 있다.

2) mTLS 적용 여부를 감리원이 직접 기술적으로 검증할 필요는 없으나, {보안점검결과}와 {테스트결과서}에 구체적인 시험 증빙(인증서 없이 접근 시 거부 화면, 만료 인증서 접근 시 오류 메시지 등)이 포함되었는지를 확인한다.

3) AI 추론 서버가 컨테이너(Docker/K8s) 환경에서 운영되는 경우, Service Mesh(예: Istio)를 통해 mTLS가 적용될 수 있다. 이 경우 애플리케이션 코드가 아닌 인프라 계층에서 인증이 처리되므로, 설계서에 인증 처리 계층이 명확히 정의되었는지 확인한다.

4) 토큰 만료시간(TTL)이 과도하게 길면(예: 24시간 이상) 토큰 탈취 시 피해가 커진다. 반대로 너무 짧으면(예: 1분 미만) 빈번한 토큰 갱신으로 성능 저하가 발생할 수 있다. 적정 TTL(예: 5~30분)이 설정되었는지 확인한다.

5) 보안점검결과서가 일반적인 웹 애플리케이션 보안 항목만 포함하고, AI 추론 서버 관련 상세 검토 항목(추론 API 접근제어, 모델 저장소 접근제어 등)이 누락된 경우가 많다. AI 특화 보안상세 검토 항목이 포함되었는지 확인한다.

AG_SD44-1-07 | 서비스 관리자페이지 IP 통제점검

시스템구조·보안 | 시험 | 시스템 구현 | 인증/인가/접근통제

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.) · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 서비스 관리자 페이지(관리 콘솔)에 대해 허용된 내부망 IP 대역에서만 접근이 가능하도록 IP 기반 접근통제가 설계대로 구현되었는지, 그리고 해당 통제의 시험이 적정하게 수행되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 서비스 관리자 페이지는 모델 재학습, 프롬프트 템플릿 수정, 금칙어 목록 관리, 응답 필터링 정책 변경 등 AI 시스템의 핵심 동작을 제어할 수 있는 고권한 기능을 포함하므로, 비인가 IP에서의 접근은 모델 탈취, 서비스 조작, 정책 변경 등 심각한 보안 위협으로 직결될 수 있다.

2. 필요성

▶ AI 서비스의 관리자 페이지는 프롬프트 템플릿 수정, 모델 교체, 금칙어 목록 변경, 응답 필터링 정책 조정 등 AI 시스템의 핵심 동작을 제어하는 기능을 제공한다. 만약 외부망 또는 비인가 IP에서 관리자 페이지에 접근할 수 있다면, 다음과 같은 AI 고유의 보안 위험이 발생한다.

▶ 프롬프트 인젝션(Prompt Injection): 공격자가 관리자 콘솔에 접근하여 시스템 프롬프트를 조작하면, AI 응답 자체가 악의적으로 변조될 수 있음

▶ 모델 탈취/조작: 관리 콘솔을 통해 모델 버전을 교체하거나 가중치(weight)를 변경하면 AI 시스템 전체의 신뢰성이 훼손됨

▶ 금칙어/필터링 정책 비활성화: 공격자가 금칙어 목록이나 응답 필터링 정책을 해제하면 유해 콘텐츠 생성이 가능해짐

▶ 사내 정보 유출: 관리 콘솔에서 접근 가능한 학습 데이터, 사용자 피드백 로그, 시스템 설정 정보 등이 외부로 노출될 위험

3. 검토방법

No	점검 항목	확인 방법
1	IP 화이트리스트 구현 적정성	설계서에 명시된 허용 IP 대역이 실제 시스템에 정확히 반영되었는지 확인
2	비인가 IP 차단 시험	허용되지 않은 IP에서 관리자 페이지 접근 시 차단되는지 시험 결과 확인

No	점검 항목	확인 방법
3	IP 통제 우회 방지 검증	프록시 경유, VPN 우회, 포트 직접 접근 등 우회 경로에 대한 차단 여부 검증
4	접근 로그 기록 적정성	관리자 페이지 접근 시도(성공·실패)가 로그로 기록되고 실패 임계 초과 시 경보가 발생하도록 설정되었는지 로그 및 경보 설정을 확인
5	전개 계획 반영 적정성	'배포/전개 계획서'에 IP 통제 적용 절차와 운영 이관 계획이 명시되어 있는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	설계서의 IP 통제 정책이 실제 시스템에 정확히 구현되었고, 비인가 IP 차단 시험 완료, 우회 방지 검증 완료, 로그/경보 설정이 확인되며, 배포·전개 계획서에 IP 통제 적용 절차와 운영 이관 계획이 명시된 경우	현행 유지
보완권고 (△)	IP 통제는 구현되었으나 우회 경로 검증 누락, 접근 로그·경보 미설정, 전개 계획 미반영 등 일부 미흡이 존재하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	IP 접근통제가 구현되지 않았거나, 비인가 IP 차단 시험이 수행되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 관리자 페이지와 일반 웹 관리자 페이지의 구분: AI 관리자 페이지는 모델 학습 파라미터 조정, 프롬프트 템플릿 변경, 금칙어 관리 등 AI 모델의 동작 자체를 변경할 수 있으므로, 일반적인 웹 애플리케이션 관리자 페이지보다 보안 수준이 높아야 한다. IP 통제 외에 다요소 인증(MFA) 적용 여부도 함께 확인하는 것이 바람직하다.

2) CIDR 범위의 적정성 판단: 허용 IP 범위가 /24(약 254개 IP) 이하인지 확인한다. /16(약 65,000개) 이상의 큰 범위는 불필요한 IP가 포함될 수 있어 보안 위험이 높다. 특별한 사유가 없는 한 관리자 페이지는 개별 IP 또는 /28~/24 수준으로 제한하는 것이 권장된다.

3) 실증 시험의 효과: 산출물 검토만으로는 IP 통제의 실효성을 100% 확인하기 어렵다. 감리 현장에서 비인가 네트워크의 PC나 모바일 단말(4G/5G)을 이용하여 관리자 페이지 URL에 직접 접근을 시도하면 차단 여부를 즉시 확인할 수 있다.

4) 전개 후 변경관리 확인: IP 통제 정책은 최초 적용 후에도 조직 변경, 네트워크 재구성 등으로 수정이 필요할 수 있다. '배포/전개 계획서'에 IP 변경 시 승인 절차, 변경 이력 기록, 정기 검토 주기 등이 명시되어 있는지 확인한다.

5) 카나리아 배포와 IP 통제 연계: AI 서비스가 카나리아(Canary) 또는 Blue-Green 배포 방식을 채택하는 경우, 신규 버전 환경에도 동일한 IP 통제 정책이 적용되는지 확인해야 한다. 배포 방식에 따른 IP 통제 적용 범위를 '배포/전개 계획서'에서 확인한다.

AG_SD44-1-08 | 구간암호화 및 인증서 유효성

시스템구조·보안 | 시험 | 시스템 구현 | 암호화

근거 | 「개인정보 보호법」 · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · KCMVP

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에서 정의된 구간 암호화(TLS) 요건이 실제 시스템에 올바르게 구현되었는지, AI 추론(Inference) API 및 모델 서빙 구간에 HTTPS 통신이 강제 적용되었는지, 인증서의 유효성이 확보되었는지를 {보안점검결과}를 중심으로 확인하는 것을 목적으로 한다.

AI 서비스는 일반 웹 서비스와 달리 대용량 프롬프트, 모델 파라미터, 사용자 컨텍스트 등 민감 데이터가 추론 API를 통해 빈번하게 전송된다.

이 과정에서 평문 전송이 이루어지거나, 만료·자체서명 인증서가 사용될 경우 중간자 공격(MITM)에 의해 프롬프트 내용 탈취, 응답 변조(환각 유도), 모델 엔드포인트 위장 등 AI 고유의 보안 위협이 발생할 수 있다.

이에 따라 구현된 암호화 설정과 인증서 상태를 실질적으로 점검할 필요가 있다.

2. 필요성

▶ AI 서비스의 통신 구간 암호화는 단순히 HTTPS를 적용하는 것에 그치지 않으며, AI 시스템 특유의 다중 구간 통신(프론트엔드 → API 게이트웨이 → 추론 서버 → 모델 서빙 → 벡터 DB 등)이 모두 안전하게 보호되고 있는지를 종합적으로 확인해야 한다.

▶ AI 추론 API 프롬프트 보호: 사용자가 입력하는 프롬프트에는 업무 기밀, 개인정보, 내부 문서 내용 등이 포함될 수 있으며, 평문 전송 시 네트워크 스니핑을 통해 탈취될 수 있다. TLS 1.2 이상 적용으로 전송 구간의 기밀성을 확보해야 한다.

▶ 모델 응답 변조 방지: 추론 서버로부터의 응답이 암호화되지 않을 경우, 공격자가 중간에서 응답을 변조하여 잘못된 정보를 사용자에게 전달(환각 증폭)할 위험이 있다. 구간 암호화를 통해 응답 무결성을 보장해야 한다.

▶ 인증서 만료·위변조에 의한 서비스 장애 예방: AI 서비스의 인증서가 만료되거나 자체 서명(Self-signed)된 경우, 클라이언트 측에서 연결을 거부하거나 보안 경고가 발생하여 서비스 가용성이 저하된다. 인증서 유효기간, 발급기관(CA), 인증서 체인을 사전 점검해야 한다.

▶ 내부 마이크로서비스 간 암호화 누락 방지: AI 시스템은 모델 서빙(예: Triton, vLLM), 벡터 DB(예: Milvus, Pinecone), RAG 파이프라인 등 복수의 내부 서비스가 연계되며, 이들 간 통신에 평문 gRPC/HTTP가 사용될 경우 내부자 위협에 취약해진다.

3. 검토방법

No	점검 항목	확인 방법
1	TLS 프로토콜 버전 적정성	AI 추론 API 및 모델 서빙 엔드포인트에 TLS 1.2 이상이 적용되었는지, SSL 3.0/TLS 1.0/1.1 등 취약 프로토콜이 비활성화되었는지 확인
2	인증서 유효성 검증	인증서 만료일, 발급 CA의 신뢰성, 인증서 체인 완전성, 와일드카드 인증서 남용 여부 확인
3	HTTPS 강제 적용	HTTP→HTTPS 리다이렉트(301/308) 또는 HSTS 헤더 적용 여부, 평문 HTTP 접근 차단 여부 확인
4	내부 서비스 간 암호화	추론 서버↔벡터 DB, API 게이트웨이↔모델 서빙 등 내부 구간 mTLS 또는 TLS 적용 여부 확인
5	암호 스위트(Cipher Suite) 적정성	취약 암호 스위트(RC4, DES, 3DES, NULL 등) 비활성화 여부, PFS(Forward Secrecy) 지원 확인
6	AI 모델 다운로드/업데이트 구간 암호화	모델 가중치 파일 다운로드 및 업데이트 경로에 TLS가 적용되었는지, 무결성 검증(해시/서명)이 병행되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 추론 API 및 모델 서빙 엔드포인트에 TLS 1.2 이상이 적용되고, 공인 CA 발급 인증서가 유효하며, HTTPS 강제 전환이 구현되고, AI 모델 가중치 파일의 다운로드·업데이트 경로에 TLS가 적용되어 해시·서명 기반 무결성 검증이 병행되며, {보안점검결과}에 이상이 없는 경우	현행 유지 (사례: TLS 1.3 적용, 공인 CA 인증서 잔여 120일, HSTS 및 내부 mTLS 적용 완료)
보완권고 (△)	외부 구간 TLS 적용은 적정하나, 내부 서비스 간 암호화 미흡, 인증서 갱신 절차 미수립, 또는 일부 취약 암호 스위트 잔존 등 보완이 필요한 경우	해당 항목 보완 요청 후 재확인 (사례: 추론서버↔벡터 DB 평문 gRPC 사용, 인증서 잔여 45일에 갱신 절차 부재)
부적합 (X)	AI 추론 API에 TLS 미적용 또는 TLS 1.0/1.1 사용, 자체서명 인증서 사용, HTTPS 미강제, {보안점검결과} 미작성 등 중대한 보안 결함이 있는 경우	시정조치 요구 및 이행결과 확인 (사례: 추론 API 평문 HTTP 통신, 자체서명 인증서 사용)

5. 감리 시 유의사항 및 실무 Tip

1) AI 추론 구간과 일반 웹 구간 구분: 일반적인 HTTPS 적용 여부는 기존 IT 감리에서 다루므로, 본 점검에서는 AI 추론 API, 모델 서빙, 벡터 DB 연계 등 AI 고유 구간의 암호화 적용 여부에 중점을 두어야 한다. 단순히 프론트엔드 HTTPS 적용만으로 '적합' 판정하지 않도록 유의한다.

2) 도구 스캔 결과의 해석: SSL Labs 등급이 A 이상이라도 내부 구간 암호화 미적용, 인증서 갱신 체계 부재 등의 문제가 있을 수 있다. 외부 스캔 결과만으로 전체 적정성을 판단하지 말고, 내부 구간은 별도 확인이 필요하다.

3) 인증서 잔여 기간의 실질적 판단: 인증서 잔여 기간이 '6개월'이라 해도, 전개 후 운영 시점까지의 기간을 고려해야 한다. 자동 갱신 체계가 없는 경우, 잔여 기간이 90일 이상이더라도 갱신 절차서 존재 여부를 반드시 확인한다.

4) gRPC 구간의 암호화 확인: AI 추론 서버가 gRPC를 사용하는 경우, gRPC는 기본적으로 HTTP/2 기반이므로 TLS가 자동 적용되지 않는다. 반드시 gRPCS(gRPC Secure, TLS 활성화된 gRPC) 모드로 설정되어 있는지 확인해야 한다.

5) 모델 가중치 전송 구간 확인: 모델 레지스트리(예: MLflow, Hugging Face Hub)에서 추론 서버로 모델 가중치를 다운로드하는 경로가 HTTPS인지, 다운로드 후 무결성 검증(SHA-256 등)이 수행되는지를 면담으로 확인한다. 평문 HTTP 다운로드는 모델 변조 공격의 경로가 될 수 있다.

AG_SD44-1-09 | DDoS_WAF 방어시험 점검

시스템구조·보안 | 시험 | 시스템 구현 | 시큐어코딩

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · OWASP LLM Top 10

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 및 구현된 DDoS 방어 체계와 WAF(Web Application Firewall)/L7 계층 차단 기능이 AI 추론 API 엔드포인트에 대한 대용량 트래픽 공격 시나리오에서 실제로 정상 동작하는지를 검증하기 위한 모의훈련 및 시험 결과의 적정성을 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 AI 서비스는 추론 요청 1건당 소요되는 GPU 연산 자원이 일반 웹 요청 대비 수십~수백 배에 달하므로, 상대적으로 적은 수의 동시 요청만으로도 서비스 가용성이 급격히 저하될 수 있다.

따라서 기존 웹 서비스 대상 DDoS 방어 시험과는 별도로, AI 추론 API 특성을 반영한 부하 시나리오와 임계값 기반 차단 동작의 유효성이 실증적으로 검증되었는지 확인할 필요가 있다.

2. 필요성

▶ AI 서비스 인프라는 GPU 기반 추론 서버의 자원 제약으로 인해 일반 웹 서비스와 비교하여 대용량 트래픽에 대한 취약성이 구조적으로 높다. 종료(시험/전개) 단계에서 설계된 방어 체계의 실효성을 모의훈련을 통해 검증하지 않으면, 운영 개시 이후 실제 공격 발생 시 AI 서비스가 무방비 상태에 놓일 수 있다.

▶ AI 추론 자원의 고갈 방지 검증: GPU 자원은 유한하며 추론 요청이 폭주할 경우 정상 사용자의 응답 지연 또는 서비스 중단으로 이어진다. 설계된 Rate Limit과 임계값 초과 시 WAF/L7 차단 규칙이 실제 부하 상황에서 정상 동작하는지 시험을 통해 실증적으로 확인해야 한다.

▶ 모의훈련의 실효성 확보: 단순히 네트워크 계층(L3/L4) 수준의 DDoS 방어만으로는 AI API에 대한 Application 계층(L7) 공격(예: Slowloris, HTTP Flood, 대량 프롬프트 요청)을 차단할 수 없다. AI 서비스 특성을 반영한 L7 계층 시험 시나리오의 포함 여부를 확인해야 한다.

▶ 배포 전 최종 안전망 확인: 운영 환경과 동등한(또는 유사한) 조건에서의 모의훈련 결과가 {테스트결과서}에 기록되어야 배포/전개 승인의 객관적 근거가 된다. 시험 미수행 또는 부실 시험은 운영 리스크를 직접적으로 증가시킨다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 API DDoS 모의훈련 시나리오 수립 여부	AI 추론 API 엔드포인트를 대상으로 한 DDoS 모의훈련 시나리오가 {테스트계획서}에 수립되었는지 확인
2	L7 계층 공격 유형 포함 여부	HTTP Flood·대량 프롬프트 요청 등 L7 계층 공격 유형이 모의훈련 시나리오에 포함되었는지 확인
3	임계값 초과 시 WAF/L7 차단 동작 시험 여부	AI 추론 API 임계값 초과 시 WAF/L7 차단 규칙이 실제 동작하는지 {테스트결과서} 차단 건수·차단율과 부하테스트 도구 로그를 대조
4	운영 유사 조건 부하 시험 수행 여부	부하테스트 도구로 운영 환경과 유사한 동시 접속·트래픽 조건에서 시험이 수행되었는지 시험 조건 기재 내용에서 확인
5	정상·공격 트래픽 분리 처리(오탐 방지) 확인 여부	시험 결과에서 정상 트래픽이 차단되지 않았는지(오탐 방지) 확인하고 오탐 발생 시 조치 내역을 담당자 면담으로 확인
6	시험 부하 수준의 실 운영 규모 반영 여부	시험 시나리오 부하 수준이 AI 추론 서비스의 실 운영 트래픽 규모를 반영했는지 산정 근거를 확인하고 면담으로 검증
7	DDoS/WAF 요구사항-시험 항목 추적성 여부	{보안요구사항}의 DDoS/WAF 항목과 {테스트결과서} 시험 항목 간 추적성이 확보되어 누락 없이 대조되는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 추론 API 대상 DDoS 모의훈련 시나리오(L7 공격 유형 포함)가 {테스트계획서}에 수립되고, 임계값 초과 시 WAF/L7 차단 동작이 운영 환경과 유사한 조건에서 부하테스트 도구로 실증 확인되었으며, 오탐 방지 및 요구사항-시험항목 간 추적성이 확보된 경우	현행 유지
보완권고 (△)	시험은 수행되었으나 L7 계층·AI 특화 공격 시나리오 일부가 누락되었거나, 차단 로그 등 구체적 증적이 부족하거나, 시험 부하 수준이 실 운영 트래픽 규모를 반영하지 못한 경우	시험 결과 미제출 또는 기준 미달 항목에 대한 재시험 요구 및 보완 권고 후 재확인
부적합 (X)	AI 추론 API 대상 DDoS/WAF 방어 시험이 수행되지 않았거나, {테스트결과서}에 차단 동작 검증 결과가 전혀 기록되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 추론 API 특화 시나리오 확인: '정상적인 HTTP 부하테스트만 수행하였습니다'라는 답변에 안주하지 말고, AI 추론 API 엔드포인트에 특화된 부하 테스트(예: Rate Limiting이 단순 HTTP 요청 건수 기준인지, Token/연산량 기반 통제인지)가 수행되었는지 확인해야 한다.

2) 증적의 구체성 확인: {테스트결과서}에 '차단 성공'이라는 서술만 있고 WAF 로그, HTTP 응답 코드(429/503), 차단 IP 목록 등 구체적 증적이 없는 경우 '보완권고(△)'로 판정해야 한다. 특히 AI API Gateway와 WAF의 차단 로그가 시간대별로 정합성을 갖추고 있는지 교차 확인한다.

3) 오탐률(False Positive Rate) 측정 요구: DDoS 방어 시스템이 정상 사용자의 AI 추론 요청을 오차단하는 비율을 반드시 측정·기록하도록 요구해야 한다. AI 서비스 특성상 단일 사용자가 연속적으로 긴 프롬프트를 전송하는 것이 정상 패턴일 수 있으므로, 단순 빈도 기반 차단은 오탐을 유발할 수 있다.

4) 도구 활용 점검과 수작업 점검의 구분: WAF 차단 동작 시험 결과(③④)는 도구(부하테스트 도구 + WAF 로그)를 활용한 점검이 필수이며, 시험 시나리오의 적정성(①②⑤⑥⑦)은 문서 검토 및 면담(수작업)으로 점검한다. 감리원이 직접 부하테스트 도구를 실행할 필요는 없으나, 도구 실행 결과 로그의 진위 여부는 확인해야 한다.

5) 운영 환경 동등성 확인: 시험이 개발/스테이징 환경에서만 수행되고 운영 환경(또는 운영 동등 환경)에서는 미수행된 경우, 시험 결과의 신뢰성이 제한적이다. 시험 환경의 인프라 구성(GPU 종류·수량, WAF 버전, 네트워크 구성)이 운영 환경과 동등함을 확인하는 문서가 있는지 점검한다.

AG_SD44-1-10 | OS미들웨어 보안패치_CVE취약점 점검

시스템구조·보안 | 시험 | 시스템 구현 | 시큐어코딩

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · OWASP LLM Top 10 · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 추론 서버·API 서버·학습 인프라 등 서비스 운영에 직접 사용되는 서버의 OS 및 미들웨어에 최신 보안 패치가 적용되었는지, 그리고 알려진 CVE(Common Vulnerabilities and Exposures) 취약점이 잔존하지 않는지를 확인하기 위한 점검 기준과 방법을 제시하는 것을 목적으로 한다.

AI 서비스 환경은 GPU 드라이버, CUDA/cuDNN 라이브러리, Python 런타임, 컨테이너 런타임(Docker ·containerd), 오케스트레이션 플랫폼(Kubernetes) 등 일반 웹 서비스와 다른 고유한 소프트웨어 스택으로 구성된다.

이들 컴포넌트는 패치 주기가 불규칙하고, 취약점 발생 시 AI 모델 탈취·추론 서버 침해·데이터 유출 등 고위험 결과로 이어질 수 있다.

본 가이드라인은 감리원이 {보안점검결과}를 기반으로 해당 위험이 통제되고 있는지 제3자 관점에서 점검하는 방법을 구체적으로 안내한다.

2. 필요성

▶ AI 서비스를 구성하는 소프트웨어 스택은 범용 웹 서버와 달리 GPU 드라이버, ML 프레임워크 (TensorFlow ·PyTorch·ONNX Runtime), API 게이트웨이, 컨테이너 런타임 등이 복합적으로 연동된다. 각 컴포넌트에서 발견된 CVE 취약점이 패치되지 않은 상태로 운영환경에 배포될 경우 다음과 같은 위험이 발생할 수 있다.

▶ AI 모델 무단 접근·탈취: 미패치 컨테이너 런타임 취약점을 통해 공격자가 컨테이너 탈출(Container Escape) 후 AI 모델 가중치에 접근할 수 있다.

▶ 학습 데이터 유출: OS 커널 또는 Python 런타임 취약점을 통해 학습 데이터셋·개인정보가 저장된 파일시스템에 접근할 수 있다.

▶ 추론 서버 서비스 중단: 미들웨어(Nginx, Gunicorn 등) 취약점 악용 시 서비스 가용성이 침해되어 AI 서비스 응답 불가 상태가 발생할 수 있다.

▶ 공급망 공격(Supply Chain Attack): 오픈소스 의존 라이브러리의 알려진 CVE를 이용한 악성코드 삽입으로 AI 모델 추론 결과 조작 위험이 존재한다.

3. 검토방법

No	점검 항목	확인 방법
1	OS 커널 보안 패치 최신화 여부	{보안점검결과}·'패치 관리 대장'에서 OS 커널 보안 패치 최신 상태(Ubuntu LTS의 ESM, RHEL/CentOS의 EPEL 기준 등)를 직접 검토
2	GPU 드라이버·CUDA 버전 보안 최신화 여부	{보안점검결과}·'서버 구성 명세서'에서 GPU 드라이버·CUDA 버전이 공개 보안 권고에 따라 최신 상태를 유지하는지 검토
3	Python 런타임 버전 지원 여부 및 보안 패치 적용 여부	'서버 구성 명세서'·{보안점검결과}에서 Python 런타임이 EOL이 아닌 지원 버전이며 보안 패치가 적용되었는지 검토
4	웹/API 미들웨어 CVE 패치 적용 여부	{보안점검결과}·'패치 관리 대장'에서 Nginx·Apache·Gunicorn·uWSGI 등 미들웨어에 Critical·High 등급 CVE 패치가 적용되었는지 검토
5	컨테이너 런타임 보안 패치 적용 여부	{보안점검결과} 검토와 사업자 면담으로 Docker·containerd·CRI-O 등 컨테이너 런타임의 보안 패치 적용 여부(탈출 취약점 중점)를 확인
6	쿠버네티스 핵심 컴포넌트 보안 패치 적용 여부	{보안점검결과}·'Kubernetes 버전 현황'에서 kube-apiserver·kubelet·etcd 등 핵심 컴포넌트의 보안 패치 적용 여부를 검토

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	OS 커널, GPU 드라이버/CUDA, Python 런타임, 미들웨어, 컨테이너 런타임, Kubernetes 등 전체 스택에 최신 보안 패치가 적용되고, Critical·High 등급 CVE가 잔존하지 않음이 {보안점검결과}로 확인된 경우	현행 유지
보완권고 (△)	주요 패치는 적용되었으나 일부 컴포넌트(GPU 드라이버, Python 런타임 등)의 패치가 지연되거나, 미조치 취약점에 대한 위험 수용 결재 문서·대안 통제가 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	Critical·High 등급 CVE가 미조치 상태로 잔존하거나, EOL 버전의 OS·Python·미들웨어가 운영환경에 사용되고 있는 경우, 또는 {보안점검결과}가 작성되지 않은 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 특화 스택의 별도 점검 목록 준비: GPU 드라이버(NVIDIA CUDA), LLM 서빙 프레임워크(vLLM, TGI 등), Python ML 라이브러리는 일반 취약점 스캐너가 탐지하지 못하는 경우가 있다. 사업자에게 AI 특화 컴포넌트 목록을 별도로 요청하여 스캔 대상 포함 여부를 확인해야 한다.

2) 스캔 보고서 발행 시점 확인: 취약점 스캔 보고서는 반드시 시험(테스트) 완료 이후, 배포·전개 직전에 발행된 것이어야 한다. 개발 초기에 실행된 스캔 보고서를 최종 보고서로 제출하는 사례가 있으므로 보고서 발행일과 현장 배포 일정을 반드시 대조 확인해야 한다.

3) EOL 버전 사용 여부의 명확한 기준 제시: '지원이 종료된 OS·Python·미들웨어를 사용하지 않아야 한다'는 요건은 감리 현장에서 분쟁이 잦다. Python 3.8 이하(EOL 완료), CentOS 7(EOL 2024.06) 등 구체적인 EOL 기준 목록을 사전에 준비하여 점검 기준을 명확히 해야 한다.

4) 미조치 취약점에 대한 위험 수용 문서 확인: 패치를 즉시 적용하기 어려운 경우(호환성 문제 등) 위험 수용(Risk Acceptance) 결재 문서가 존재하는지 확인한다. 단순히 '패치 예정'이라는 구두 답변만으로는 '보완권고(△)' 이상 판정이 어렵다. 위험 수용 결재 문서에는 위험 수용 사유, 대안 통제 방안, 패치 적용 목표 일정이 포함되어야 한다.

5) 컨테이너 이미지 레지스트리 점검 포함 여부: Kubernetes 환경에서 AI 서비스를 운영하는 경우, 운영 중인 컨테이너 이미지 레지스트리(Harbor, ECR, GCR 등)에서 Pull되는 이미지에 대한 CVE 스캔 여부도 확인한다. 빌드 시점에 스캔하였더라도, 배포 시점의 Base Image에 신규 CVE가 발생했을 수 있으므로 배포 직전 재스캔이 필요하다.

AG_SD44-1-11 | 망분리_폐쇄환경 모델 훈련점검

시스템구조·보안 | 시험 | 시스템 구현 | 망분리/폐쇄망

근거 | 「전자정부법」 · 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · NIST AI RMF · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 요구정의 및 설계 단계에서 수립된 망분리 폐쇄 환경 기반의 AI 모델 훈련 인프라가 실제 시스템으로 구현되었는지, 외부 인터넷으로의 아웃바운드 통신이 설계대로 차단되어 있는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

공공기관의 AI 모델 훈련(Fine-tuning, 전이학습 포함) 과정에서 학습 데이터 및 모델 가중치(Weight)가 외부로 유출되면, 내부 업무 데이터·개인정보가 유출되거나 모델 탈취(Model Extraction) 공격에 노출될 수 있다.

따라서 훈련 환경이 물리적·논리적으로 외부 네트워크와 완전히 격리되었는지, 방화벽 아웃바운드 룰이 설계서({시스템구성도})와 일치하는지를 시험/실증적으로 확인하는 것이 핵심이다.

2. 필요성

▶ AI 모델 훈련은 대량의 학습 데이터를 GPU 서버에서 반복적으로 처리하는 과정으로, 일반적인 웹 서비스 운영과는 달리 훈련 과정 자체가 보안 위험의 원천이 된다. 특히 생성형 AI 모델은 학습 데이터의 패턴을 모델 내부에 기억(Memorization)하는 특성이 있어, 모델 가중치가 유출될 경우 학습 데이터 자체가 간접적으로 유출되는 결과를 초래할 수 있다.

▶ 학습 데이터·모델 가중치 유출 방지: AI 모델 훈련 환경이 외부 인터넷에 연결된 상태에서 운영되면, 훈련 중 모델 파라미터·체크포인트 파일이 악의적 스크립트나 설정 오류를 통해 외부 서버로 전송(Exfiltration)될 위험이 존재한다.

▶ 망분리 규정 준수 실증: 공공기관은 국가정보보안기본지침 및 전자정부법에 따라 업무망과 인터넷망을 분리해야 하며, AI 훈련 인프라는 업무망 내 폐쇄 구간에 위치해야 한다. 이를 시험 단계에서 실증하지 않으면 감리 종료 후 보안사고 발생 시 책임 소재가 불분명해진다.

▶ 국제 기준 정합성 확인: NIST AI RMF는 AI 시스템의 보안과 회복력(Secure and Resilient)을 핵심 신뢰성 속성으로 정의하며, 모델 추출·데이터 포이즈닝 등 AI 고유 공격에 대한 보호 조치가 구현되었는지 확인을 요구한다.

▶ 설계-구현 일관성 검증: 요구정의 단계에서 작성된 보안요구사항이 실제 구현물에 반영되었는지를 종료 단계에서 대조·검증함으로써 요구사항 추적의 완결성을 확보한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 훈련 서버의 망분리 폐쇄 구간 배치 여부	{시스템구성도} 검토와 담당자 면담으로 AI 모델 훈련 서버(GPU 서버)가 명시된 망분리 폐쇄 구간(별도 VLAN 또는 물리 분리 구간)에 실제 배치되었는지 확인
2	훈련 환경 아웃바운드 트래픽 기본 차단 설정 여부	방화벽 정책·산출물 검토와 담당자 면담으로 훈련 환경에서 외부 인터넷(0.0.0.0/0)으로의 아웃바운드 트래픽이 기본 차단(Default Deny)으로 설정되었는지 확인
3	아웃바운드 예외 허용 룰의 정당한 승인 여부	{보안요구사항} 문서 확인과 담당자 면담으로 아웃바운드 예외 허용 룰의 목적·대상 IP·포트·프로토콜이 근거에 따라 정당하게 승인되었는지 확인
4	망분리 차단의 실제 통신 시험 증빙 여부	{테스트계획서}·{테스트결과서}에서 훈련 환경→인터넷 구간의 실제 통신 시도(ping·curl·TCP 연결 등)가 차단되었음을 증빙하는 시험 결과가 존재하는지 확인
5	학습 데이터 반입 경로 한정 및 직접 다운로드 차단 여부	{데이터관리계획서}·{전처리설계서}에서 학습 데이터 반입이 물리 매체·일방향 전송장비(Data Diode)로 한정되고 네트워크 직접 다운로드(pip install·wget 등)가 차단되었는지 확인
6	모델 가중치 파일 반출 절차 수립 여부	산출물 검토와 담당자 면담으로 훈련 완료 모델 가중치(Weight) 파일의 반출 절차와 반출 시 승인·기록 절차가 존재하는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{시스템구성도}와 실제 구축 환경이 일치하고, 훈련 환경→인터넷 아웃바운드가 Default Deny로 설정되어 있으며, 통신 차단 시험 결과가 모든 항목에서 '차단'으로 확인되고, 데이터 반입·모델 반출 절차가 수립되어 있는 경우	현행 유지
보완권고 (△)	일부 미흡 사항이 있으나 보완 조치로 해소 가능한 경우 (예: 아웃바운드 예외 룰의 사유가 문서화되지 않았으나 구두 확인 가능, 시험 결과서에 일부 포트 시험이 누락된 경우)	해당 항목 보완 요청 후 재확인
부적합 (X)	훈련 서버가 인터넷 접근이 가능한 구간에 배치되어 있거나, 아웃바운드 Default Deny가 미설정된 경우, 통신 차단 시험에서 외부 통신 성공 기록이 존재하거나 데이터 반입·반출 절차가 전혀 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 설계서와 실제 환경의 괴리 주의: {시스템구성도}에는 폐쇄 구간으로 설계되어 있으나, 구축 과정에서 편의상 인터넷 연결을 유지한 채 훈련을 진행하는 사례가 빈번하다. '실제 통신 차단 시험 결과'가 존재하는지를 반드시 확인해야 하며, 설계서만으로 적정 판정을 내려서는 안 된다.

2) AI 특화 위협과 기존 위협의 구분: 본 가이드라인은 'AI 모델 훈련 환경'이라는 AI 고유 인프라의 보안을 점검하는 것이다. 일반적인 방화벽 점검(인바운드 룰, 웹 방화벽 정책 등)은 기존 IT 감리 영역에서 다루므로, 여기서는 훈련 환경의 아웃바운드 차단과 데이터·모델 유출 방지에 초점을 맞추어야 한다.

3) 예외 허용 룰의 정당성 검증: 아웃바운드에 예외 허용 룰이 존재하는 경우, '왜 허용해야 하는가'에 대한 정당한 사유와 승인 기록이 반드시 있어야 한다. 예를 들어, 내부 NTP 서버 동기화, 라이선스 인증 서버 통신 등은 정당한 사유가 될 수 있으나, 외부 PyPI 저장소(pypi.org)로의 직접 접근은 폐쇄망 환경에서 허용되어서는 안 된다.

4) 모델 가중치 반출 = 데이터 유출 가능성: 생성형 AI 모델은 학습 데이터의 패턴을 가중치에 기억하므로, 모델 가중치 파일의 비인가 반출은 학습 데이터 유출과 동일한 위험을 갖는다. 모델 반출 절차에 '보안 검증(개인정보 포함 여부 점검 등)' 단계가 포함되어 있는지 확인해야 한다.

5) 클라우드 환경에서의 망분리 특수성: 클라우드(CSP) 기반의 AI 훈련 환경인 경우, 물리적 망분리가 아닌 논리적 격리(VPC, Private Subnet, Security Group)를 통해 구현될 수 있다. 이 경우 {시스템구성도}에 VPC 구성, 서브넷 배치, NAT Gateway 미사용 여부가 명시되어 있는지 확인하고, 클라우드 보안 그룹의 아웃바운드 룰이 Default Deny로 설정되어 있는지 확인해야 한다.

AG_SD44-1-12 | 클라우드 보안설정_방화벽 점검

시스템구조·보안 | 시험 | 시스템 구현 | 망분리/폐쇄망

근거 | CSAP · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에서 정의된 클라우드 보안 설정(방화벽, VPC 등)이 실제 시스템에 적정하게 구현·가동되었는지, 그리고 AI 서비스 운영에 필요한 포트 이외의 모든 외부 인바운드 접근이 100% 차단되어 있는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히 생성형 AI 서비스는 추론(Inference) API 엔드포인트, 모델 서빙 포트, GPU 클러스터 간 내부 통신 포트 등 일반 웹 시스템과 다른 고유한 네트워크 통신 경로를 사용하며, 이러한 AI 전용 포트가 외부에 불필요하게 노출될 경우 모델 탈취(Model Extraction), 프롬프트 인젝션(Prompt Injection) 경유, 추론 결과 도청 등 AI 고유 보안 위협의 공격 경로로 악용될 수 있다.

따라서 {보안점검결과}를 근거로 보안 설정의 설계 대비 구현 일치 여부와 AI 특화 포트의 차단 적정성을 확인한다.

2. 필요성

▶ 클라우드 환경에서 운영되는 AI 서비스는 일반적인 웹/WAS 서버 환경과 달리, 모델 추론 서버(예: TGI, vLLM, Triton 등), 벡터 데이터베이스, 오케스트레이터(예: LangChain 서버), GPU 노드 간 통신(NCCL/RDMA) 등 다층적인 네트워크 구성요소가 존재하므로, 이러한 구성요소 각각에 대해 방화벽·VPC 보안 설정의 적정성을 개별적으로 점검할 필요가 있다.

▶ AI 추론 API 엔드포인트 보호: 생성형 AI의 추론 API가 외부에 직접 노출되면 비인가 사용자의 대량 호출을 통한 모델 탈취(Model Extraction) 공격이 가능하다. 방화벽/VPC를 통해 허용된 경로만으로 접근을 제한해야 한다.

▶ GPU 클러스터 내부 통신 격리: AI 학습·추론에 사용되는 GPU 노드 간 통신(NCCL, gRPC 등)이 외부 네트워크에 노출되면 학습 데이터 및 모델 파라미터가 도청될 위험이 있다. VPC 서브넷 분리를 통한 내부 통신 격리가 필수적이다.

▶ AI 서비스 공격 표면(Attack Surface) 최소화: 일반 IT 시스템 대비 AI 시스템은 모델 서빙 포트, 벡터 DB 포트, 모니터링 대시보드 포트 등 추가적인 통신 포트를 사용하므로, 필요 최소 포트만 개방하고 나머지를 차단하는 점검이 운영 전 필수적으로 수행되어야 한다.

▶ CSAP 및 공공 클라우드 보안 기준 준수: 공공기관 AI 서비스가 클라우드 환경에서 운영되는 경우, CSAP(클라우드 보안인증) 기준에 따른 네트워크 보안 설정 적합성을 시험/전개 단계에서 최종 확인해야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	VPC 네트워크 분리 구현 적정성	AI 추론 서버, GPU 클러스터, 벡터 DB 등이 설계대로 별도 서브넷에 분리 배치되었는지 확인
2	방화벽(Security Group) 규칙 적정성	AI 서비스 필요 포트만 허용되고, 불필요 포트가 100% 차단되었는지 {보안점검결과} 대비 확인
3	AI 추론 API 엔드포인트 외부 노출 차단	추론 서버(vLLM, TGI 등) 포트가 외부 인터넷에서 직접 접근 불가능한지 확인
4	GPU 노드 간 통신 포트 격리	NCCL/gRPC 등 GPU 내부 통신 포트가 외부 서브넷에서 접근 불가능한지 확인
5	NACL(Network ACL) 설정 적정성	VPC 서브넷 레벨 NACL이 설계서 대비 정확하게 적용되었는지 확인
6	설계 대비 구현 일치성 확인	'클라우드 보안 설정서' 및 {시스템구성도} 대비 실제 설정 일치 여부 확인
7	AI 모니터링/관리 포트 접근 제한	Prometheus·Grafana 등 AI 모니터링 대시보드가 외부 인터넷에서 접근 불가능하고 VPN 또는 내부망으로만 접근이 제한되는지 방화벽 설정과 접속 시험으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	① VPC/서브넷 분리와 서브넷 레벨 NACL 설정이 설계서와 100% 일치하고, ② AI 서비스 필요 포트 외 모든 외부 인바운드가 차단되었으며, ③ AI 추론 API가 API Gateway를 경유하여 접근 제어되고, ④ GPU 내부 통신이 완전 격리된 경우	현행 유지
보완권고 (△)	대부분 적정하나 일부 미흡 사항이 존재하는 경우. 예: ① 모니터링 포트 1~2개가 외부에 노출되어 있으나 인증은 적용된 경우, ② 포트 매트릭스에 누락된 포트가 1~2건 존재하나 실질적 위험도가 낮은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	① AI 추론 서버 포트가 외부 인터넷에 직접 노출된 경우, ② Security Group에 0.0.0.0/0 허용 규칙이 AI 핵심 포트에 적용된 경우, ③ VPC 서브넷 분리가 설계서와 불일치하는 경우, ④ {보안점검결과}가 미제출된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 전용 포트 식별이 핵심: 일반 IT 감리에서 점검하는 80/443 포트 외에, AI 추론 서버(8080, 8443, 50051 등), 벡터 DB(6333, 19530 등), GPU 통신(NCCL 동적 포트), 모니터링(3000, 9090, 5000) 등 AI 고유 포트를 사전에 'AI 인프라 아키텍처 설계서' 또는 '포트 매트릭스'에서 식별한 후 점검해야 한다. 포트 목록 없이 점검하면 AI 특화 위협을 놓칠 수 있다.

2) 설계서와 실제 설정의 1:1 대조 필수: 클라우드 환경은 콘솔에서 손쉽게 보안 설정을 변경할 수 있어, 설계서 작성 이후 개발 과정에서 임시로 규칙을 완화한 뒤 복원하지 않는 경우가 빈번하다. 반드시 '포트 매트릭스'와 실제 Security Group 규칙을 항목별로 대조하고, 불일치 건수를 기록해야 한다.

3) 0.0.0.0/0 규칙 전수 확인: Security Group 인바운드 규칙에서 Source가 0.0.0.0/0(모든 IP 허용)으로 설정된 항목은 AI 서비스 포트에 대해 원칙적으로 부적정 사유에 해당한다. 다만 CDN이나 외부 사용자 대상 서비스 포트(HTTPS 443 등)는 예외로 인정할 수 있으며, 이 경우 WAF/DDoS 방어 적용 여부를 추가 확인해야 한다.

4) 클라우드 콘솔 화면 캡처 요청: 감리원이 직접 클라우드 콘솔에 접속하기 어려운 경우, 사업자에게 VPC 구성도, Security Group 규칙 목록, NACL 설정 화면의 캡처본을 요청하고, 캡처 일시가 감리 시점과 근접한지 확인한다. 구 버전 캡처를 제출하는 사례를 방지하기 위해 화면 내 시간 정보를 함께 확인한다.

5) AI 모델 서빙과 일반 웹 서빙의 구분: 생성형 AI 시스템은 일반 웹 서버(Nginx, Apache) 뒤에 별도의 추론 서버(vLLM, TGI, Triton 등)가 존재하는 다계층 구조이다. 감리 시 '웹 서버 포트만 확인하면 된다'는 인식을 경계하고, 추론 서버 계층까지 점검 범위에 포함해야 한다.

AG_SD44-1-13 | 모델 신뢰성 테스트 점검

시스템구조·보안 | 시험 | 인공지능 모델 개발 | AI QA체계

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 사전에 수립한 AI 모델 신뢰성 테스트 계획이 실제로 빠짐없이 수행되었는지, 그리고 그 결과가 {테스트결과서}에 완전하고 검증 가능하게 기록되었는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

감리원은 AI 시스템의 성능 수준을 직접 평가하는 것이 아니라, 계획된 테스트 시나리오 100%가 빠짐없이 실행되고 결과가 적정하게 기록되었는지의 '문서 완결성'과 '절차 준수 여부'를 중심으로 점검한다.

2. 필요성

▶ AI 모델은 일반 소프트웨어와 달리 동일 입력에 대해서도 확률적으로 다른 출력을 생성할 수 있으며, 환각(Hallucination), 편향(Bias), 적대적 입력(Adversarial Input)에 대한 취약성 등 일반 소프트웨어 테스트 방식으로 충분히 검증되지 않는 고유한 위험이 존재한다. 이에 따라 AI 모델 신뢰성 테스트의 이행 여부를 별도로 점검할 필요가 있다.

▶ 테스트 완결성 보증: 계획된 시나리오 중 일부가 누락될 경우 검증되지 않은 위험 영역이 운영 환경에 노출될 수 있다. 계획 대비 수행률 100%를 확인하는 것은 이러한 위험의 공백을 방지하기 위한 최소한의 조치이다.

▶ 결과 기록의 재현 가능성 확보: 테스트 결과가 모호하거나 불완전하게 기록될 경우 향후 결함 재현, 이의제기, 유지보수 시 근거 자료로 활용이 불가능해진다. 감리원은 기록의 구체성과 추적 가능성을 점검한다.

▶ 감리원의 역할 명확화: AI 신뢰성 테스트에서 감리원은 기술적 성능 평가자가 아닌 절차 준수 및 산출물 완결성 검토자로서, {테스트결과서}의 형식적·내용적 적정성을 점검한다.

▶ 규정 및 기준 준수 확인: TTA 신뢰할 수 있는 인공지능 개발안내서 및 NIST AI RMF는 AI 시스템 배포 전 신뢰성 검증(TEVV) 수행 및 결과 기록을 요구하며, 정보시스템 감리기준은 시험 단계 산출물의 적정성 점검을 감리원의 의무로 규정하고 있다.

3. 검토방법

No	점검 항목	확인 방법
1	테스트 시나리오 계획 대비 수행률 100% 여부	{테스트계획서} 대비 {테스트결과서}의 시나리오 수량이 일치(계획 대비 수행률 100%)하는지 시나리오 목록 수량을 대조
2	시나리오별 입력·예상·실제 결과 기록 여부	각 시나리오별 입력값·예상 결과·실제 결과가 구체적으로 기록되었는지 {테스트결과서} 샘플 항목(전체의 최소 20% 이상)에서 검토
3	테스트 수행일시·수행자 기록 여부	테스트 수행일시·수행자(서명 또는 식별 가능한 정보)가 {테스트결과서} 헤더 및 각 시나리오 기록 항목에 기록되었는지 확인
4	판정 기준 문서화·결과 일치 여부	Pass/Fail 판정 기준이 'AI 모델 신뢰성 테스트 계획서'에 사전 문서화되고 {테스트결과서} 결과와 일치하는지 비교 검토
5	Fail 시나리오 결함 등록·조치 기록 여부	Fail 판정 시나리오의 결함이 '결함관리대장'에 등록되고 조치 완료 여부가 기록되었는지 대조

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	계획된 모든 시나리오가 수행되어 수행률 100%가 확인되고, 시나리오별 입력·예상·실제 결과와 수행일시·수행자가 기록되었으며, Fail 항목이 결함관리대장에 등록·조치된 경우	현행 유지
보완권고 (△)	테스트는 수행되었으나 일부 시나리오의 결과 기록 미흡, 수행자 정보 누락, Fail 항목의 조치 이력 불완전 등 문서 완결성이 부족한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	계획 대비 시나리오가 상당수 미수행되었거나 {테스트결과서}가 미제출되어 신뢰성 검증 이행을 확인할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 수행률 100% 주장에 대한 증빙 요구: 사업자가 "모든 시나리오를 수행했다"고 주장하더라도, {테스트결과서}에 시나리오별 수행 일시와 결과가 구체적으로 기록되어 있지 않으면 이를 검증할 방법이 없다. 감리원은 반드시 목록 기반으로 계획 대비 결과를 직접 대조하고, 기록이 미흡한 항목에 대해서는 보완 요청 또는 지적을 명확히 해야 한다.

2) 요구사항 추적성 교차 검증: 단순 수행 여부뿐만 아니라 해당 테스트 시나리오가 제안요청서(RFP) 또는 요구사항정의서 상의 신뢰성/비기능 요구사항과 정확히 추적(Mapping)되는지 추적표를 통해 교차 검증해야 한다. (감리수행가이드 기준 점검)

3) AI 특화 테스트 항목의 명칭 다양성 주의: 환각 테스트, 편향 테스트 등 AI 특화 시나리오는 사업자마다 '품질 테스트', '모델 안정성 검증', 'QA 점검' 등 다양한 이름으로 기술될 수 있다. 감리원은 명칭이 아닌 실질 내용을 기준으로 AI 특화 테스트 항목이 포함되었는지를 판단해야 한다. 면담 시 "환각 방지를 위한 테스트 케이스가 있었는가"처럼 내용 중심의 질문을 사용한다.

AG_SD44-1-14 | 성능테스트 통과여부

시스템구조·보안 | 시험 | 시스템 구현 | 테스트시나리오

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE · OWASP LLM Top 10 · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 감리원이 자동화 테스트 스크립트 실행 결과, 자동화 성능 테스트 통과 여부 및 리포트를 검토할 때, AI 고유 품질 위험이 테스트 과정에 적절히 반영되었는지 여부를 점검하기 위한 검토 방향, 검토 항목, 검토 결과 작성 기준을 제시함을 목적으로 한다.

특히 감리원은 환각·편향성·프롬프트 인젝션 등 AI 특유의 리스크를 포함한 시험 시나리오 구성 여부, 사전 설계된 성능 지표(KPI) 대비 실측치의 합격 여부, 그리고 자동화 테스트 결과 산출물의 기재 완전성이라는 세 가지 핵심 관점에서 테스트결과서와 관련 산출물을 검토한다.

2. 필요성

▶ AI 서비스는 일반 소프트웨어와 달리 입력 데이터와 모델의 상호작용에 의해 출력이 비결정적으로 생성되므로, 기능 테스트만으로는 품질을 보증하기 어렵다. 동일한 입력에도 맥락에 따라 다른 결과가 나올 수 있으며, 훈련 데이터의 편향이 서비스 출력에 그대로 반영될 위험이 있다.

▶ 감리원은 다음 두 가지 위험에 집중하여 검토한다.

▶ AI 전용 시험 시나리오 부재 위험: 환각, 프롬프트 인젝션, 편향성 등 AI 특유의 위험 항목이 테스트케이스에 포함되지 않으면, 실제 서비스 오픈 후 예상하지 못한 장애 또는 민원이 발생할 수 있다.

▶ 성능 기준의 모호성 위험: 요구사항 단계에서 정의된 AI 성능 KPI(정확도, 응답시간, 처리량 등)가 테스트 결과서에 명확히 대조되지 않으면, 검수 단계에서 합격·불합격 판정 분쟁이 발생하거나 품질 미달 서비스가 전개될 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 전용 시험 시나리오의 반영 여부 확인	'테스트계획서'와 '시험결과서'에서 환각·편향성·프롬프트 인젝션 등 AI 전용 시험 시나리오가 테스트케이스로 포함되었는지 확인
2	AI 성능 지표(KPI) 기준치 충족 여부 확인	'AI품질기준정의서'의 성능 KPI 기준치와 '시험결과서'의 실측값을 대조하여 충족 여부를 확인

No	점검 항목	확인 방법
3	요구사항 추적 매트릭스(RTM) 시험 항목 연계 확인	요구사항 추적 매트릭스(RTM)에서 AI 성능 요구사항이 시험 항목과 누락 없이 연계되는지 '테스트계획서'와 대조
4	Red Teaming(레드팀) 수행 결과 반영 확인	레드팀 수행 결과가 '시험결과서'에 반영되고 발견 취약점의 조치 내역이 기재되었는지 확인하고, 수행 주체의 독립성을 담당자 면담으로 확인
5	Human-in-the-Loop 검증 절차 이행 여부 확인	Human-in-the-Loop 검증 절차가 '테스트계획서'에 정의되고 '시험결과서'에 이행 결과가 기재되었는지 확인
6	자동화 테스트 스크립트의 범위·실행 결과 검토	자동화 테스트 스크립트의 적용 범위와 실행 결과를 CI 로그 또는 테스트 도구 리포트로 확인하고, 미자동화 항목의 수동 시험 여부를 확인
7	AI 편향성 및 공정성 시험 결과 검토	'AI품질기준정의서'의 편향성·공정성 기준과 '시험결과서'의 집단별 측정 결과를 대조
8	배포·전개 계획(롤백 및 장애 대응 포함) 수립 확인	'배포·전개 계획서'에 롤백 절차와 장애 대응 방안이 수립되었는지 확인하고, 승인 책임자 지정 여부를 담당자 면담으로 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 전용 시험 시나리오(환각·편향성·프롬프트 인젝션)가 테스트케이스에 포함되고, 성능 KPI 기준치를 충족하였으며, RTM 연계가 완전하고, Human-in-the-Loop 검증 절차가 이행되었으며, 집단별 편향성·공정성 시험 결과가 기준을 충족하고, 롤백 절차와 장애 대응 방안을 포함한 배포·전개 계획이 수립된 경우	현행 유지
보완권고 (△)	AI 성능 지표 일부(응답속도 등)가 목표치 대비 10% 미만 미달이나 개선 계획이 명시된 경우, 또는 자동화 테스트 커버리지가 계획 대비 90% 이상 100% 미만인 경우	해당 항목 보완 요청 후 개선 계획 이행 여부 재확인
부적합 (X)	AI 전용 시험 시나리오 미작성, 성능 KPI 기준치 10% 이상 미달, RTM 내 AI 기능 미매핑, Red Teaming 미수행 중 하나 이상에 해당하는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 성능 기준의 수치화 확인: "AI의 응답 품질이 우수하여야 한다"와 같은 표현은 검수 분쟁의 원인이 된다. 감리원은 {테스트결과서}와 'AI 성능 테스트 리포트'에서 목표치(예: 정확도 92% 이상, 평균 응답 시간 2초 이내)가 수치로 기재되어 있는지, 그리고 실측치와 대조한 합격 판정이 명시되어 있는지 반드시 확인한다.

2) 테스트 환경과 운영 환경의 동등성 검증: 성능 테스트(응답속도, 처리량 등) 리포트를 검토할 때 단순히 목표 수치를 넘었는지만 보지 말고, 테스트가 수행된 환경(GPU 스펙, 동시 접속자 수, 벡터 DB 데이터 건수 등)이 운영 환경과 동등한 조건인지 확인한다. 환경 차이가 있는 경우 그 영향에 대한 분석 근거를 요청한다.

3) AI 전용 시나리오와 일반 기능 시나리오 구분: 단순 기능 테스트(예: 버튼 클릭, 화면 전환)와 AI 고유 위험 테스트(환각, 편향성, 프롬프트 인젝션)는 구분하여 점검한다. {테스트케이스}에 AI 전용 분류 체계가 없는 경우, 감리원은 항목 분류 보완을 권고한다.

AG_SD44-1-15 | APM 성능 저하 탐지 알림발송

시스템구조·보안 | 시험 | 운영·모니터링 | 모니터링/재학습

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MANAGE · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에서 정의된 APM(Application Performance Monitoring) 기반 성능 저하 탐지 및 알림 발송 요건이 실제 시스템에 적정하게 구현되었는지를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 추론 서버의 GPU 사용률이나 응답 지연(Latency) 등 AI 고유의 성능 지표가 모니터링 대상에 포함되었는지, GPU VRAM 사용률 90% 초과, 초당 토큰 생성 수(Tokens/sec) 저하 등 AI 특화 임계값 도달 시 이메일/SMS 등 자동 알림이 실제로 발송되는지를 {테스트결과서} 및 시스템 실행 결과를 통해 확인한다.

2. 필요성

▶ 생성형 AI 시스템은 일반 웹 애플리케이션과 달리 GPU 연산 부하, 모델 추론 지연, 대규모 동시 요청 처리 등 성능 저하 요인이 복합적으로 발생한다. 이러한 성능 이상을 실시간으로 탐지하고 담당자에게 즉시 알리는 체계가 적정하게 구현되지 않으면, 장애 대응이 지연되어 서비스 중단 또는 사용자 피해로 이어질 수 있다.

▶ AI 추론 성능 저하의 조기 탐지: 생성형 AI 모델은 프롬프트 복잡도, 토큰 길이, 동시 요청 수에 따라 응답 시간이 급격히 변동할 수 있으며, 기존 웹 APM과 달리 GPU 메모리 사용률, 추론 큐 대기 시간 등 AI 고유 지표를 함께 모니터링해야 정확한 성능 저하 원인을 식별할 수 있다.

▶ 장애 대응 시간 단축: 임계값 초과 시 자동 알림이 즉시 발송되지 않으면, 운영자가 성능 저하를 인지하는 시점이 지연되어, 사용자 불만과 서비스 신뢰도 하락으로 이어진다.

▶ 시험 단계에서의 실증 필요성: 설계 단계에서 정의한 임계값과 알림 정책이 실제 테스트를 통해 정상 작동함을 검증해야 운영 전환 시 장애 미탐지 위험을 방지할 수 있다.

▶ AI 모델 버전 변경 시 성능 회귀 감지: 모델 업데이트나 프롬프트 템플릿 변경 후 응답 시간이 급격히 증가하는 성능 회귀(Performance Regression)를 자동으로 감지하고 알리는 체계가 필요하다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 추론 서버 성능 지표의 APM 모니터링 대상 포함 여부	{AI품질기준정의서}·{테스트결과서}에서 GPU 사용률·추론 지연·큐 대기 등 AI 추론 서버 성능 지표가 APM 모니터링 대상에 포함되었는지 확인
2	자원별 임계값의 구체적 수치 정의 여부	{AI품질기준정의서}·{테스트결과서}에서 CPU·메모리·GPU 등 자원별 임계값이 구체적 수치(예: CPU 80%, GPU 메모리 90%)로 정의되었는지 확인
3	임계값 초과 알림 채널 설정 및 발송 시험 여부	임계값 초과 상황을 유발하거나 테스트 알림을 발송해 이메일·SMS·메신저 채널로 실제 수신되는지 확인하고 발송 이력을 APM 도구에서 대조
4	추론 응답 시간(Latency) 임계값 설정 및 별도 알림 발생 여부	AI 모델 추론 응답 시간 임계값이 APM에 설정되었는지 확인하고, 초과 시 자원 알림과 구분된 별도 알림이 발생하는지 시험 결과로 확인
5	부하 테스트 시나리오의 AI 고유 조건 포함 여부	{테스트계획서}·{테스트결과서}에서 대규모 동시 추론·긴 토큰 요청 등 AI 고유 조건이 부하 테스트 시나리오에 포함되었는지 확인
6	알림 수신자의 AI 인프라 담당자 지정 적정성	{시스템아키텍처설계서}·{인프라구성도}에서 알림 수신자가 GPU 운영자를 포함한 AI 인프라 담당자로 적절히 지정되었는지 확인
7	모델 버전 변경 후 성능 회귀 탐지 베이스라인 체계 수립 여부	{AI품질기준정의서}·{테스트결과서}에서 재학습·파인튜닝 후 성능 회귀 탐지를 위한 베이스라인 비교 체계가 수립되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 추론 성능 지표(GPU 사용률, 추론 Latency 등)가 APM 모니터링 대상에 포함되고, 구체적 임계값이 정의되었으며, 초과 시 알림 발송 테스트가 성공적으로 완료되고, 알림 수신자가 AI 인프라 담당자(GPU 운영자 포함)로 지정되었으며, 모델 버전 변경(재학습·파인튜닝) 후 성능 회귀 탐지를 위한 베이스라인 비교 체계가 수립된 경우	현행 유지
보완권고 (△)	임계값은 정의되었으나 AI 고유 지표(GPU, Latency)가 부분적으로 누락되었거나, 알림 테스트가 일부 채널에서만 수행된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 추론 성능 지표가 APM에 등록되지 않았거나, 임계값 초과 시 알림이 발송되지 않는 경우, 부하 테스트에 AI 고유 조건이 전혀 반영되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 성능 요건의 수치화: '성능을 모니터링한다'와 같은 모호한 문구는 향후 분쟁의 원인이 된다. 'GPU 사용률 85% 초과 시 1분 이내 알림 발송', '추론 응답시간 5초 초과 시 경고' 등 측정 가능한 기준으로 구체화하도록 유도해야 한다.

2) AI 고유 지표와 기존 IT 지표의 구분: 기존 IT 인프라 모니터링(CPU, 메모리, 디스크)은 일반 IT 감리에서 다루므로, 본 가이드라인에서는 GPU 사용률, 추론 Latency, 추론 큐 대기시간 등 AI 고유 지표의 모니터링 여부에 중점을 두어야 한다.

3) 시스템 접근이 제한될 경우의 대응: 감리원이 APM 대시보드에 직접 접근할 수 없는 경우, 사업자에게 임계값 초과 시 알림 발송 스크린샷, 알림 수신 이메일/SMS 스크린샷을 증빙 자료로 요청하여 검증한다.

4) 모델 버전 변경과 성능 회귀 탐지 연계: AI 시스템은 모델 업데이트가 빈번하므로, 배포 전후 성능 베이스라인 비교 체계가 수립되어 있는지 확인하는 것이 중요하다. 단순히 전통적인 성능 테스트만으로는 AI 모델 변경에 따른 성능 영향을 포착하기 어렵다.

5) 운영 전환 산출물 확인: 종료 단계에서는 구현 결과뿐만 아니라, 운영 단계에서 활용할 '운영 전환 계획서'에 모니터링 체계 이관 방안이 포함되었는지도 점검해야 한다. 모니터링 체계가 구축되었더라도 운영자에게 인수되지 않으면 무용지물이 된다.

AG_SD44-1-16 | 모니터링 요건 점검

시스템구조·보안 | 시험 | 운영·모니터링 | 모니터링/재학습

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MANAGE · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, AI 시스템이 운영 환경에 배포된 이후에도 모델 성능이 지속적으로 유지될 수 있도록 하는 지속적 모니터링 관리 계획이 {사업수행계획서}에 구체적으로 수립되어 있는지를 제3자적 관점에서 점검하는 것이 목적이다.

AI 모델은 배포 이후 입력 데이터의 변화, 사용 패턴의 변화, 외부 환경 변화 등으로 인해 성능이 점진적으로 저하되는 모델 드리프트(Model Drift)가 발생한다.

이를 조기에 감지하고 대응하기 위한 모니터링 체계가 계획되어 있는지를 검토하여, 사업 종료 후 AI 시스템의 신뢰성 저하와 서비스 품질 하락을 사전에 예방하는 데 초점을 둔다.

2. 필요성

▶ 일반 정보시스템과 달리 AI 시스템은 배포 이후에도 모델 성능이 시간이 지남에 따라 저하될 수 있으며, 이를 방치하면 서비스 품질 저하, 오결정 확대, 이용자 피해 발생으로 이어진다. 요구정의 단계에서 모니터링 체계가 계획되지 않으면 다음과 같은 리스크가 현실화된다.

▶ 모델 성능 저하가 발생해도 담당자가 인지하지 못하여 오결정이 장기간 지속된다.

▶ 모니터링 기준값(Baseline)과 임계값(Threshold)이 없어 성능 저하 판단 자체가 불가능해진다.

▶ 재학습(Retraining) 시점을 결정하는 트리거 기준이 없어 모델 갱신이 주먹구구식으로 이루어진다.

▶ 이상 발생 시 알림 체계가 없어 대응이 지연되고 피해가 확대된다.

3. 검토방법

No	점검 항목	확인 방법
1	모니터링 대상 성능 지표(KPI) 정의 여부	배포 후 지속적으로 측정할 AI 성능 지표(KPI)가 {사업수행계획서} 또는 "AI 모니터링 계획서"에 명시되어 있는지 확인
2	성능 기준값(Baseline) 및 임계값 설정 여부	{사업수행계획서} 또는 "성능 지표 정의서"에 각 KPI별 초기 기준값(Baseline)과 경보 임계값(Threshold)이 수치로 명시되어 있는지 확인

No	점검 항목	확인 방법
3	모니터링 주기 및 수행 방법 명시 여부	모니터링 수행 주기(일별·주별·월별 등)와 구체적인 수행 방법이 명시되어 있는지 확인
4	이상 감지 및 알림 체계 정의 여부	성능 임계값 초과 또는 이상 상황 발생 시 알림이 전달되는 체계(채널·대상·시한)가 정의되어 있는지 확인
5	모델 재학습·재배포 트리거 정의 여부	성능 저하 수준이 일정 기준에 도달하였을 때 모델 재학습 또는 재배포를 개시하는 조건(트리거)이 {사업수행계획서}에 정의되어 있는지 확인
6	모니터링 담당 조직 및 보고 체계 명시 여부	모니터링 활동을 수행할 담당 조직(부서·역할)과 결과를 보고하는 체계(보고 대상·주기·양식)가 {사업수행계획서}에 명시되어 있는지 확인
7	RFP 요건과 {사업수행계획서} 일치성	{제안요청서(RFP)}에 명시된 모니터링 관련 요건이 {사업수행계획서}에 빠짐없이 반영되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 특화 성능 지표(KPI), 수치 기반 기준값 및 임계값, 모니터링 주기·방법, 이상 알림 체계, 재학습 트리거, 담당 조직·보고 체계가 {사업수행계획서}에 모두 구체적으로 명시된 경우	현행 유지
보완권고 (△)	AI 특화 KPI 누락, 기준값·임계값 미수치화, 모니터링 주기 미명시, 재학습 트리거 미정의, 알림 대상·채널·시한 불명확 등 핵심 요소 중 일부가 결여된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	지속적 모니터링 계획 자체가 수립되지 않았거나, 도구 도입 계획만 있고 KPI·임계값·담당 조직 등 실질적 모니터링 계획 요소가 전무한 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) "모니터링 계획"과 "모니터링 도구 도입 계획"을 혼동하지 말라: "성능 모니터링 대시보드를 구축한다"고만 기술된 경우 이는 도구 도입 계획이지 모니터링 계획이 아니다. 무엇을(KPI), 어떤 기준으로(기준값·임계값), 얼마나 자주(주기), 누가(담당 조직), 이상 발생 시 어떻게(알림·에스컬레이션) 모니터링할 것인지가 모두 기술되어 있는지 확인하고, 하나라도 빠진 경우 보완권고로 판정한다.

2) 수치 없는 임계값 기술은 반드시 구체화를 권고하라: "성능이 현저히 저하되면 알림을 발송한다"와 같이 임계값이 수치로 명시되지 않은 경우, 이상 상황의 판단이 담당자의 주관에 전적으로

의존하게 되어 사업 종료 후 분쟁의 씨앗이 된다. "정확도 85% 미만 시 1차 알림, 80% 미만 시 즉시 재학습 개시"와 같이 2단계 이상의 수치 기반 임계값으로 구체화하도록 권고한다.

3) AI 고유 지표와 일반 IT 지표를 구분하여 점검하라: CPU·메모리·네트워크 등 인프라 지표만 계획되어 있고 모델 정확도, 드리프트 지표, 응답 품질 등 AI 특화 KPI가 없는 경우는 AI 모니터링 계획으로 인정하지 않는다.

4) 외부 SaaS 기반 AI 서비스의 경우 모델 관리 책임이 전적으로 공급업체에 있다면, SaaS SLA 계약에 모니터링 조건이 포함되어 있는지 확인한 후 그 결과를 판정에 반영하고 기록한다.

AG_SD44-1-17 | 모델 백업버전 복구-롤백테스트

시스템구조·보안 | 시험 | 운영·모니터링 | 모델관리

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MANAGE · ISO/IEC 27001

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 AI 모델에 대한 백업 및 버전 복구(롤백) 테스트를 계획대로 수행하였는지 여부를 제3자적 관점에서 점검하는 것을 목적으로 한다.

특히, AI 모델은 일반 소프트웨어와 달리 가중치·파라미터·학습 데이터·프롬프트 설정 등이 복합적으로 구성되어 있어, 단순한 소스 코드 롤백만으로는 이전 버전의 동작을 완전히 복구하기 어렵다.

이에 따라 {테스트결과서}를 통해 롤백 절차의 구체성, 복구 후 정상 동작 확인 여부, 그리고 AI 특유 구성 요소(모델 파일, 설정값, 추론 파라미터 등)의 버전 일치성이 검증되었는지를 확인한다.

2. 필요성

▶ AI 시스템 배포 이후 오작동, 성능 저하, 또는 안전·윤리 기준 위반이 발생할 경우, 신속하게 이전 버전으로 복구하여 서비스 연속성과 신뢰성을 유지하는 것이 필수적이다. 그러나 사전에 복구 절차가 검증되어 있지 않으면 실제 장애 발생 시 혼란과 서비스 중단이 장기화될 위험이 있다. 구체적인 필요성은 다음과 같다.

▶ AI 모델의 구성 복잡성: AI 모델은 모델 가중치 파일, 파인튜닝 파라미터, RAG 연계 설정, 프롬프트 템플릿, 추론 설정 등 다수의 구성요소로 이루어져 있다. 이들 구성요소가 하나라도 누락되거나 버전이 불일치하면 이전 버전과 동일한 동작을 보장할 수 없다.

▶ 운영 중 AI 위험 대응 필요성: 배포 후 환각(Hallucination), 편향 발현, 프롬프트 인젝션 취약 등 예기치 않은 위험이 확인될 경우, 즉시 이전 안전 버전으로 전환할 수 있는 복구 체계가 갖추어져 있어야 한다.

▶ 공공 AI 서비스의 연속성 요건: 공공 서비스에 활용되는 AI 시스템은 높은 가용성이 요구되므로, 배포 이전에 복구(롤백) 시나리오를 시험하고 그 결과를 문서화하여야 한다.

▶ TTA 및 NIST AI RMF의 관리 요건 반영: TTA 신뢰할 수 있는 인공지능 개발안내서 및 NIST AI RMF MANAGE 기능은 AI 시스템의 배포 후 사고 대응 및 복구 계획 수립과 검증을 명시적으로 요구한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 모델 백업 계획의 구체성 확인	{테스트결과서} 및 '배포 및 롤백 절차서'에 AI 모델의 구성요소별 백업 계획이 명시되어 있는지 확인
2	버전 관리 체계 수립 여부	'AI 모델 버전 관리 계획서'를 통해 모델 버전 식별 체계가 적정하게 수립되어 있는지 확인
3	롤백 트리거 조건 정의 여부	'배포 및 롤백 절차서' 또는 {테스트결과서}에 롤백을 실행할 기준 조건이 구체적으로 정의되어 있는지 확인
4	롤백 테스트 시나리오 구성의 적정성	{테스트결과서}에 롤백 테스트 시나리오가 AI 모델 특성을 반영하여 구성되어 있는지 확인
5	롤백 후 정상 동작 확인 기록의 충실성	{테스트결과서}에 롤백 실행 이후 이전 버전의 정상 동작이 확인되었는지 기록이 충실한지 점검
6	복구 목표 시간(RTO) 달성 여부	'배포 및 롤백 절차서'에 정의된 복구 목표 시간(RTO: Recovery Time Objective)이 실제 테스트를 통해 달성되었는지 확인
7	구성요소 버전 일치성 검증	롤백 후 모델 파일, 파라미터, 프롬프트, 연계 설정 등 AI 모델 구성요소의 버전이 동일한 이전 버전으로 일치하게 복구되었는지를 산출물과 실제 배포 환경의 비교를 통해 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 모델 구성요소별(가중치, 파라미터, 프롬프트, 연계 설정) 백업 계획과 버전 관리 체계가 수립되고, 롤백 트리거 조건·테스트 시나리오·롤백 후 정상 동작 확인·RTO 달성이 {테스트결과서}로 검증된 경우	현행 유지
보완권고 (△)	백업·롤백 체계는 수립되었으나 일부 구성요소(프롬프트 템플릿, 연계 설정 등)의 버전 관리가 누락되었거나, 롤백 조건이 정성적 표현에 그치거나 RTO 목표값과 실제 측정값의 비교가 수행되지 않은 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	AI 모델 백업·버전 관리 체계가 없거나 롤백 테스트가 수행되지 않아, 장애 발생 시 이전 버전으로의 복구를 보장할 수 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 소프트웨어 롤백과 AI 모델 롤백의 구분: 사업자가 "소스코드 버전 관리(Git)를 통해 롤백이 가능하다"고 주장하는 경우가 많다. 그러나 AI 모델 롤백은 소스코드만이 아니라 모델 가중치 파일, 파인튜닝 설정, 프롬프트 템플릿, RAG 지식베이스 버전까지 일치해야 한다. 감리원은 이 차이를 명확히 인식하고 AI 구성요소 전체의 버전 관리 및 롤백 범위를 점검해야 한다.

2) 롤백 조건의 수치화 유도: "성능이 저하되면 롤백한다"는 모호한 표현은 분쟁의 씨앗이 된다. 감리 시 "응답 정확도 N% 이하", "오류율 N% 이상 3분 지속" 등 측정 가능한 기준으로 구체화하도록 요구하여야 한다.

3) Human-in-the-Loop 검토자 역할 확인: AI 롤백 결정은 자동화된 기준만으로 처리되어서는 안 된다. 특히 공공 AI 서비스는 안전·윤리적 문제가 발생한 경우 반드시 인간 검토자가 롤백 여부를 최종 확인하는 절차가 있어야 한다. {테스트결과서}에 Human-in-the-Loop 검토 기록이 포함되어 있는지 확인한다.

4) 외부 API 기반 AI의 버전 관리 한계 인식: OpenAI, HyperCLOVA 등 외부 API를 활용하는 경우, 공급사가 API 버전을 변경하면 사업자 단독으로 이전 버전을 완전히 복구하는 것이 어렵다. 이 경우 '배포 및 롤백 절차서'에 API 버전 고정 설정(예: model=gpt-4-0613 등 버전 명시) 또는 대체 조치가 명시되어 있는지 확인해야 한다.

5) 복구 목표 시간(RTO) 실효성 점검: {테스트결과서}에 롤백 소요 시간이 기록되어 있어도, 그 시간이 설계서의 RTO 목표와 비교되지 않으면 의미가 없다. 감리원은 반드시 설계 목표값과 실제 측정값을 대조하여 미달성 여부 및 조치 계획을 확인한다.

AG_SD44-1-18 | 서버 리소스 쿼터점검

시스템구조·보안 | 시험 | 시스템 구현 | GPU/NPU연산

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · OWASP LLM Top 10 · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서, 설계 단계에서 정의된 서버 리소스 쿼터(Quota) 통제 정책과 단일 프로세스의 GPU 메모리 독점 차단 설정이 실제 시스템에 올바르게 구현되었는지를 {테스트결과서} 등 산출물을 통해 검증하기 위한 점검 방향과 기준을 제시한다.

AI 추론/학습 워크로드는 대량의 GPU 메모리와 컴퓨팅 리소스를 소비하므로, 특정 프로세스가 리소스를 독점할 경우 다른 서비스의 응답 지연이나 장애로 이어질 수 있다.

본 점검은 이러한 리소스 독점 위험이 시험을 통해 검증되었는지, 설계대로 통제 메커니즘이 동작하는지를 확인하는 데 목적이 있다.

2. 필요성

▶ 생성형 AI 시스템은 일반 정보시스템과 달리 모델 추론 시 수십 GB 단위의 GPU 메모리를 점유하며, 학습(Fine-tuning) 작업 시에는 이보다 훨씬 많은 리소스를 소비한다.

▶ 단일 프로세스가 GPU 전체 메모리를 독점하면 동일 서버에서 구동되는 다른 AI 서비스나 배치 작업이 OOM(Out of Memory) 오류로 중단될 수 있다.

▶ CPU, 메모리, 네트워크 대역폭 등 서버 리소스에 대한 쿼터(Quota) 제한이 없으면, 특정 컨테이너 또는 Pod가 노드 전체 자원을 소진하여 다른 워크로드의 스케줄링이 불가능해지는 노이지 네이버(Noisy Neighbor) 문제가 발생한다.

▶ 따라서 종료(시험/전개) 단계에서 이러한 통제 메커니즘의 실제 작동 여부를 시험 결과를 통해 검증할 필요가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	서버 리소스 쿼터(Quota) 초과 통제 시험	컨테이너/Pod별 CPU, 메모리, GPU 리소스 쿼터(Request/Limit) 설정이 설계대로 적용되었는지 시험 결과를 확인
2	쿼터 초과 시 제한 동작 검증	리소스 쿼터를 초과하는 요청 시 시스템이 설계대로 요청 거부, 스케줄링 대기, 또는 프로세스 종료 등의 제한 동작을 수행하는지 시험 결과를 확인

No	점검 항목	확인 방법
3	GPU 메모리 독점 차단 설정 검증	단일 프로세스가 GPU 메모리 전체를 점유하지 못하도록 상한 제한이 설정되었는지, 해당 제한이 실제 동작하는지 시험 결과를 확인
4	리소스 독점 시나리오 부하시험	복수의 AI 워크로드가 동시에 실행되는 환경에서 특정 워크로드의 리소스 독점이 다른 워크로드에 미치는 영향을 시험한 결과가 있는지 확인
5	설계 대비 구현 일치성	{시스템아키텍처설계서}에 정의된 리소스 통제 정책과 실제 구현/시험 결과 간의 일치 여부를 대조 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	리소스 쿼터(Quota) 초과 통제 및 GPU 메모리 독점 차단 설정이 설계대로 구현되었고, 시험이 수행되어 모든 시나리오에서 Pass 판정을 받은 경우	현행 유지
보완권고 (△)	주요 시험은 수행되었으나, 일부 시나리오 누락(예: 동시 다중 워크로드 부하시험 미수행), 시험 증빙 부족, 또는 설계 대비 경미한 설정값 차이가 있는 경우	개선 권고 사항을 구체적으로 기술하고 해당 항목 보완 요청 후 재확인
부적합 (X)	리소스 쿼터 관련 시험이 수행되지 않았거나 GPU 메모리 독점 차단 시험이 전혀 없는 경우, 또는 시험 결과 Fail이 다수이나 조치 이력이 없는 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 시험 시나리오의 현실성 확인: 리소스 쿼터 초과 시험이 단순히 '설정값 확인'에 그치지 않고, 실제로 쿼터를 초과하는 부하를 발생시켜 시스템 반응을 확인하는 능동적 시험(Active Test)이 수행되었는지를 확인한다. 단순 설정 확인만으로는 실제 운영 환경에서의 동작을 보장할 수 없다.

2) 설계서와 시험 설정값의 정량적 대조: 단순 CPU/Memory Quota가 아닌, 쿠버네티스 환경 내 GPU 자원 분할 기술(예: NVIDIA MIG(Multi-Instance GPU), vGPU, Time-slicing 설정)이 아키텍처 설계서와 실제 인프라 설정(Manifest 파일 등)에 일치하게 반영되었는지, 단일 Pod가 할당량 이상의 VRAM을 점유하지 못하도록 제한되어 있는지 정량적으로 대조 확인한다.

3) GPU 독점 차단의 AI 특수성: GPU 메모리 독점 차단은 일반 IT 인프라에는 없는 AI 특화 상세 검토 항목이다. 사업자가 '일반 서버 리소스 모니터링으로 대체 가능'이라고 주장할 수 있으나, GPU 메모리는 일반 시스템 메모리와 관리 체계가 다르므로 별도의 시험이 반드시 필요하다.

4) 시험 증빙의 객관성: 시험 결과의 객관성을 확보하기 위해, 시험 수행 시점의 시스템 리소스 현황을 보여주는 증빙(모니터링 대시보드 화면 캡처, 명령어 실행 결과 등)이 {테스트결과서}에 첨부되어 있는지 확인한다. 텍스트 기술만으로는 시험의 실제 수행 여부를 판단하기 어렵다.

5) 운영 단계 전환 준비 확인: 종료 단계에서는 시험 결과뿐 아니라, 운영 단계에서 리소스 독점을 지속적으로 통제하기 위한 준비 사항도 확인한다. 'AI 인프라 운영 매뉴얼'에 리소스 임계값 초과 시 알림 설정, 자동 스케일링 정책, 긴급 대응 절차 등이 기술되어 있는지 검토한다.

AG_SD44-1-19 | 운영매뉴얼 목차 설계 점검

시스템구조·보안 | 시험 | 운영·모니터링 | 안정화/하자보수

근거 | 「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MANAGE · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서 {운영자매뉴얼}의 목차가 요구사항 분석 단계에서 정의된 내용을 빠짐없이 구체적으로 반영하였는지를 감리원의 관점에서 점검하는 데 목적이 있다.

특히 AI 시스템에 특화된 두 가지 핵심 항목—모델 재학습(반복 튜닝) 절차와 AI 관련 장애 대응 절차—이 {운영자매뉴얼} 목차에 명시되어 있는지를 확인함으로써, 운영 단계에서 발생할 수 있는 매뉴얼 공백으로 인한 위험을 사전에 차단하는 것이 핵심 점검 목표이다.

2. 필요성

▶ 일반적인 정보시스템은 배포 이후 소프트웨어 버전이 고정되는 반면, AI 시스템은 서비스 운영 중에도 데이터 편향·성능 저하·출력 오류 등의 이유로 모델을 주기적으로 재학습·재배포해야 한다. 이러한 특성 때문에 {운영자매뉴얼}에는 기존 IT 시스템 운영 절차 외에 AI 고유의 운영 항목이 반드시 포함되어야 한다.

▶ 설계 단계에서 {운영자매뉴얼}의 목차가 부실하게 작성되면, 이후 구현·시험 단계를 거쳐 최종 산출물로 납품되는 매뉴얼에서 해당 항목이 아예 누락될 가능성이 높다. 감리원은 이 시점에 목차 수준의 완결성을 검토함으로써 구현 이후의 재작업 비용을 최소화할 수 있다.

▶ TTA 신뢰할 수 있는 인공지능 개발안내서(생성AI 기반 서비스 분야)는 운영 단계에서 AI 모델의 모니터링, 재학습, 사용자 피드백 반영 절차를 문서화할 것을 명시하며, NIST AI RMF의 MANAGE 4.1 및 4.3은 사고 대응, 복구, 변경관리 계획의 문서화를 요구하고 있다. 감리원은 이들 기준이 설계 단계 산출물인 {운영자매뉴얼} 목차에 반영되었는지 확인할 필요가 있다.

3. 검토방법

No	점검 항목	확인 방법
1	모델 재학습 절차 목차 포함 여부	{운영자매뉴얼} 목차 내 "모델 재학습", "학습 갱신", "모델 갱신" 또는 유사 항목 존재 여부를 직접 확인
2	AI 관련 장애 대응 절차 목차 포함 여부	{운영자매뉴얼} 목차 내 "AI 장애", "모델 이상", "비정상 응답 처리" 또는 유사 항목 존재 여부 확인
3	AI 출력 모니터링 및 품질 기준 목차 포함 여부	{운영자매뉴얼} 목차 내 "AI 성능 모니터링", "출력 품질 관리" 또는 유사 항목 존재 여부 확인

No	점검 항목	확인 방법
4	요구사항과 {운영자매뉴얼} 목차 간 추적성 확인	{요구사항정의서}의 AI 관련 운영 요구사항 항목과 {운영자매뉴얼} 목차를 1:1 대조하여 미반영 항목 식별
5	사용자 피드백 및 이의제기 처리 절차 목차 포함 여부	{운영자매뉴얼} 목차 내 "사용자 피드백", "AI 결과 이의제기", "재검토 요청 처리" 항목 존재 여부 확인
6	모델 버전 관리 및 롤백 절차 목차 포함 여부	{운영자매뉴얼} 목차 내 "모델 버전 관리", "롤백", "이전 모델 복구" 항목 존재 여부 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	{운영자매뉴얼} 목차에 모델 재학습 절차, AI 관련 장애 대응 절차, AI 출력 모니터링·품질 기준, 사용자 피드백·이의제기 처리, 모델 버전 관리·롤백 절차가 모두 포함되어 있고, {요구사항정의서}의 운영 요건과 목차 간 추적성이 확보된 경우	현행 유지
보완권고 (△)	목차에 주요 항목은 존재하나 환각 대응·편향 출력 탐지 등 일부 AI 특화 항목이 누락되었거나, 해당 절이 "추후 상세 작성 예정"으로만 기재되어 내용이 미확정인 경우	해당 항목 보완 요청 및 최종 납품 전 완성 일정 확인 후 재확인
부적합 (X)	{운영자매뉴얼} 목차에 모델 재학습 절차 또는 AI 관련 장애 대응 절차가 포함되지 않았거나, 일반 IT 시스템 운영 항목만으로 구성되어 AI 고유 운영 항목이 전혀 반영되지 않은 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 목차 제목의 모호성에 주의: "7장. 운영 절차"와 같이 포괄적인 장 제목만 있고 AI 관련 하위 항목이 없는 경우, 겉보기에는 목차가 완성된 것처럼 보이지만 실제로는 AI 특화 운영 항목이 누락된 것이다. 감리원은 반드시 2단계 이하 목차(소절 수준)까지 확인해야 한다.

2) "추후 작성" 항목의 처리 방법: 목차에 해당 절이 존재하지만 "추후 상세 작성 예정"이라고만 기재된 경우, 목차에 항목은 있으나 내용이 미확정된 상태이다. 해당 항목은 '보완권고(△)'로 판정하고, 최종 납품 전 완성 일정을 사업자로부터 확인받아야 한다.

3) 재학습 절차와 장애 대응 절차의 구분: 서버 다운과 같은 IT 인프라 장애와 달리, 시스템은 정상 가동되나 AI가 엉뚱한 답변을 지속해서 생성하는 'AI 품질 장애(모델 오류)' 상황을 별도의 장애 유형으로 정의하고 대응(예: 롤백, 안전모드 전환)하는 목차가 있는지 점검해야 한다. IT 시스템 장애와 AI 장애의 개념 차이를 감리원과 사업자가 명확히 인식하도록 해야 한다.

4) AI 고유 장애 유형의 식별: 감리원은 {운영자매뉴얼} 목차에 "환각(Hallucination) 발생 시 처리", "편향 출력 탐지 및 대응" 등 AI 특화 장애 유형이 명시되어 있는지 확인해야 한다. 일반 IT 장애(시스템 다운, 네트워크 오류)만 다루는 장애 대응 절차는 AI 서비스 운영에 부적합하다.

5) 추적성 매트릭스 활용: 점검 결과를 감리 보고서에 첨부할 때 추적성 매트릭스를 활용하면 미반영 항목을 명확히 제시할 수 있다. "일부 항목 누락"과 같은 모호한 표현 대신, "OP-AI-003 항목({운영자매뉴얼} 7장에 미반영)"과 같이 구체적으로 기술하는 것이 분쟁 방지에 유효하다.

AG_SD44-1-20 | 오픈소스 라이선스 충돌점검

시스템구조·보안 | 시험 | 시스템 구현 | 저작권/라이선스

근거 | 「저작권법」·「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

AI 서비스 구축 사업의 구현 완료된 시스템에 포함된 오픈소스 소프트웨어(OSS)의 라이선스 유형이 상용 서비스 배포 조건과 충돌하지 않는지를 검증하는 것이 목적이다.

특히 AI 시스템은 학습 프레임워크(PyTorch, TensorFlow 등), 추론 엔진(ONNX Runtime, TensorRT 등), 데이터 전처리 라이브러리, 사전학습 모델(Pre-trained Model) 등 다수의 오픈소스 구성요소가 중첩적으로 결합되며, 이들 구성요소 간 라이선스 조건의 상호 충돌(GPL/AGPL의 Copyleft 전파 등)이 발생할 경우 상용 서비스 배포가 법적으로 제한될 수 있다.

본 가이드라인은 이러한 AI 고유의 오픈소스 라이선스 리스크가 {결과보고서}에 적정하게 반영되었는지를 점검하기 위한 감리 기준을 제시한다.

2. 필요성

▶ AI 시스템은 일반 정보시스템과 달리 수십~수백 개의 오픈소스 라이브러리가 의존성 체인(Dependency Chain)으로 연결되며, 사전학습 모델 자체에도 별도의 라이선스(예: Llama Community License, CreativeML Open RAIL-M 등)가 부여된다. 이러한 복합적인 라이선스 구조에서 GPL/AGPL 등 Copyleft 라이선스가 하나라도 포함되면 전체 서비스의 배포 조건에 영향을 미칠 수 있다.

▶ 공공 AI 서비스가 이러한 라이선스 충돌을 간과한 채 배포될 경우, 발주기관이 법적 분쟁에 노출되거나, 서비스 중단 또는 전면 재개발이 필요한 심각한 사업 리스크로 이어질 수 있다. 따라서 오픈소스 라이선스 충돌 여부를 체계적으로 검증하고 그 결과를 {결과보고서}에 명확히 기록하는 것이 필수적이다.

3. 검토방법

No	점검 항목	확인 방법
1	SBOM 작성 완전성 및 법무 검토 결과 확인	AI 프레임워크, 추론 엔진, 전처리 라이브러리, 유틸리티 등 전체 구성요소가 SBOM에 빠짐없이 등재되어 있는지, 모델 라이선스의 상업적 이용 제약(예: Llama의 MAU 제한)과 외부 데이터의 저작권 위배 소지가 법무 검토 결과서에서 이상 없음으로 확인되었는지 검토
2	AI 모델 라이선스 식별	사전학습 모델(Foundation Model), 파인튜닝 모델, 임베딩 모델 등의 라이선스 유형이 식별되고, 상용 이용 가능 여부가 검토되었는지 확인

No	점검 항목	확인 방법
3	GPL 계열 충돌 점검	GPL/LGPL/AGPL 라이선스 구성요소가 상용 배포 코드와 정적/동적 링크로 결합되어 Copyleft 전파가 발생하는지 분석 결과 확인
4	라이선스 의무사항 이행	저작권 고지, 라이선스 전문 포함, 소스코드 제공 안내 등 개별 라이선스가 요구하는 의무사항 이행 계획 또는 결과 확인
5	SCA 도구 활용 점검	자동화된 SCA(Software Composition Analysis) 도구를 활용하여 라이선스 충돌을 점검한 결과가 있는지, 도구 점검 범위가 AI 구성요소를 포함하는지 확인
6	충돌 해소 조치 확인	라이선스 충돌이 발견된 경우, 대체 라이브러리 교체 또는 격리(분리 실행) 등 해소 조치가 수행되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	SBOM이 AI 구성요소를 포함하여 완전하게 작성되었고, GPL/AGPL 계열 라이선스 충돌이 없으며, 모든 라이선스 의무사항이 이행되었고, SCA 도구 점검이 수행되어 {결과보고서}에 적정하게 기록된 경우	현행 유지
보완권고 (△)	SBOM은 작성되었으나 일부 AI 구성요소(예: 모델 서빙 도구, 전처리 라이브러리 등)가 누락되었거나, 라이선스 충돌 분석이 부분적으로 수행되었거나, SCA 도구 점검 범위가 AI 구성요소를 완전히 포괄하지 못한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	SBOM이 미작성되었거나, GPL/AGPL 계열 라이선스가 상용 코드에 Copyleft 전파를 유발하는 상태로 방치되었거나, 비상업적 전용 모델이 상용 서비스에 포함된 상태에서 해소 조치가 없는 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 모델 라이선스와 코드 라이선스의 분리 확인: AI 시스템에서는 학습 코드(예: Apache 2.0)와 모델 가중치(예: Non-Commercial License)에 서로 다른 라이선스가 적용되는 경우가 빈번하다. SBOM에 코드 라이선스만 기재하고 모델 라이선스를 누락하는 사례가 많으므로, 반드시 모델 가중치의 라이선스 조건을 별도로 확인해야 한다.

2) AGPL의 네트워크 서비스 조항 주의: 공공 AI 서비스가 API 형태로 제공되는 경우, AGPL 라이선스 구성요소가 포함되면 전체 서비스 소스코드를 공개해야 할 수 있다. 일반 GPL에서는 문제되

지 않는 SaaS/API 서비스 형태가 AGPL에서는 소스코드 공개 의무를 발생시키므로 특별히 주의해야 한다.

3) SCA 도구의 AI 구성요소 탐지 한계 인식: 대부분의 SCA 도구는 npm, pip, Maven 등 표준 패키지 관리자 기반 의존성 분석에 최적화되어 있다. HuggingFace Model Hub에서 다운로드한 모델, 직접 빌드한 CUDA 커널, Jupyter Notebook 내 인라인 코드 등은 SCA 도구의 자동 탐지 범위를 벗어나므로, 이러한 구성요소는 수작업 목록화 등 별도의 보완 점검이 필요하다.

4) 전이 의존성(Transitive Dependency)까지 점검 범위에 포함: AI 프레임워크는 수십 개의 하위 라이브러리에 의존하며, 이 하위 라이브러리가 다시 다른 라이브러리에 의존하는 다단계 의존성 구조를 가진다. 직접 의존성(Direct Dependency)만 점검하면 GPL 계열 라이선스가 전이 의존성에 숨어 있는 경우를 놓칠 수 있다.

5) 라이선스 충돌 시 대안 확인: GPL 계열 충돌이 발견된 경우, 동일 기능의 Permissive 라이선스(MIT, BSD, Apache 2.0 등) 대체 라이브러리가 존재하는지, 또는 프로세스 격리(별도 컨테이너로 분리 실행)를 통해 Copyleft 전파를 차단할 수 있는지를 확인하도록 유도한다.

AG_SD44-1-21 | 위험관리대응매뉴얼 점검

시스템구조·보안 | 시험 | 운영·모니터링 | 안정화/하자보수

근거 | 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조·제34조 · 「인공지능기본법 이행 가이드라인」(과기정통부, 2026.1.) · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 사업자가 운영 단계 돌입 전에 작성하는 {운영자매뉴얼} 및 'AI 위험관리 대응 매뉴얼'이 AI 특유의 장애·편향·오출력 사고에 대한 대응 절차를 충분하고 실행 가능한 수준으로 담고 있는지를 감리원이 제3자적 관점에서 검토하는 것을 목적으로 한다.

특히, 일반적인 IT 장애 대응 절차와 구별되는 AI 특화 사고유형(환각·편향 출력·프롬프트 인젝션 등)에 대한 탐지·격리·복구·재발방지 절차가 산출물에 명시되었는지, Human-in-the-Loop 체계 및 긴급 서비스 중단(kill switch) 절차가 구체적으로 기술되어 있는지 여부를 확인한다.

2. 필요성

▶ AI 시스템은 일반 소프트웨어와 달리 데이터·모델의 변화에 따라 출력 품질이 지속적으로 변동하며, 편향 증폭, 환각, 프롬프트 인젝션 등 AI 고유의 사고 유형이 운영 중 언제든 발생할 수 있다. 이러한 사고는 사용자 피해·기관 신뢰 훼손·법적 책임으로 이어질 수 있음에도, 기존 IT 장애 대응 매뉴얼은 이러한 AI 특화 사고 유형을 다루지 못한다.

▶ 종료 단계는 실제 운영 이전의 마지막 점검 기회로, 이 단계에서 {운영자매뉴얼} 및 'AI 위험관리 대응 매뉴얼'의 완전성을 확인하지 않으면 운영 중 사고 발생 시 대응 공백이 생길 수 있다. 감리원은 매뉴얼의 AI 특화 대응 절차 포함 여부 및 실행 가능성을 문서 검토와 인터뷰를 통해 검증하여야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 사고 유형별 대응 절차 명시 여부	{운영자매뉴얼} 또는 'AI 위험관리 대응 매뉴얼'이 환각·편향 출력·프롬프트 인젝션·드리프트 등 AI 고유 사고 유형별 대응 절차를 독립적으로 기술하고 있는지 확인
2	Human-in-the-Loop 체계 및 전환 기준 명시	AI 판단을 사람이 검토·승인하는 개입 시점과 기준(예: 신뢰도 임계값 이하 시 인간 검토 요청)이 매뉴얼에 명시되었는지 확인한다. 완전 자율 운영 구간과 인간 개입 구간이 명확히 구분되어 있는지 확인

No	점검 항목	확인 방법
3	Kill Switch 및 서비스 중단 절차 완전성	이상 출력 또는 보안 위협 감지 시 AI 서비스를 즉각 차단할 수 있는 Kill Switch 절차와 권한 체계(누가, 언제, 어떻게 발동하는지)가 구체적으로 기술되어 있는지 확인
4	사고 탐지 기준 및 임계값 정량화 여부	'AI 시스템 모니터링 계획서' 또는 {운영자매뉴얼}에 사고 탐지를 위한 정량적 기준(예: 환각 발생률 X% 초과, 편향 지표 Y 이상 시 경보)이 명시되어 있는지 확인
5	편향 사고 대응 및 재발방지 절차 구체성	편향 출력 사고 발생 시 ① 탐지 방법, ② 영향 범위 분석, ③ 서비스 임시 중단 또는 출력 필터 적용, ④ 모델 재학습·재배포 절차, ⑤ 피해 사용자 통지 절차가 순서대로 기술되어 있는지 확인
6	사고 이력 기록 및 보고 체계 명시	사고 발생 시 이력 로그 기록 방식, 보고 대상(내부 책임자, 발주기관), 보고 시한이 매뉴얼에 명시되어 있는지 확인한다. 사고 기록이 추적 가능한 형태로 보존되도록 절차가 수립되어 있는지 확인
7	요구사항-매뉴얼 간 추적성(RTM) 연결 확인	앞 단계(분석·설계)에서 도출된 AI 위험 요구사항이 {운영자매뉴얼}의 대응 절차와 연결되어 있는지 RTM 또는 교차 참조표를 통해 확인한다. 누락된 위험 항목이 있으면 미흡으로 판정

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	AI 사고 유형별 대응 절차(탐지·격리·복구·재발방지)가 모두 명시되고, HITL 개입 기준·Kill Switch 절차·사고 보고 체계가 구체적으로 기술되어 있으며, 요구사항과의 추적성이 확인된 경우	현행 유지 (우수 사례는 특이사항으로 기록 가능)
보완권고 (△)	AI 사고 유형 중 1개 이상이 대응 절차에서 누락되거나, Kill Switch 절차가 불명확하거나, 탐지 기준이 정성적으로만 기술된 경우	해당 항목 보완 요청 후 재확인 (구체적 보완 사항을 점검결과 개선방향에 명시)
부적합 (X)	AI 특화 대응 절차가 전무하거나, 일반 IT 장애 대응 절차를 그대로 유용한 경우, 또는 HITL·Kill Switch 관련 내용이 전혀 없는 경우	시정조치 요구 및 이행결과 확인 (필요 시 재감리 여부 검토 권고)

5. 감리 시 유의사항 및 실무 Tip

1) AI 사고 유형의 경계 구분: 환각·편향·드리프트는 서로 원인과 대응 방법이 다르다. 매뉴얼에서 이들을 단일 "오류" 항목으로 묶어 처리하는 경우, AI 특화 대응 절차로 인정하지 않는다. 유형별 독립 항목 구성 여부를 반드시 확인한다.

2) 사업관리/품질보증 영역의 경계 준수: 감리원은 매뉴얼의 AI 모델 성능·기술적 탐지 알고리즘의 적절성을 평가하지 않는다. 점검 범위는 "절차의 존재와 완전성"에 한정한다. 모델 성능 수치의 타당성 평가는 응용시스템 감리원 영역이다.

3) Kill Switch 실효성 확인: 문서에 기술되어 있더라도 담당자가 발동 절차를 실제로 인지하지 못하면 실효성이 없다. 인터뷰를 통해 운영 담당자의 실제 이해도를 확인한다.

4) 운영 단계 점검사항과 혼동 주의: 본 가이드라인은 매뉴얼 "문서의 완전성"을 점검하는 것이며, 운영 중 실제 모니터링 수행 여부는 점검 범위에서 제외한다. 단, 운영을 위한 모니터링 계획이 수립되었는지는 포함된다.

5) 탐지 기준 수치화 유도: "이상 출력 발생 시 조치"와 같은 모호한 기술은 실제 운영에서 분쟁의 원인이 된다. 감리 현장에서 "환각 발생률 3% 이상 시 서비스 일시 정지" 등과 같이 정량적 임계값으로 구체화하도록 권고한다.

AG_SD45-1-01 | CI/CD 배포 파이프라인 자동화 점검

시스템구조·보안 | 전개 | 운영·모니터링 | CI/CD파이프라인

근거 | 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업에서, 설계된 CI/CD 배포 파이프라인이 실제 시스템으로 구현되어 정상 동작하는지, 빌드-테스트-배포 전 과정에 AI 모델 보안 점검이 자동으로 연동되어 있는지를 제3자적 관점에서 확인하는 것을 목적으로 한다.

특히, AI 시스템의 CI/CD 파이프라인은 일반 소프트웨어와 달리 AI 모델 아티팩트(가중치 파일, 설정 파일, 프롬프트 템플릿 등)의 버전 관리와 무결성 검증, 모델 배포 전 성능·안전성·편향성 자동 테스트, 컨테이너 이미지 취약점 스캔 등 AI 고유의 보안 파이프라인에 통합되어 있는지를 확인한다.

2. 필요성

▶ AI 서비스의 CI/CD 파이프라인은 일반 애플리케이션 배포와 달리 AI 모델 아티팩트(학습 가중치, 토크나이저, 프롬프트 템플릿 등)를 함께 관리해야 하므로, 배포 과정에서 모델 파일 위변조, 악성 코드 삽입, 무허가 모델 교체 등의 고유한 위험이 존재한다.

▶ 또한, AI 모델은 배포 전 성능 검증(정확도, 응답 품질), 안전성 검증(탈옥 테스트, 금칙어 필터링), 편향성 검증(공정성 테스트) 등이 자동화된 파이프라인 내에서 수행되어야 하며, 이를 통과하지 못한 모델이 운영 환경에 배포되는 것을 방지하는 게이트(Gate) 메커니즘이 필수적이다. 이러한 검증 절차가 빠지면 장애, 보안 사고, 품질 미달 모델의 무단 배포 등 심각한 운영 리스크로 이어질 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	CI/CD 파이프라인 구현 및 자동화 동작 확인	설계서 대비 빌드-테스트-배포 전 과정의 자동화 구현 여부, 파이프라인 실행 로그 확인
2	AI 모델 아티팩트 버전 관리 및 무결성 검증	모델 가중치/토크나이저/프롬프트 버전 해시 검증, 배포 시 무결성 체크 자동화 여부
3	배포 전 AI 모델 보안·안전성 자동 테스트 연동	탈옥 테스트, 금칙어 필터링, 편향성 검증, 컨테이너 취약점 스캔 등 보안 게이트 연동 여부

No	점검 항목	확인 방법
4	배포 승인 절차 및 롤백 계획 수립	'배포 계획서'에서 모델 배포 승인 책임자가 지정되고, 블루/그린 배포 전략과 즉시 롤백 절차가 정의되었는지 확인하며, 최근 배포 건의 승인·롤백 이력을 대조
5	배포 파이프라인 접근통제 및 감사추적성	파이프라인 실행 권한 통제, 배포 이력 로깅, 책임 추적 경로 확보

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	CI/CD 파이프라인이 설계대로 구현되어 정상 동작하고, AI 모델 보안·안전성 게이트가 자동 연동되며, 배포 승인/롤백 절차가 수립되어 있는 경우	현행 유지
보완권고 (△)	파이프라인은 구현되었으나, AI 모델 보안 게이트 일부 미연동 또는 롤백 절차가 미비한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	CI/CD 파이프라인이 미구현되었거나 수동 배포만 수행되고, 보안 검증 연동이 전혀 없는 경우	시정조치 요구 및 이행결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 모호한 배포 요건의 수치화: 'CI/CD를 적용한다', '자동 배포를 한다'와 같은 모호한 문구는 분쟁의 원인이 된다. '모델 배포 전 탈옥 테스트 100건 이상 자동 수행', 'Critical CVE 발견 시 배포 자동 차단' 등 측정 가능한 기준으로 구체화하도록 유도해야 한다.

2) AI 고유 배포 위험과 일반 SW 배포의 구분: 일반 애플리케이션 CI/CD 점검은 IT 감리 영역에서 다루므로, 본 가이드라인에서는 AI 모델 아티팩트의 버전 관리, 모델 무결성 검증, 탈옥/편향성/금칙어 테스트 자동화 등 AI 고유의 배포 보안 요소에 중점을 두어야 한다.

3) 파이프라인 실행 로그 확인 시 실제 동작 검증: 사업자가 제출한 산출물(파이프라인 설계서)만으로는 실제 동작 여부를 판단하기 어렵다. 가능한 경우 CI/CD 도구(Jenkins, GitLab CI, GitHub Actions 등)의 실행 이력 화면을 실제 확인하여 파이프라인 자동화 동작을 검증해야 한다.

4) 모델 버전 관리의 특수성 이해: AI 모델 가중치 파일은 수GB~수백GB에 달하므로 Git 같은 일반 버전 관리 도구로는 관리가 불가능하다. MLflow, DVC, Hugging Face Hub 등 AI 모델 전용 레지스트리를 통한 버전 관리 체계가 수립되어 있는지 확인해야 한다.

5) 보안 게이트 실패 시 배포 차단 여부 확인: 보안 테스트(탈옥, 취약점 스캔 등)가 파이프라인에 형식적으로만 연동되어 있고 실제로는 실패 시에도 배포가 진행되는 경우가 있다. 보안 게이트 실패 시 배포가 실제로 차단되는지 파이프라인 실행 이력을 통해 반드시 검증해야 한다.

AG_SD45-1-02 | 컨테이너이미지 보안점검

시스템구조·보안 | 전개 | 운영·모니터링 | 시큐어코딩

근거 | 「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.) · 「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.) · ISO/IEC 27001

1. 목적

본 가이드라인은 AI 서비스 구축 사업에서, 컨테이너 기반으로 배포되는 AI 추론(Inference) 서비스 이미지에 악성코드가 포함되어 있지 않은지, 그리고 런타임 환경에서 구동될 컨테이너 이미지에 알려진 보안 취약점(CVE)이 존재하지 않는지를 체계적으로 확인하기 위한 감리 점검 기준을 제시한다.

AI 서비스는 모델 서빙(Model Serving), 전처리/후처리 파이프라인, API Gateway 등 다수의 컨테이너 이미지가 조합되어 운영되므로, 단일 이미지가 아닌 서비스를 구성하는 전체 이미지 셋(Image Set)에 대한 보안 검증이 완료되었는지를 확인하는 것이 핵심이다.

악성코드가 내포된 이미지가 운영 환경에 배포될 경우, AI 모델 가중치(Weight) 탈취, 학습 데이터 유출, 추론 결과 조작 등 AI 특유의 보안 위협이 발생할 수 있다.

2. 필요성

▶ AI 서비스 컨테이너 이미지는 일반 웹 애플리케이션과 달리 GPU 드라이버, CUDA 라이브러리, AI 프레임워크(PyTorch, TensorFlow 등), 모델 서빙 엔진(Triton, vLLM 등) 등 다양한 오픈소스 컴포넌트를 포함하고 있어, 각 계층(Layer)에 알려진 취약점(CVE)이 잔존할 가능성이 높다.

▶ AI 모델 가중치 파일(.pt, .onnx, .safetensors 등)이 컨테이너 내부에 탑재되는 경우가 많아, 이미지가 악성코드에 감염되면 모델 자체의 무결성이 훼손되거나 백도어(Backdoor)가 삽입될 위험이 있다.

▶ 공급망 공격(Supply Chain Attack)을 통해 베이스 이미지나 AI 프레임워크 배포 채널 자체가 오염될 수 있으므로, 이미지 출처와 무결성에 대한 검증이 필요하다.

▶ 이러한 점검이 누락될 경우, 운영 환경에서 AI 추론 결과의 신뢰성 저하, 개인정보 및 학습 데이터 유출, 서비스 가용성 훼손 등의 보안 사고로 이어질 수 있다.

3. 검토방법

No	점검 항목	확인 방법
1	컨테이너 이미지 악성코드 미검출	AI 서비스를 구성하는 전체 컨테이너 이미지에 대해 악성코드 스캔이 수행되었고, 악성코드가 검출되지 않았는지 확인

No	점검 항목	확인 방법
2	런타임 이미지 취약점 스캔 결과 위험 없음	운영 환경에 배포될 최종 이미지를 대상으로 CVE 취약점 스캔을 수행하고, Critical/High 등급 취약점이 조치 완료(패치 또는 예외 승인)되었는지 확인
3	AI 프레임워크/라이브러리 취약점 점검	컨테이너 이미지 내 AI 프레임워크(PyTorch, TensorFlow 등), 모델 서빙 엔진, GPU 드라이버 등 AI 특화 컴포넌트의 알려진 취약점 점검 여부 확인
4	배포 파이프라인 내 보안 스캔 통합	CI/CD 배포 파이프라인에 컨테이너 이미지 보안 스캔이 설계대로 구현되어 있고, 스캔 미통과 시 배포가 차단되는지 확인
5	베이스 이미지 출처 및 무결성 확인	AI 서비스 컨테이너의 베이스 이미지가 신뢰할 수 있는 공식 레지스트리에서 가져왔으며, 이미지 서명(Image Signing) 또는 해시(Digest) 검증이 수행되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	전체 컨테이너 이미지에 대해 악성코드 스캔 및 CVE 취약점 스캔이 수행되었고, Critical/High 취약점이 모두 조치 완료되었으며, 배포 파이프라인에 보안 스캔이 통합되어 자동 차단 정책이 적용된 경우	현행 유지
보완권고 (△)	보안 스캔은 수행되었으나, 일부 이미지(사이드카, 모니터링 에이전트 등)가 스캔 대상에서 누락되었거나, Medium 등급 이하 취약점의 조치 계획이 미수립된 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	컨테이너 이미지 보안 스캔이 수행되지 않았거나, Critical/High 등급 취약점이 미조치 상태로 배포가 진행된 경우, 또는 배포 파이프라인에 보안 스캔 단계가 미구현된 경우	부적합 사항 시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 전체 이미지 식별이 핵심: AI 서비스는 모델 서빙, 전처리/후처리, 벡터 DB, 캐시, 모니터링 등 다수의 컨테이너로 구성되므로, 일부 이미지만 스캔하고 전체를 완료했다고 주장하는 경우를 주의해야 한다. 반드시 '컨테이너 이미지 목록(Inventory)'과 스캔 보고서의 이미지 태그/다이제스트를 대조하여 누락을 확인한다.

2) 스캔 시점과 배포 이미지의 일치 여부 확인: 스캔은 최종 빌드된 이미지에 대해 수행되어야 한다. 스캔 후 새 레이어가 추가되거나 이미지가 재빌드된 경우, 스캔 결과가 실제 배포 이미지와 불일치할 수 있다. 스캔 보고서의 이미지 Digest와 배포 대상 이미지 Digest의 일치 여부를 반드시 확인한다.

3) AI 특화 취약점에 주목: 일반적인 OS 패키지 취약점 외에, AI 프레임워크(PyTorch의 torch.load 역직렬화 취약점 등), 모델 포맷별 보안 이슈(Pickle 기반 .pt 파일의 임의 코드 실행 등), GPU 드라이버 취약점 등 AI 특유의 보안 위협이 스캔 범위에 포함되었는지 별도로 확인한다.

4) 예외 승인 건의 적정성 검토: Critical/High 취약점 중 패치가 불가능하여 예외 승인(Risk Acceptance)된 건이 있는 경우, 승인 사유의 합리성(패치 미제공, AI 프레임워크 호환성 등)과 보완 조치(WAF 정책 강화, 네트워크 격리 등)의 실효성을 검토한다.

5) 파이프라인 보안 게이트 실효성 확인: CI/CD 파이프라인에 보안 게이트가 있더라도, 수동 우회(Manual Override)가 가능하거나 스캔 실패 시에도 배포가 진행되는 설정이면 실질적 보안 게이트로 기능하지 못한다. 시연이나 파이프라인 설정 파일(Jenkinsfile, GitLab CI YAML 등) 검토를 통해 실효성을 확인한다.

AG_SD45-1-03 | 모델 드리프트 감지 및 재학습 트리거 기준 요건

시스템구조·보안 | 전개 | 운영·모니터링 | 모니터링/재학습

근거 | EU AI Act · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 전개 단계에서, 운영 중 AI 모델의 성능 저하(드리프트)를 탐지하기 위한 기준과 재학습을 실행하는 트리거 조건이 요구사항과 인수산출물에 측정 가능한 수준으로 정의되고 구현되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 드리프트 모니터링 대시보드, 알림 구성 파일, CI/CD 설정과 감사 로그 스토리지 설정을 근거로 지표의 실시간 표출, 임계 초과 시 자동 알림, 재학습 파이프라인 자동화, 감사 로그 보존을 확인한다.

감리원의 점검은 드리프트 탐지 알고리즘의 타당성이 아니라 감지·대응 체계의 구축과 실제 동작 증빙에 한정한다.

2. 필요성

▶ 현행 '성능저하 임계값' 기준은 응답 시간 관점에 한정되어 있어, 정확도·데이터 분포 변화와 같은 AI 모델 품질 드리프트는 별도의 지표와 임계값으로 정의되어야 한다.

▶ 드리프트 감지 없이 운영되는 AI는 시간 경과에 따라 정확도가 자연 저하되며, 그 결과가 행정 오류로 누적되어도 민원이 발생하기 전까지 인지되지 않는다.

▶ EU AI Act 제15조는 고위험 AI에 대해 정확도·강건성·사이버보안 수준의 지속 모니터링을 요구하며, 국내 AI 기본법의 고영향 AI 조항 역시 운영 중 관리 의무를 규정한다.

▶ TTA 신뢰할 수 있는 인공지능 개발안내서는 운영 단계의 지속 모니터링과 재학습 체계 수립을 명시하며, 감지 결과가 재학습으로 연계되는 절차의 문서화를 권고한다.

▶ 대시보드만 구축하고 알림 수신자와 대응 기한이 지정되지 않으면 탐지 체계는 실효성을 갖지 못한다. 감지 이후의 책임과 절차가 함께 정의되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	드리프트 모니터링 대시보드 구축	드리프트 모니터링 대시보드에 정확도·데이터 분포 변화 등 지표 4종 이상이 실시간 표출되는지 시연으로 확인하고 설계 문서와 대조
2	알림(Alert) 자동화 설정	임계 초과 시 담당자에게 자동 알림이 발송되도록 설정되었는지 알림 구성 파일에서 수신자·임계값·채널을 확인 (구성 파일 검토)
3	재학습 파이프라인 자동화	재학습 파이프라인이 자동 실행되도록 구성되었는지 코드 저장소와 CI/CD 설정에서 트리거 조건과 승인 단계를 확인 (코드·CI/CD 검토)
4	감사 로그 자동 보존	드리프트 감지·재학습 이력이 감사 로그로 1년 이상 자동 보존되는지 스토리지의 보존 정책 설정과 실제 적재 이력을 확인 (스토리지 검토)

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	드리프트 모니터링 대시보드(지표 4종 이상 실시간 표출), 임계 초과 시 담당자 자동 알림, 재학습 파이프라인 자동화(1-Click 또는 무인 재학습), 감사 로그 1년 이상 보존이 모두 구현되어 확인된 경우	현행 유지
보완권고 (△)	모니터링 대시보드는 구축되었으나 알림 자동화·재학습 트리거 기준이 일부 미비하거나, 감사 로그 보존 기간이 기준(1년)에 미달하는 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	드리프트 감지 체계가 구축되지 않았거나, 재학습 트리거 기준이 측정 가능한 수준으로 전혀 정의되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) 드리프트는 '무엇을 측정하는가'보다 '누가 언제 보는가'가 더 중요하다: 대시보드만 구축하고 운영 책임자가 지정되지 않으면 실효성이 없다. 감리원은 임계 초과 시 알림 수신자(직위·조직), 확인 기한, 미조치 시 에스컬레이션 경로가 운영 산출물에 명시되었는지 확인한다.

2) 응답시간 임계값과 모델 품질 드리프트를 혼동하지 않는다: 기존 성능저하 임계값은 지연시간 관점에 한정되므로, 정확도·정밀도와 입력 데이터 분포 변화(PSI 등) 같은 품질 지표가 별도로 정의되어야 한다. 감리원은 지표 4종 이상이 수치 임계값과 함께 정의되었는지 확인한다.

3) 재학습 트리거는 자동 실행과 승인 후 실행을 구분한다: 조건 충족 시 무조건 자동 재학습하도록 설계하면 오염·편향 데이터까지 학습할 위험이 있다. 감리원은 트리거 발생 이후 데이터 검증과 승인 절차가 파이프라인에 포함되어 있는지 확인한다.

4) 정답 레이블 지연을 고려한 지표 설계를 확인한다: 운영 중에는 실제 정답이 뒤늦게 확보되어 정확도를 즉시 산출할 수 없는 경우가 많다. 감리원은 레이블 확보 이전에도 사용 가능한 대리 지표(입력 분포 변화, 예측 신뢰도 분포 변화 등)가 정의되었는지 확인한다.

5) 대시보드 시연은 실운영 데이터로 요구한다: 더미 데이터로 구성된 화면은 운영 전환 후 지표가 표출되지 않는 사례가 많다. 감리원은 최근 30일 실데이터의 표출 여부와, 임계 초과 이력 및 그에 따른 조치 기록이 남아 있는지를 확인한다.

AG_SD45-1-04 | AI 서비스 로그 관리 및 프롬프트 감사 추적

시스템구조·보안 | 전개 | 운영·모니터링 | 모니터링/재학습

근거 | EU AI Act · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「개인정보 보호법」 · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 전개 단계에서, 운영 로그가 체계적으로 수집·보관되고 LLM 기반 서비스에서 프롬프트 및 응답 이력을 감사 추적할 수 있는 체계가 구축되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

구체적으로는 로그 스키마 정의서, 실제 적재된 로그, 로그 저장소의 권한 설정과 보존 정책을 근거로 ① 스키마의 완전성, ② 개인정보 마스킹, ③ Trace ID 기반 감사 추적성, ④ 접근 통제와 무결성, ⑤ 보존·파기 정책의 법령 부합을 확인한다.

감리원의 점검은 로그 수집 도구의 기술적 우수성이 아니라 추적 가능성과 통제의 실제 동작에 한정한다.

2. 필요성

▶ 공공 AI 서비스의 의사결정 과정에 대한 감사 요건이 법령과 행정 내부 통제 기준으로 강화되고 있으며, 로그는 사후 검증의 유일한 근거가 된다.

▶ AI 오류나 부적절한 응답이 발생했을 때 원인 분석과 책임 추적을 위해서는 프롬프트·응답·모델 버전이 함께 기록된 상세 로그가 필요하다. 응답만 저장된 로그로는 재현이 불가능하다.

▶ EU AI Act 제12조는 고위험 AI 시스템에 대한 자동 로그 기록 의무를 규정하며, 국내 AI 기본법의 고영향 AI 조항 역시 이에 준하는 기록 의무를 정한다.

▶ 프롬프트 유출이나 인젝션 공격 사고가 발생하면 감사 로그가 사실상 유일한 추적 수단이 된다. 로그를 생성하는 운영자가 삭제 권한까지 보유하면 증거로서의 효력을 잃는다.

▶ 이용자가 프롬프트에 개인정보를 직접 입력하는 경우가 많아, 마스킹이 적용되지 않은 로그는 원문보다 넓은 범위로 개인정보를 축적하는 저장소가 된다.

3. 검토방법

No	점검 항목	확인 방법
1	로그 스키마 완전성	로그 스키마에 요청 Trace ID, 사용자 ID, 모델·프롬프트 버전, RAG 인용 출처가 포함되었는지 스키마 정의서로 확인하고 실제 적재 로그를 샘플 검증

No	점검 항목	확인 방법
2	개인정보 마스킹	프롬프트에 포함된 개인정보가 저장 이전 단계에서 마스킹되는지 샘플 입력을 넣어 적재된 로그의 결과를 확인
3	감사 추적성	임의의 요청 1건을 지정하여 Trace ID로 프롬프트·응답·모델 버전까지 로그에서 추적되는지 시연으로 확인
4	접근 통제·무결성	로그 저장소의 열람·수정·삭제 권한이 운영자와 분리되고 위·변조 방지 조치가 적용되었는지 권한 설정을 확인
5	보존·삭제 정책 법령 부합	로그 보존 기간이 관련 법령·내부 규정과 부합하는지 근거 조항과 대조하고, 만료 시 자동 파기 절차가 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	로그 스키마가 완전하고(요청 Trace ID, 사용자 ID, 모델·프롬프트 버전 메타데이터, RAG 인용 출처 포함), 개인정보 마스킹·접근통제·보존 정책이 법령에 부합하게 구축되어 감사 추적이 시연으로 확인된 경우	현행 유지
보완권고 (△)	로그 수집 체계는 구축되었으나 일부 항목(버전 메타데이터 저장, RAG 인용 기록, 개인정보 마스킹 등)이 누락되었거나 사용자·요청 단위 검색 기능이 미흡한 경우	해당 항목 보완 요청 후 재확인
부적합 (X)	프롬프트·응답 이력이 기록되지 않거나, 감사 추적 체계와 로그 접근통제·보존 정책이 구축되지 않은 경우	시정조치 요구 및 이행 결과 확인

5. 감리 시 유의사항 및 실무 Tip

1) AI 로그는 '저장'보다 '찾기'가 더 어렵다: 로그가 적재되어 있어도 상관관계 키가 없으면 사고 시 추적이 불가능하다. 감리원은 임의의 요청 1건을 지정하여 Trace ID로 프롬프트·응답·모델 버전·RAG 인용 출처까지 1시간 내 복원되는지를 시연으로 확인한다.

2) 프롬프트 로그 자체가 개인정보 저장소가 된다: 이용자가 프롬프트에 주민등록번호·병력 등을 직접 입력하는 사례가 많아, 마스킹이 없으면 로그가 가장 큰 유출 지점이 된다. 감리원은 저장 이전 단계에서 마스킹이 적용되는지를 샘플 입력 테스트로 확인한다.

3) 모델·프롬프트 버전 메타데이터 누락에 주의한다: 동일한 질의라도 모델 버전이 바뀌면 응답이 달라지므로, 버전 정보가 없는 로그는 사후 재현과 책임 규명이 불가능하다. 감리원은 로그 스키마에 모델 ID·모델 버전·프롬프트 템플릿 버전 필드가 존재하는지 확인한다.

4) 로그 무결성과 권한 분리를 함께 확인한다: 로그를 생성하는 운영자가 삭제·수정 권한까지 보유하면 감사 증거로서 효력이 없다. 감리원은 쓰기 전용(WORM) 저장, 해시 체인 등 위·변조 방지 조치와 로그 접근 권한의 분리 현황을 확인한다.

5) 보존기간이 법령과 충돌하지 않는지 교차 확인한다: 감사 목적의 장기 보존과 개인정보 보호법상 최소 보관 원칙은 충돌할 수 있다. 감리원은 보존기간의 근거 법령·내부 규정과 만료 시 자동 파기 절차가 실제로 구현되었는지를 함께 확인한다.

AG_SD45-1-05 | 시스템 리소스 가용성 점검

시스템구조·보안 | 전개 | 운영·모니터링 | 추론인프라

근거 | 「지능정보기술 감리 실무 가이드」(NIA, 2023.2.) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)

1. 목적

본 점검 가이드라인은 AI 서비스 구축 사업의 전개 단계에서 인프라의 리소스 가용성과 부하 대응 체계가 적정 수준으로 확보되었는지를 감리원이 제3자적 관점에서 점검하는 것을 목적으로 한다.

AI 워크로드는 일반 웹 서비스 대비 GPU·CPU 집약도가 현저히 높고, LLM(대규모 언어 모델) 추론 요청은 단일 요청당 수백 밀리초에서 수십 초의 처리 시간을 소모한다.

따라서 인프라 용량의 적정성, 부하 시험 결과의 신뢰성, 오토스케일링 정책의 실효성을 종합적으로 검토하는 것이 필수적이다.

감리원은 'AI 인프라 용량 산정서', {성능테스트결과서}, 오토스케일링 구성 파일, '시스템 구성도'를 근거로 용량 산정의 정량성과 부하 시험의 충분성, 자동 확장·이중화 설계의 실효성을 확인하며, 인프라 최적화 기법 자체의 기술적 타당성 평가는 점검 범위에서 제외한다.

2. 필요성

▶ AI 서비스는 일반 서비스 대비 GPU/CPU 집약적 워크로드로 인프라 리소스 과부하 위험이 높으며, 서비스 중단 시 공공 업무에 직결되는 심각한 장애를 유발할 수 있다.

▶ LLM 추론은 요청당 높은 컴퓨팅 자원을 소모하므로 동시 접속자 증가 시 성능 저하·서비스 중단 위험이 있으며, 이에 대한 사전 검증이 구축 단계에서 반드시 이루어져야 한다.

▶ 정보시스템 감리기준은 전개 단계에서 용량 산정의 적정성 및 부하 시험 결과를 주요 상세 검토 항목으로 명시하고 있으며, AI 서비스의 특수성을 반영한 보완 점검이 요구된다.

▶ 인프라 자원의 과소 산정은 서비스 장애를, 과대 산정은 불필요한 예산 낭비를 초래하므로 정량적 근거에 기반한 용량 계획 수립이 필요하다.

▶ 오토스케일링 정책의 미비는 트래픽 급증 시 자원 배분 실패로 이어지며, 특히 AI 추론 요청의 burst 패턴에 대한 대응 체계가 설계되어야 한다.

3. 검토방법

No	점검 항목	확인 방법
1	AI 워크로드 용량 산정 근거의 정량성	'AI 인프라 용량 산정서'에서 GPU·CPU·메모리 소요량이 동시 요청 수, 최대 토큰 길이, KV Cache 메모리를 반영한 산정식과 함께 제시되었는지 확인하고, 실측 데이터의 출처와 측정 시점을 확인
2	부하 시험 시나리오의 3구간 구성	{성능테스트결과서}에 정상·피크·스파이크 3구간 시나리오가 정의되고, 각 구간의 동시 접속자 수·요청률이 운영 목표치와 함께 기재되었는지 확인
3	스케일 아웃 트리거의 복합 메트릭 구성	오토스케일링 구성 파일에서 GPU 사용률·요청 큐 길이·응답시간을 결합한 복합 메트릭이 트리거로 설정되었는지 확인. 단일 CPU 메트릭만 사용된 경우 부적합으로 판정
4	쿨다운(Cooldown) 시간 정의의 적정성	오토스케일링 정책에 쿨다운 시간이 수치로 정의되고 산정 근거가 제시되었는지 확인하고, 과소 설정에 따른 Flapping 또는 과대 설정에 따른 대응 지연 여부를 {성능테스트결과서}와 대조
5	최소 인스턴스 유지 정책과 Cold Start 방지	최소 유지 인스턴스 수가 0보다 크게 설계되었는지 구성 파일에서 확인하고, Cold Start 발생 시 초기 요청 지연이 {성능테스트결과서}의 응답시간 목표를 초과하지 않는지 확인
6	이중화 및 장애 자동 복구 설계	'시스템 구성도'와 인프라 설계서에서 로드밸런서·추론 서버 이중화와 가용 영역(AZ) 분산 배포가 반영되었는지 확인하고, 헬스체크 실패 시 자동 복구 절차의 시험 결과를 확인
7	리소스 모니터링 및 임계 초과 대응 체계	운영 모니터링 대시보드에서 GPU·CPU·메모리 사용률이 실시간 표출되는지 시연으로 확인하고, 임계 초과 시 알림 수신자와 대응 절차가 '운영 절차서'에 정의되었는지 확인

4. 점검결과 판정기준

판정	판정 기준	주요 사례 / 조치
적합 (O)	용량산정·부하시험·오토스케일링이 정량적 근거와 함께 수행되고, 로드밸런서·추론 서버 이중화와 가용 영역(AZ) 분산 배포 및 헬스체크 실패 시 자동 복구가 설계·시험되며, GPU·CPU·메모리 사용률의 실시간 모니터링과 임계 초과 시 알림 수신자·대응 절차가 정의되어, 결과가 요건을 충족하고 산출물에 구체적으로 명시되어 있는 경우	현행 유지 (사례: 용량산정에 KV Cache 포함, 3구간 부하시험 수행, 오토스케일링 트리거·쿨다운 정량 정의)

판정	판정 기준	주요 사례 / 조치
보완권고 (△)	용량산정은 수행했으나 부하시험 일부 시나리오가 누락되었거나, 오토스케일링 정책이 선언적 수준에 그친 경우	해당 항목 보완 요청 후 재확인 (사례: 피크·스파이크 시험 누락, 쿨다운 시간 미정의, KV Cache 미반영)
부적합 (X)	용량산정 근거가 부재하거나, 부하시험이 수행되지 않았거나, 오토스케일링이 설계되지 않은 경우	시정조치 요구 및 이행결과 확인 (사례: 자원산정 없이 임의 배정, 부하시험 결과서 미제출, 모니터링 체계 부재)

5. 감리 시 유의사항 및 실무 Tip

1) GPU 자원은 과소·과대 산정 모두 문제이다: 과소 산정은 서비스 장애를, 과대 산정은 불필요한 예산 낭비를 초래한다. 감리원은 실측 근거 없이 산정된 경우 부하 시험 결과를 필수 증빙으로 요구하고, 운영 환경과 시험 환경의 사양·구성이 일치하는지를 함께 검증한다.

2) LLM 용량 산정에 KV Cache 메모리가 반영되었는지 확인한다: 모델 가중치 크기만으로 GPU 메모리를 산정하면 동시 요청 증가 시 KV Cache가 메모리를 잠식하여 OOM 장애가 발생한다. 감리원은 용량 산정서에 동시 요청 수와 최대 토큰 길이를 반영한 산정 근거가 있는지 확인한다.

3) 오토스케일링 트리거를 CPU 단일 메트릭으로 설정하지 않았는지 확인한다: GPU 추론 서버는 CPU 사용률이 낮은 상태에서도 GPU 큐가 포화될 수 있다. 감리원은 GPU 사용률·요청 큐 길이·응답시간을 결합한 복합 메트릭 기반 트리거가 설정되었는지 확인하며, 단일 CPU 메트릭만 사용한 경우 부적합으로 판정한다.

4) 쿨다운 시간과 최소 인스턴스 정책은 수치로 정의되어야 한다: 쿨다운이 짧으면 스케일링이 반복되는 Flapping이, 길면 과부하 대응 지연이 발생한다. 또한 0으로 스케일 다운하면 Cold Start로 초기 요청이 타임아웃된다. 감리원은 쿨다운 시간과 최소 유지 인스턴스 수가 근거와 함께 정의되었는지 확인한다.

5) 부하 시험이 정상·피크·스파이크 3구간으로 수행되었는지 확인한다: 평균 부하만 시험하면 AI 추론 요청의 burst 패턴 대응 능력을 검증할 수 없다. 감리원은 최대 부하와 스파이크 시나리오가 포함되었는지, 로드밸런서·추론 서버 이중화와 AZ 분산, 헬스체크 자동 복구 시험 결과가 함께 제시되었는지 확인한다.

APPENDIX

부록

부록

A. 전체 가이드 목록표 · B. 인용 법령·표준 현황 · C. 약어·용어집 · D. 사업특성별 점검항목 조정 매트릭스 · E. 현행 감리 점검항목 연계표 · F. 기술 유형별 커버리지 현황

부록 A. 전체 가이드 목록표 (195건)

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD41-1-01	민감정보 유출차단 DLP 연동 요건	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-02	백업-DR 체계요건	시스템구조·보안 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-1-03	모델성능 정확도 목표 요건	시스템구조·보안 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-1-04	목표AI시스템구성도완전성점검	시스템구조·보안 / 요구정의·분석	시스템 구현	–
AG_SD41-1-05	안전모드 작동 인프라 요건 점검	시스템구조·보안 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-1-06	시스템 도입 설치 환경 분석 요건점검	시스템구조·보안 / 요구정의·분석	시스템 구현	–
AG_SD41-1-07	시스템 용량 산정기준 점검	시스템구조·보안 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-1-08	컨테이너 보안 요건점검	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-09	CI/CD 파이프라인 보안 요건	시스템구조·보안 / 요구정의·분석	시스템 구현	–
AG_SD41-1-10	목적기반LegacyAPI설계점검	시스템구조·보안 / 요구정의·분석	시스템 구현	–
AG_SD41-1-11	DDoS_API_RateLimit 요건점검	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-12	SecurityGuideline모델 공격 방어 요건	시스템구조·보안 / 요구정의·분석	인공지능 모델 개발	보안·프라이버시
AG_SD41-1-13	모델저장소 Registry 보안요건	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-14	비정상 API호출 모델추출 공격요건 점검	시스템구조·보안 / 요구정의·분석	인공지능 모델 개발	보안·프라이버시
AG_SD41-1-15	AI 공급망 보안(AI-BOM) 요건 점검	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-16	데이터 최적화 방어요건 및 적대적 데이터 최적화 공격방어 요건	시스템구조·보안 / 요구정의·분석	데이터 수집·처리	보안·프라이버시
AG_SD41-1-17	API 접근통제 인증요건 점검	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD41-1-18	암호화 전송 요건	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-19	오픈소스 라이브러리 보안 요건	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-20	망분리 환경 모델 훈련 보안요건 점검	시스템구조·보안 / 요구정의·분석	시스템 구현	보안·프라이버시
AG_SD41-1-21	AI신뢰성 테스트	시스템구조·보안 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-1-22	성능저하 임계값 요건 점검	시스템구조·보안 / 요구정의·분석	운영·모니터링	안전성·강건성
AG_SD41-1-23	성능저하 알림 모니터링 인프라 요건	시스템구조·보안 / 요구정의·분석	운영·모니터링	안전성·강건성
AG_SD41-1-24	알림 라우팅체계 점검	시스템구조·보안 / 요구정의·분석	운영·모니터링	안전성·강건성
AG_SD41-2-01	편향식별 기능 요건 점검	응용시스템 / 요구정의·분석	인공지능 모델 개발	공정성·편향
AG_SD41-2-02	AI모델명세 투명성 점검	응용시스템 / 요구정의·분석	인공지능 모델 개발	투명성·설명가능성
AG_SD41-2-03	XAI 설명가능성 점검	응용시스템 / 요구정의·분석	인공지능 모델 개발	투명성·설명가능성
AG_SD41-2-04	신뢰점수 노출 요건	응용시스템 / 요구정의·분석	인공지능 모델 개발	투명성·설명가능성
AG_SD41-2-05	환각 점검 요건	응용시스템 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-2-06	부적절 프롬프트 차단 요건	응용시스템 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-2-07	편견 차별 단어 필터링 요건	응용시스템 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-2-08	금융 컴플라이언스 5대 필수 설명 의무	응용시스템 / 요구정의·분석	인공지능 모델 개발	안전성·강건성
AG_SD41-2-09	윤리적 우려사항 보고 채널	응용시스템 / 요구정의·분석	생명주기 관리	투명성·설명가능성
AG_SD41-2-10	다국어 및 접근성 지원 요건	응용시스템 / 요구정의·분석	시스템 구현	–
AG_SD41-2-11	예외처리 정책 요건	응용시스템 / 요구정의·분석	시스템 구현	–
AG_SD41-2-12	접속 로깅 요건	응용시스템 / 요구정의·분석	시스템 구현	보안·프라이버시

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD41-2-13	시스템 오작동 시 기본값 요건	응용시스템 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-2-14	Fail-safe 전환 요건	응용시스템 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-2-15	UX로그 수집 요건	응용시스템 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-3-01	특성분석 및 선정기준 요건	데이터베이스 / 요구정의·분석	데이터 수집·처리	공정성·편향
AG_SD41-3-02	공통코드 명세요건 및 AI모델 공통분류 코드체계 요구 여부	데이터베이스 / 요구정의·분석	데이터 수집·처리	투명성·설명가능성
AG_SD41-3-03	데이터계보 추적요건 점검	데이터베이스 / 요구정의·분석	데이터 수집·처리	투명성·설명가능성
AG_SD41-3-04	메타데이터 확보 관리 요건	데이터베이스 / 요구정의·분석	데이터 수집·처리	투명성·설명가능성
AG_SD41-3-05	AI 의사결정 감사 추적(Audit Trail) 요건 점검	데이터베이스 / 요구정의·분석	데이터 수집·처리	투명성·설명가능성
AG_SD41-3-06	민감정보 및 개인 식별번호 비식별화 마스킹 및 가명-익명처리요건	데이터베이스 / 요구정의·분석	데이터 수집·처리	보안·프라이버시
AG_SD41-3-07	기존데이터 마이그레이션 및 레거시 데이터 이관 검증 요건	데이터베이스 / 요구정의·분석	데이터 수집·처리	안전성·강건성
AG_SD41-3-08	학습-검증 테스트 분할 비율 및 데이터셋 분할 기준 명시 여부	데이터베이스 / 요구정의·분석	데이터 수집·처리	안전성·강건성
AG_SD41-3-09	AI 학습데이터 품질기준 명시	데이터베이스 / 요구정의·분석	데이터 수집·처리	안전성·강건성
AG_SD41-3-10	벡터 DB 및 RAG 파이프라인 요구정의 적정성 점검	데이터베이스 / 요구정의·분석	데이터 수집·처리	–
AG_SD41-3-11	데이터 분할 계획 및 검증용(Hold-out) 데이터 확보	데이터베이스 / 요구정의·분석	데이터 수집·처리	안전성·강건성
AG_SD41-4-01	편향성 정량지표 점검	품질보증 / 요구정의·분석	인공지능 모델 개발	공정성·편향
AG_SD41-4-02	데이터 버전관리 요건 및 모델버전 종속 데이터 형상관리 요건	품질보증 / 요구정의·분석	데이터 수집·처리	투명성·설명가능성

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD41-4-03	예외대응 요건	품질보증 / 요구정의·분석	시스템 구현	안전성·강건성
AG_SD41-4-04	AI산출물 품질보증 계획	품질보증 / 요구정의·분석	생명주기 관리	안전성·강건성
AG_SD41-4-05	산출물 작성표준	품질보증 / 요구정의·분석	생명주기 관리	–
AG_SD42-1-01	개인정보 비식별화 절차 설계	시스템구조·보안 / 설계	데이터 수집·처리	보안·프라이버시
AG_SD42-1-02	성능지표설계	시스템구조·보안 / 설계	인공지능 모델 개발	안전성·강건성
AG_SD42-1-03	성능 임계값 설계점검	시스템구조·보안 / 설계	운영·모니터링	안전성·강건성
AG_SD42-1-04	안전모드 Fail-safe 트리거 설계 점검	시스템구조·보안 / 설계	시스템 구현	안전성·강건성
AG_SD42-1-05	CI/CD 배포 파이프라인 보안 설계점검	시스템구조·보안 / 설계	시스템 구현	–
AG_SD42-1-06	모델 공격-방어 설계점검	시스템구조·보안 / 설계	인공지능 모델 개발	보안·프라이버시
AG_SD42-1-07	AI서비스 관리자페이지 접근통제 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-08	API 접근 인증-인가 통제설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-09	모델저장소 접근통제점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-10	HSM 연동 키관리 아키텍처 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-11	구간암호화 TLS 인증서 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-12	DDoS 공격방어 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-13	보안약점 예방 아키텍처 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-14	웹셸 악성파일 업로드 차단 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-15	클라우드 인프라 보안 그룹 설계점검	시스템구조·보안 / 설계	시스템 구현	보안·프라이버시
AG_SD42-1-16	테스트환경 구축 계획	시스템구조·보안 / 설계	시스템 구현	안전성·강건성

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD42-1-17	자동화 테스트 설계점검	시스템구조·보안 / 설계	시스템 구현	안전성·강건성
AG_SD42-1-18	SLM 경량 모델 활용 요건	시스템구조·보안 / 설계	인공지능 모델 개발	–
AG_SD42-1-19	모델추론서버 스케일링(Auto-scaling) 점검	시스템구조·보안 / 설계	시스템 구현	–
AG_SD42-2-01	UI 색상배치 편향배제 다크패턴 설계	응용시스템 / 설계	시스템 구현	공정성·편향
AG_SD42-2-02	편향모니터링 지표 산식설계	응용시스템 / 설계	인공지능 모델 개발	공정성·편향
AG_SD42-2-03	편향유발 입력 사전차단 로직설계	응용시스템 / 설계	인공지능 모델 개발	공정성·편향
AG_SD42-2-04	환각발언 예방 파이프라인설계	응용시스템 / 설계	인공지능 모델 개발	공정성·편향
AG_SD42-2-05	AI결정 오버라이드 기능설계	응용시스템 / 설계	인공지능 모델 개발	투명성·설명가능성
AG_SD42-2-06	AI한계점 안내 프론트엔드 설계	응용시스템 / 설계	시스템 구현	투명성·설명가능성
AG_SD42-2-07	XAI 설명가능성 적용 설계	응용시스템 / 설계	인공지능 모델 개발	투명성·설명가능성
AG_SD42-2-08	신뢰점수 노출 설계	응용시스템 / 설계	인공지능 모델 개발	투명성·설명가능성
AG_SD42-2-09	AI결정 오버라이드 수동변경 기능점검	응용시스템 / 설계	인공지능 모델 개발	투명성·설명가능성
AG_SD42-2-10	개인화 데이터처리 알고리즘 설계	응용시스템 / 설계	인공지능 모델 개발	보안·프라이버시
AG_SD42-2-11	안전모드 Fail-safe 로직 설계	응용시스템 / 설계	시스템 구현	안전성·강건성
AG_SD42-2-12	악성 프롬프트 필터링 로직 설계	응용시스템 / 설계	인공지능 모델 개발	보안·프라이버시
AG_SD42-2-13	실시간 성능지표 수집 로직 설계	응용시스템 / 설계	시스템 구현	안전성·강건성
AG_SD42-2-14	타시스템 API연계 프로토콜설계	응용시스템 / 설계	시스템 구현	–
AG_SD42-2-15	AI모델 아키텍처 점검	응용시스템 / 설계	인공지능 모델 개발	–
AG_SD42-2-16	Agentic안전정지Loop 통제점검	응용시스템 / 설계	인공지능 모델 개발	–

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD42-3-01	편향방지 피처엔지니어링 점검	데이터베이스 / 설계	데이터 수집·처리	공정성·편향
AG_SD42-3-02	학습데이터 버전-모델 매핑점검	데이터베이스 / 설계	데이터 수집·처리	투명성·설명가능성
AG_SD42-3-03	비식별화-가명화 뷰 설계점검	데이터베이스 / 설계	데이터 수집·처리	보안·프라이버시
AG_SD42-3-04	데이터 정합성 검증 배치 설계점검	데이터베이스 / 설계	데이터 수집·처리	안전성·강건성
AG_SD42-3-05	개인정보 암호화 컬럼 설계점검	데이터베이스 / 설계	데이터 수집·처리	보안·프라이버시
AG_SD42-3-06	테스트용 데이터관리 점검	데이터베이스 / 설계	데이터 수집·처리	안전성·강건성
AG_SD42-3-07	공공데이터 연계·활용 요건 점검	데이터베이스 / 설계	데이터 수집·처리	공정성·편향
AG_SD42-3-08	DB 성능 최적화 실행계획 설계점검	데이터베이스 / 설계	데이터 수집·처리	–
AG_SD42-3-09	EDB-벡터 DB-하이브리드 데이터모델 정합성	데이터베이스 / 설계	데이터 수집·처리	–
AG_SD42-3-10	성능지표 데이터마트 설계점검	데이터베이스 / 설계	데이터 수집·처리	–
AG_SD42-3-11	정제 전후 적재공간 분리점검	데이터베이스 / 설계	데이터 수집·처리	–
AG_SD43-1-01	적대적 사례 공격방어 점검	시스템구조·보안 / 구현	인공지능 모델 개발	보안·프라이버시
AG_SD43-1-02	모델저장소 비인가 접근차단 점검	시스템구조·보안 / 구현	시스템 구현	보안·프라이버시
AG_SD43-1-03	모델·파일스토리지 암호화 점검	시스템구조·보안 / 구현	시스템 구현	보안·프라이버시
AG_SD43-1-04	시스템 보안약점 예방 필터 계층점검	시스템구조·보안 / 구현	시스템 구현	보안·프라이버시
AG_SD43-1-05	파일업로드 악성코드 방지점검	시스템구조·보안 / 구현	시스템 구현	보안·프라이버시
AG_SD43-2-01	튜토리얼 및 사용자가이드 작동 점검	응용시스템 / 구현	시스템 구현	–
AG_SD43-2-02	마스킹 비식별화 뷰 정상 조회 점검	응용시스템 / 구현	데이터 수집·처리	보안·프라이버시
AG_SD43-2-03	프롬프트 엔지니어링 관리체계 점검	응용시스템 / 구현	인공지능 모델 개발	–

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD43-2-04	캐싱 히트율 및 속도개선	응용시스템 / 구현	시스템 구현	–
AG_SD43-2-05	비정형 문서 처리 AI 요건 점검	응용시스템 / 구현	인공지능 모델 개발	–
AG_SD43-2-06	AI 민원·상담 서비스 특화 요건 점검	응용시스템 / 구현	인공지능 모델 개발	–
AG_SD43-3-01	편견금칙어 사전 적용 점검	데이터베이스 / 구현	데이터 수집·처리	안전성·강건성
AG_SD43-3-02	공통코드 데이터정합성 점검	데이터베이스 / 구현	데이터 수집·처리	투명성·설명가능성
AG_SD43-3-03	데이터계보 이력저장 확인점검	데이터베이스 / 구현	데이터 수집·처리	투명성·설명가능성
AG_SD43-3-04	레거시데이터 이행 ETL 검증 점검	데이터베이스 / 구현	데이터 수집·처리	안전성·강건성
AG_SD43-3-05	이상데이터 실제 제거율점검	데이터베이스 / 구현	데이터 수집·처리	안전성·강건성
AG_SD43-3-06	데이터정합성 Null형식 배치검증 점검	데이터베이스 / 구현	데이터 수집·처리	안전성·강건성
AG_SD43-3-07	외부크롤링 적재 무결성점검	데이터베이스 / 구현	데이터 수집·처리	공정성·편향
AG_SD43-3-08	모델 추론결과 프롬프트 응답 로깅점검	데이터베이스 / 구현	데이터 수집·처리	–
AG_SD43-3-09	정제 전후DB분리 환경 점검	데이터베이스 / 구현	데이터 수집·처리	–
AG_SD43-3-10	데이터 소스변경 모니터링 알림구현 적정성 점검	데이터베이스 / 구현	데이터 수집·처리	안전성·강건성
AG_SD44-1-01	민감정보 유출 차단점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-02	성능 평가지표 달성여부	시스템구조·보안 / 시험	인공지능 모델 개발	안전성·강건성
AG_SD44-1-03	시스템 목표용량 점검	시스템구조·보안 / 시험	시스템 구현	안전성·강건성
AG_SD44-1-04	안전모드 이중화 전환 점검	시스템구조·보안 / 시험	시스템 구현	안전성·강건성
AG_SD44-1-05	API 게이트웨이 인증_인가 통제	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-06	내부 서버간 제로트러스트 인증	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD44-1-07	서비스 관리자페이지 IP 통제점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-08	구간암호화 및 인증서 유효성	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-09	DDoS_WAF 방어시험 점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-10	OS미들웨어 보안패치_CVE취약점 점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-11	망분리_폐쇄환경 모델 훈련점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-12	클라우드 보안설정_방화벽 점검	시스템구조·보안 / 시험	시스템 구현	보안·프라이버시
AG_SD44-1-13	모델 신뢰성 테스트 점검	시스템구조·보안 / 시험	인공지능 모델 개발	안전성·강건성
AG_SD44-1-14	성능테스트 통과여부	시스템구조·보안 / 시험	시스템 구현	안전성·강건성
AG_SD44-1-15	APM 성능 저하 탐지 알림발송	시스템구조·보안 / 시험	운영·모니터링	안전성·강건성
AG_SD44-1-16	모니터링 요건 점검	시스템구조·보안 / 시험	운영·모니터링	안전성·강건성
AG_SD44-1-17	모델 백업버전 복구-롤백테스트	시스템구조·보안 / 시험	운영·모니터링	–
AG_SD44-1-18	서버 리소스 쿼터점검	시스템구조·보안 / 시험	시스템 구현	–
AG_SD44-1-19	운영매뉴얼 목차 설계 점검	시스템구조·보안 / 시험	운영·모니터링	안전성·강건성
AG_SD44-1-20	오픈소스 라이선스 충돌점검	시스템구조·보안 / 시험	시스템 구현	–
AG_SD44-1-21	위험관리대응매뉴얼 점검	시스템구조·보안 / 시험	운영·모니터링	안전성·강건성
AG_SD44-2-01	편향 모니터링 지표 정상 표출점검	응용시스템 / 시험	인공지능 모델 개발	공정성·편향
AG_SD44-2-02	XAI 설명가능성 기능 동작확인 점검	응용시스템 / 시험	인공지능 모델 개발	투명성·설명가능성
AG_SD44-2-03	환각(Hallucination) 억제테스트 통과점검	응용시스템 / 시험	인공지능 모델 개발	안전성·강건성
AG_SD44-2-04	비동기 추론 프로그레스 Bar 상태 동작 점검	응용시스템 / 시험	시스템 구현	–

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD44-2-05	시스템 오작동 기본값 응답점검	응용시스템 / 시험	시스템 구현	안전성·강건성
AG_SD44-2-06	안전모드 Fail-safe 전환 테스트 점검	응용시스템 / 시험	시스템 구현	안전성·강건성
AG_SD44-2-07	실시간 지연시간 테스트 점검	응용시스템 / 시험	시스템 구현	안전성·강건성
AG_SD44-2-08	프롬프트 인젝션 방어 필터 동작점검	응용시스템 / 시험	인공지능 모델 개발	보안·프라이버시
AG_SD44-2-09	데이터 파이프라인 재학습용 동작 점검	응용시스템 / 시험	데이터 수집·처리	안전성·강건성
AG_SD44-2-10	실시간 성능지표 로깅 점검	응용시스템 / 시험	시스템 구현	안전성·강건성
AG_SD44-2-11	외부 연계시스템 통합시험 점검	응용시스템 / 시험	시스템 구현	–
AG_SD44-2-12	API 연계데이터 정합성 점검	응용시스템 / 시험	시스템 구현	–
AG_SD44-3-01	보호변수 배제 분리 보관 확인점검	데이터베이스 / 시험	데이터 수집·처리	공정성·편향
AG_SD44-3-02	학습데이터 버전관리 점검	데이터베이스 / 시험	데이터 수집·처리	투명성·설명가능성
AG_SD44-3-03	개인정보 비식별화 및 파기 점검	데이터베이스 / 시험	데이터 수집·처리	보안·프라이버시
AG_SD44-3-04	데이터 옵트아웃 동의 처리점검	데이터베이스 / 시험	데이터 수집·처리	보안·프라이버시
AG_SD44-3-05	대용량백업 스케줄러 실행점검	데이터베이스 / 시험	데이터 수집·처리	안전성·강건성
AG_SD44-3-06	악성노이즈-중독데이터 필터점검	데이터베이스 / 시험	데이터 수집·처리	보안·프라이버시
AG_SD44-3-07	관리자-일반사용자 DB 권한 통제점검	데이터베이스 / 시험	데이터 수집·처리	보안·프라이버시
AG_SD44-3-08	파일-스토리지 민감컬럼 암호화점검	데이터베이스 / 시험	데이터 수집·처리	보안·프라이버시
AG_SD44-3-09	성능평가 마트 OLAP 집계검증 대시보드 통계 정확성점검	데이터베이스 / 시험	데이터 수집·처리	안전성·강건성
AG_SD44-3-10	샘플링클래스 불균형 해소점검	데이터베이스 / 시험	데이터 수집·처리	공정성·편향
AG_SD44-3-11	데이터 모델ERD물리DB 일치점검	데이터베이스 / 시험	데이터 수집·처리	–

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_SD44-3-12	실시간 스트리밍적재 파이프라인점검	데이터베이스 / 시험	데이터 수집·처리	–
AG_SD44-3-13	인덱스-파티셔닝 효율성 점검	데이터베이스 / 시험	데이터 수집·처리	–
AG_SD44-4-01	편향-차별훈련 이행 점검	품질보증 / 시험	인공지능 모델 개발	공정성·편향
AG_SD44-4-02	AI 신뢰성 테스트·자동화 평가 파이프라인 점검	품질보증 / 시험	인공지능 모델 개발	안전성·강건성
AG_SD45-1-01	CI/CD 배포 파이프라인 자동화 점검	시스템구조·보안 / 전개	운영·모니터링	–
AG_SD45-1-02	컨테이너이미지 보안점검	시스템구조·보안 / 전개	운영·모니터링	보안·프라이버시
AG_SD45-1-03	모델 드리프트 감지 및 재학습 트리거 기준 요건	시스템구조·보안 / 전개	운영·모니터링	안전성·강건성
AG_SD45-1-04	AI 서비스 로그 관리 및 프롬프트 감사 추적	시스템구조·보안 / 전개	운영·모니터링	안전성·강건성
AG_SD45-1-05	시스템 리소스 가용성 점검	시스템구조·보안 / 전개	운영·모니터링	–
AG_PM04-1-01	KPI 점검가이드라인	사업관리 / 착수단계	생명주기 관리	–
AG_PM04-1-02	이해관계자 매핑	사업관리 / 착수단계	생명주기 관리	–
AG_PM04-1-03	AI 제약사항점검	사업관리 / 착수단계	생명주기 관리	안전성·강건성
AG_PM04-1-04	위험관리계획수립	사업관리 / 착수단계	생명주기 관리	안전성·강건성
AG_PM04-1-05	AI 리스크 분류 체계 점검요건	사업관리 / 착수단계	생명주기 관리	안전성·강건성
AG_PM04-1-06	데이터 라벨링 도급관리	사업관리 / 착수단계	데이터 수집·처리	–
AG_PM04-1-07	데이터 라벨링 요구정의	사업관리 / 착수단계	데이터 수집·처리	공정성·편향
AG_PM04-1-08	라벨링데이터 교차검수 요건 및 다수 검수자 검증 프로세스 반영	사업관리 / 착수단계	데이터 수집·처리	공정성·편향

관리번호	가이드명	감리영역 / 단계	생명주기	신뢰성 관점
AG_PM04-1-09	라벨링 편향방지 및 주관적 판단 배제 라벨링 기준 요건	사업관리 / 착수단계	데이터 수집·처리	공정성·편향
AG_PM04-1-10	데이터 수집 편향배제 및 특정 출처 편중방지 수집 요건	사업관리 / 착수단계	데이터 수집·처리	공정성·편향
AG_PM04-2-01	라벨링 교차검수 상태값 설계점검	사업관리 / 실행통제단계	데이터 수집·처리	공정성·편향
AG_PM04-2-02	라벨링 교차검수 상태코드 매핑점검	사업관리 / 실행통제단계	데이터 수집·처리	공정성·편향
AG_PM04-2-03	라벨러 분포 인적 다양성 확보결과 점검	사업관리 / 실행통제단계	데이터 수집·처리	공정성·편향
AG_PM04-2-04	외부 데이터 출처 명시·라이선스 기록 및 저작권 위반 방지	사업관리 / 실행통제단계	데이터 수집·처리	–
AG_PM04-2-05	사용자매뉴얼 목차 설계	사업관리 / 실행통제단계	운영·모니터링	–
AG_PM04-2-06	하도급 보안통제	사업관리 / 실행통제단계	생명주기 관리	보안·프라이버시
AG_PM04-3-01	평가단 XAI 평가결과	사업관리 / 종료단계	인공지능 모델 개발	투명성·설명가능성
AG_PM04-3-02	윤리기준 최종검토	사업관리 / 종료단계	생명주기 관리	투명성·설명가능성
AG_PM04-3-03	하도급업체 보안점검	사업관리 / 종료단계	생명주기 관리	보안·프라이버시
AG_PM04-3-04	이해관계자 인수테스트	사업관리 / 종료단계	생명주기 관리	–

부록 B. 인용 법령·표준 현황

본 부록은 본 가이드의 점검항목이 근거로 삼거나 감리 수행 시 준수 여부를 확인하여야 하는 법령·가이드라인·표준의 현황을 정리한 것이다. 성격이 서로 다른 근거를 하나의 목록으로 제시하면 감리원이 「감리를 하는 근거」와 「감리에서 확인할 대상」을 혼동할 수 있으므로, 다음과 같이 구분하여 제시한다.

B.1 감리 수행의 근거 — 정보시스템 감리 자체의 법적 근거

B.2 AI 관련 법령 — 감리 시 사업의 준수 여부를 확인하여야 할 실체법

B.3 공공 AI 관련 가이드라인·안내서 — 점검 기준으로 활용하는 행정 지침

B.4 국내외 표준·프레임워크 — 점검항목 설계의 참조 기준

B.5 근거–점검항목 대조표 — 근거별로 이를 인용한 점검항목의 관리코드

B.5의 대조표는 법령·지침이 개정되었을 때 갱신하여야 할 점검항목을 특정하는 수단이다. 개별 점검항목의 「근거」 항목과 상호 참조된다.

B.1 감리 수행의 근거

법령·고시	소관	감리 관련 내용
전자정부법	행정안전부	행정기관등의 정보시스템 감리 근거. 감리 대상·감리법인·감리원 자격의 상위 근거
소프트웨어 진흥법	과학기술정보통신부	소프트웨어사업 감리 및 감리법인 등록·관리 근거
정보시스템 감리기준 (행정안전부 고시)	행정안전부	감리 절차, 감리보고서 양식, 점검결과 판정 체계. 본 가이드의 점검결과 판정기준은 이 고시의 체계를 준용한다

※ B.2 이하의 법령·지침은 감리를 수행하는 근거가 아니라, 감리 과정에서 사업의 준수 여부를 확인하여야 할 대상이다. 두 성격을 구분하여 인용한다.

B.2 AI 관련 법령

법령	소관	공포·시행	하위법령	감리 시 확인 요건
인공지능 발전과 신뢰 기반 조성 등에 관한 기본법	과학기술정보통신부	2025.1.21 제정 (2026.1.22 시행) 2026.1.20 개정 (2026.7.21 시행)	시행령 (2026.1.22 시행, 개정 2026.7.21 시행)	제31조 투명성 확보 — 고영향·생성형 AI의 사전 고지 및 생성물 표시 제32조 안전성 확보 의무 제33조 고영향 AI 해당 여부 확인 제34조 고영향 AI 사업자의 책무 제35조 고영향 AI 영향평가 개정으로 도입된 AI 제품·서비스 확인 제도(공공조달 시 우선 고려) 적용 여부
인공지능 및 데이터 기반 행정 활성화에 관한 법률	행정안전부	2026.2.27 개정 (2026.8.28 시행,	시행령·시행규칙	AI 공통기반의 구축·운영 및 활용 여부 학습용 데이터의 관리 체계 행정기관 AI 활용 서비스의 목록 관리

법령	소관	공포·시행	하위법령	감리 시 확인 요건
(구 데이터기반행정 활성화에 관한 법률)		일부 조항 2027. 2.28 시행)		공공기관 AI 영향평가의 사전 실시 및 결과 공표
지능정보화 기본법	과학기술정보통신부	2026.1.22 시행 (인공지능기본법 부칙에 의한 타법개정)	시행령·시행규칙	공공 정보화사업 추진 체계, 지능정보서비스의 접근성 및 이용 보장 요건
개인정보 보호법	개인정보보호위원회	현행 유효	시행령·시행규칙	학습데이터 수집의 적법 근거, 가명·익명처리, 자동화된 결정에 대한 정보주체의 권리, 보유기간 경과 후 파기
저작권법	문화체육관광부	현행 유효	시행령	학습데이터의 저작권 처리 및 라이선스 확보, 외부 수집 데이터의 출처 기록
신용정보의 이용 및 보호에 관한 법률	금융위원회	현행 유효	시행령	금융 분야 AI 서비스의 설명의무 및 자동화평가 대응권 보장
장애인차별금지 및 권리구제 등에 관한 법률	보건복지부	현행 유효	시행령	정보 접근성 보장 — 한국형 웹 콘텐츠 접근성 지침(KWCAG) 준수

B.3 공공 AI 관련 가이드라인·안내서

문서명	발행기관	발간	감리 활용 지점
공공부문 AI 도입·활용 가이드	행정안전부·한국지능정보사회진흥원	2026.6.	기획–예산–계약–구축–운영 5단계 표준절차. 발주기관이 범정부 AI 공통기반(AI 모델·GPU 공동활용)을 검토하였는지, RAG 우선 전략과 환각 저감 설계가 제안요청서·설계산출물에 반영되었는지 점검
인공지능기본법 이행 가이드라인 5종 (투명성 확보 / 안전성 확보 / 고영향 AI 판단 / 고영향 AI 사업자 책무 / AI 영향평가)	과학기술정보통신부	2026.1.	인공지능기본법 제31~35조 의무의 구체적 이행 기준. 고영향 AI 해당 여부의 판단 근거와 영향평가 수행 증빙을 점검할 때의 직접 준거
국가·공공기관 AI보안 가이드북	국가정보원	2025.12.	15개 AI 보안위협과 30개 보안대책, AI 시스템 구축유형별 중점대책. 데이터 수집부터 폐기까지 전 수명주기의 보안통제 적정성 점검
인공지능(AI) 보안 안내서	과학기술정보통신부·한국인터넷진흥원	2025.12.	113개 보안 요구사항과 개발자·서비스제공자·이용자별 체크리스트. 사업자 산출물의 보안요구사항 이행 여부를 항목 단위로 대조
AI 보안 위협 대응 매뉴얼 AI 보안 레드티밍 가이드	과학기술정보통신부·한국인터넷진흥원	2026.7.	프롬프트 인젝션, 권한 오남용, 데이터 유출에 대한 대응 절차. AI 서비스 보안 검증(레드티밍) 수행 여부와 그 결과의 조치 이력 점검
생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서	개인정보보호위원회	2025.8.	목적설정–전략수립–학습/개발–적용·관리 4단계 생애주기별 개인정보 처리 요구사항. 공개된 개인정보 활용의 적법 근거 검토와 배포 전 프라이버시 테스트 수행 여부 점검
AI 프라이버시 리스크 관리 모델	개인정보보호위원회	2024.12.	AI 생애주기별 프라이버시 리스크의 식별·평가·완화 절차. 위험관리계획의 프라이버시 항목 적정성 점검

문서명	발행기관	발간	감리 활용 지점
국가 망 보안체계(N2SF) 보안 가이드라인	국가정보원	2025.	공공 AI 시스템의 망 구성 및 분리 설계 준거. 폐쇄망 환경의 모델 학습·추론 구성 적정성 점검
지능정보기술 감리 실무 가이드	한국지능정보사회진흥원	2023.2.	지능정보기술 사업의 감리 실무 준거. 본 가이드의 점검항목과 현행 감리 점검항목의 연계 기준

B.4 국내외 표준·프레임워크

표준·프레임워크	발행	내용 및 활용
NIST AI RMF 1.0 및 생성형 AI 프로파일	미국 국립표준기술연구소	GOVERN·MAP·MEASURE·MANAGE 4개 함수 기반의 AI 위험관리 체계. 본 가이드 점검항목의 위험 식별·측정 관점 설계 근거
신뢰할 수 있는 인공지능 개발안내서	한국정보통신기술협회	국내 신뢰 AI 요건의 기준. 공정성·투명성·안전성 요구사항의 구체화 근거
ISO/IEC 27001, ISO 19115 등	ISO/IEC	정보보안 경영시스템 및 메타데이터 표준
OWASP LLM Top 10 등	OWASP	LLM 응용의 대표 보안 위협 목록. 프롬프트 인젝션·모델 탈취 등 점검항목의 위협 분류 근거
EU AI Act	유럽연합	2024.8. 발효, 고위험 AI 의무의 단계적 적용. 국제 규제 동향 참조
GDPR	유럽연합	2018.5. 시행. 자동화된 처리에 대한 정보주체 권리 등 참조
한국형 웹 콘텐츠 접근성 지침(KWCAG)	행정안전부·한국정보통신기술협회	웹 접근성 준수 수준 판정 기준
클라우드 서비스 보안인증(CSAP)	과학기술정보통신부·한국인터넷진흥원	공공 클라우드 이용 시 보안인증 획득 여부 확인 기준
암호모듈 검증제도(KCMVP)	국가정보원	암호화 구현 시 검증필 암호모듈 사용 여부 확인 기준

※ 위 법령·가이드라인·표준의 제·개정 현황은 본 가이드 발간 시점을 기준으로 한 것이므로, 감리 수행 시에는 국가법령정보센터 및 해당 발행기관의 최신 공표 내용을 재확인하여 적용한다.

B.5 근거–점검항목 대조표

본문 점검항목이 인용한 근거를 역방향으로 정리한 것이다. 법령·지침이 개정되면 이 표를 통해 갱신 대상 점검항목을 특정한다.

근거	건수	점검항목 관리코드
「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」	50건	AG_PM04-1-03, AG_PM04-1-04, AG_PM04-1-05, AG_PM04-2-05, AG_PM04-3-01, AG_PM04-3-02, AG_SD41-1-05, AG_SD41-1-09, AG_SD41-1-11, AG_SD41-1-16, AG_SD41-2-01, AG_SD41-2-02, AG_SD41-2-03, AG_SD41-2-04, AG_SD41-2-07, AG_SD41-2-08, AG_SD41-2-11, AG_SD41-2-13, AG_SD41-2-14, AG_SD41-3-05, AG_SD41-4-01, AG_SD42-1-03, AG_SD42-1-04, AG_SD42-2-01, AG_SD42-2-02, AG_SD42-2-03, AG_SD42-2-04, AG_SD42-2-05, AG_SD42-2-06, AG_SD42-2-07, AG_SD42-2-08, AG_SD42-2-09, AG_SD42-2-11, AG_SD42-2-16, AG_SD43-1-01, AG_SD43-2-01, AG_SD43-2-06, AG_SD43-3-01, AG_SD44-1-04, AG_SD44-1-21, AG_SD44-2-01, AG_SD44-2-02, AG_SD44-2-05, AG_SD44-2-06, AG_SD44-3-01, AG_SD44-3-06, AG_SD44-4-01, AG_SD44-4-02, AG_SD45-1-03, AG_SD45-1-04

근거	건수	점검항목 관리코드
「인공지능 및 데이터 기반 행정 활성화에 관한 법률」	16건	AG_PM04-1-05, AG_PM04-1-10, AG_PM04-2-02, AG_SD41-1-04, AG_SD41-3-02, AG_SD41-3-03, AG_SD41-3-04, AG_SD41-3-05, AG_SD41-3-07, AG_SD41-3-09, AG_SD41-4-02, AG_SD42-3-02, AG_SD42-3-07, AG_SD43-3-03, AG_SD44-2-09, AG_SD44-3-02
「지능정보화 기본법」	4건	AG_SD41-2-09, AG_SD41-2-10, AG_SD42-2-06, AG_SD43-2-05
「개인정보 보호법」	38건	AG_PM04-1-03, AG_PM04-1-04, AG_PM04-1-06, AG_PM04-2-06, AG_PM04-3-03, AG_SD41-1-01, AG_SD41-1-09, AG_SD41-1-10, AG_SD41-1-16, AG_SD41-1-18, AG_SD41-1-20, AG_SD41-2-01, AG_SD41-2-12, AG_SD41-2-15, AG_SD41-3-06, AG_SD42-1-01, AG_SD42-1-11, AG_SD42-2-07, AG_SD42-2-10, AG_SD42-2-13, AG_SD42-3-03, AG_SD42-3-05, AG_SD42-3-06, AG_SD42-3-11, AG_SD43-1-03, AG_SD43-2-02, AG_SD43-2-05, AG_SD43-2-06, AG_SD43-3-09, AG_SD44-1-01, AG_SD44-1-08, AG_SD44-2-10, AG_SD44-3-01, AG_SD44-3-03, AG_SD44-3-04, AG_SD44-3-07, AG_SD44-3-08, AG_SD45-1-04
「저작권법」	6건	AG_PM04-2-04, AG_SD41-1-15, AG_SD42-1-18, AG_SD42-3-07, AG_SD43-3-07, AG_SD44-1-20
「신용정보의 이용 및 보호에 관한 법률」	2건	AG_SD41-1-10, AG_SD41-2-08
「장애인차별금지 및 권리구제 등에 관한 법률」	6건	AG_PM04-1-02, AG_SD41-2-07, AG_SD41-2-10, AG_SD43-2-05, AG_SD43-2-06, AG_SD44-3-01
「전자정부법」	15건	AG_PM04-1-03, AG_SD41-1-02, AG_SD41-2-12, AG_SD41-3-02, AG_SD41-3-07, AG_SD42-1-07, AG_SD42-1-13, AG_SD42-1-14, AG_SD42-2-14, AG_SD42-3-08, AG_SD44-1-11, AG_SD44-2-10, AG_SD44-2-12, AG_SD44-3-11, AG_SD44-3-13
「공공부문 AI 도입·활용 가이드」(행정안전부·NIA, 2026.6.)	31건	AG_PM04-1-01, AG_PM04-2-05, AG_SD41-1-03, AG_SD41-1-04, AG_SD41-1-05, AG_SD41-1-06, AG_SD41-1-07, AG_SD41-1-22, AG_SD41-2-05, AG_SD41-3-05, AG_SD41-3-10, AG_SD41-4-05, AG_SD42-1-18, AG_SD42-1-19, AG_SD42-2-04, AG_SD42-2-14, AG_SD42-2-15, AG_SD42-3-07, AG_SD42-3-09, AG_SD43-2-01, AG_SD43-2-04, AG_SD43-2-06, AG_SD44-1-03, AG_SD44-1-15, AG_SD44-1-16, AG_SD44-1-18, AG_SD44-1-19, AG_SD44-2-03, AG_SD44-2-07, AG_SD44-2-11, AG_SD45-1-05
「인공지능기본법 이행 가이드라인」(과기정통부, 2026.1.)	4건	AG_PM04-3-02, AG_SD41-2-02, AG_SD43-2-01, AG_SD44-1-21
「국가·공공기관 AI보안 가이드북」(국가정보원, 2025.12.)	43건	AG_PM04-1-06, AG_PM04-2-06, AG_PM04-3-03, AG_SD41-1-02, AG_SD41-1-08, AG_SD41-1-10, AG_SD41-1-11, AG_SD41-1-13, AG_SD41-1-14, AG_SD41-1-16, AG_SD41-1-17, AG_SD41-1-19, AG_SD41-1-20, AG_SD41-2-06, AG_SD41-2-12, AG_SD41-3-10, AG_SD42-1-05, AG_SD42-1-06, AG_SD42-1-07, AG_SD42-1-08, AG_SD42-1-09, AG_SD42-1-10, AG_SD42-1-12, AG_SD42-1-14, AG_SD42-1-15, AG_SD42-2-16, AG_SD42-3-05, AG_SD43-1-01, AG_SD43-1-02, AG_SD43-1-03, AG_SD43-1-04, AG_SD43-1-05, AG_SD44-1-06, AG_SD44-1-07, AG_SD44-1-08, AG_SD44-1-09, AG_SD44-1-10, AG_SD44-1-11, AG_SD44-1-12, AG_SD44-3-06, AG_SD44-3-07, AG_SD45-1-01, AG_SD45-1-02
「인공지능(AI) 보안 안내서」(과기정통부·KISA, 2025.12.)	37건	AG_PM04-2-06, AG_SD41-1-08, AG_SD41-1-09, AG_SD41-1-11, AG_SD41-1-13, AG_SD41-1-15, AG_SD41-1-16, AG_SD41-1-17, AG_SD41-1-19, AG_SD41-1-24, AG_SD41-2-06, AG_SD41-2-11, AG_SD42-1-05, AG_SD42-1-06, AG_SD42-1-08, AG_SD42-1-09, AG_SD42-1-10, AG_SD42-1-11, AG_SD42-1-12, AG_SD42-1-13, AG_SD42-1-14, AG_SD42-2-12, AG_SD43-1-02, AG_SD43-1-04, AG_SD43-1-05, AG_SD44-1-05, AG_SD44-1-06, AG_SD44-1-07, AG_SD44-1-08, AG_SD44-1-09, AG_SD44-1-10, AG_SD44-1-12, AG_SD44-1-18, AG_SD44-2-05, AG_SD44-3-06, AG_SD45-1-01, AG_SD45-1-02
「AI 보안 레드티밍 가이드」(과기정통부·KISA, 2026.7.)	16건	AG_SD41-1-10, AG_SD41-1-12, AG_SD41-1-21, AG_SD41-2-06, AG_SD42-1-06, AG_SD42-1-13, AG_SD42-1-17, AG_SD42-2-12, AG_SD42-2-16, AG_SD

근거	건수	점검항목 관리코드
		43-1-01, AG_SD43-1-04, AG_SD43-2-03, AG_SD44-1-05, AG_SD44-1-07, AG_SD44-2-08, AG_SD44-2-11
「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회, 2025.8.)	13건	AG_SD41-1-01, AG_SD41-2-12, AG_SD41-2-15, AG_SD41-3-06, AG_SD41-3-10, AG_SD42-1-01, AG_SD42-2-10, AG_SD42-3-03, AG_SD44-1-01, AG_SD44-2-10, AG_SD44-3-03, AG_SD44-3-04, AG_SD45-1-04
「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회, 2024.12.)	7건	AG_PM04-1-04, AG_SD42-1-01, AG_SD42-2-10, AG_SD42-3-03, AG_SD43-2-02, AG_SD44-1-01, AG_SD44-3-03
「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원, 2025.)	6건	AG_SD41-1-20, AG_SD42-1-15, AG_SD42-1-18, AG_SD44-1-06, AG_SD44-1-11, AG_SD44-1-12
「지능정보기술 감리 실무 가이드」(NIA, 2023.2.)	53건	AG_PM04-1-01, AG_PM04-1-03, AG_PM04-3-04, AG_SD41-1-04, AG_SD41-1-06, AG_SD41-1-07, AG_SD41-1-22, AG_SD41-1-23, AG_SD41-3-02, AG_SD41-3-07, AG_SD41-4-04, AG_SD41-4-05, AG_SD42-1-02, AG_SD42-1-16, AG_SD42-1-17, AG_SD42-1-19, AG_SD42-2-09, AG_SD42-2-14, AG_SD42-3-04, AG_SD42-3-06, AG_SD42-3-08, AG_SD43-2-04, AG_SD43-3-02, AG_SD43-3-04, AG_SD43-3-05, AG_SD43-3-06, AG_SD43-3-08, AG_SD44-1-02, AG_SD44-1-03, AG_SD44-1-13, AG_SD44-1-14, AG_SD44-1-20, AG_SD44-2-01, AG_SD44-2-02, AG_SD44-2-03, AG_SD44-2-04, AG_SD44-2-05, AG_SD44-2-06, AG_SD44-2-07, AG_SD44-2-08, AG_SD44-2-09, AG_SD44-2-10, AG_SD44-2-11, AG_SD44-2-12, AG_SD44-3-03, AG_SD44-3-05, AG_SD44-3-09, AG_SD44-3-10, AG_SD44-3-11, AG_SD44-3-12, AG_SD44-3-13, AG_SD45-1-01, AG_SD45-1-05
NIST AI RMF	96건	AG_PM04-1-01, AG_PM04-1-02, AG_PM04-1-04, AG_PM04-1-06, AG_PM04-1-07, AG_PM04-1-08, AG_PM04-1-09, AG_PM04-1-10, AG_PM04-2-01, AG_PM04-2-02, AG_PM04-2-03, AG_PM04-2-04, AG_PM04-3-01, AG_PM04-3-04, AG_SD41-1-01, AG_SD41-1-03, AG_SD41-1-09, AG_SD41-1-12, AG_SD41-1-14, AG_SD41-1-15, AG_SD41-1-19, AG_SD41-1-21, AG_SD41-1-23, AG_SD41-1-24, AG_SD41-2-02, AG_SD41-2-03, AG_SD41-2-04, AG_SD41-2-05, AG_SD41-2-09, AG_SD41-2-13, AG_SD41-2-14, AG_SD41-3-01, AG_SD41-3-03, AG_SD41-3-04, AG_SD41-3-06, AG_SD41-3-08, AG_SD41-3-09, AG_SD41-3-11, AG_SD41-4-01, AG_SD41-4-02, AG_SD41-4-03, AG_SD41-4-04, AG_SD42-1-02, AG_SD42-1-03, AG_SD42-1-04, AG_SD42-1-05, AG_SD42-1-07, AG_SD42-1-09, AG_SD42-1-16, AG_SD42-1-17, AG_SD42-1-19, AG_SD42-2-02, AG_SD42-2-03, AG_SD42-2-05, AG_SD42-2-06, AG_SD42-2-07, AG_SD42-2-08, AG_SD42-2-09, AG_SD42-2-11, AG_SD42-2-12, AG_SD42-2-13, AG_SD42-2-15, AG_SD42-2-16, AG_SD42-3-01, AG_SD42-3-09, AG_SD42-3-10, AG_SD43-1-01, AG_SD43-2-02, AG_SD43-3-01, AG_SD43-3-03, AG_SD43-3-07, AG_SD43-3-08, AG_SD43-3-10, AG_SD44-1-02, AG_SD44-1-03, AG_SD44-1-04, AG_SD44-1-11, AG_SD44-1-13, AG_SD44-1-14, AG_SD44-1-15, AG_SD44-1-16, AG_SD44-1-17, AG_SD44-1-19, AG_SD44-2-02, AG_SD44-2-09, AG_SD44-3-02, AG_SD44-3-04, AG_SD44-3-05, AG_SD44-3-06, AG_SD44-3-07, AG_SD44-3-08, AG_SD44-3-09, AG_SD44-3-10, AG_SD44-3-12, AG_SD44-4-01, AG_SD44-4-02
「신뢰할 수 있는 인공지능 개발안내서」(TTA)	103건	AG_PM04-1-02, AG_PM04-1-07, AG_PM04-1-08, AG_PM04-1-09, AG_PM04-1-10, AG_PM04-2-01, AG_PM04-2-02, AG_PM04-2-03, AG_PM04-2-05, AG_PM04-3-01, AG_PM04-3-02, AG_PM04-3-04, AG_SD41-1-01, AG_SD41-1-03, AG_SD41-1-05, AG_SD41-1-12, AG_SD41-1-14, AG_SD41-1-21, AG_SD41-1-22, AG_SD41-1-23, AG_SD41-1-24, AG_SD41-2-01, AG_SD41-2-02, AG_SD41-2-03, AG_SD41-2-04, AG_SD41-2-05, AG_SD41-2-07, AG_SD41-2-09, AG_SD41-2-11, AG_SD41-2-13, AG_SD41-2-14, AG_SD41-2-15, AG_SD41-3-01, AG_SD41-3-03, AG_SD41-3-04, AG_SD41-3-06, AG_SD41-3-08, AG_SD41-3-09, AG_SD41-3-10, AG_SD41-3-11, AG_SD41-4-01, AG_SD41-4-02, AG_SD41-4-03, AG_SD41-4-04, AG_SD41-4-05, AG_SD42-1-02, AG_SD42-1-03, AG_SD42-1-04, AG_SD42-1-05, AG_SD42-1-12, AG_SD42-1-16, AG_SD42-1-18, AG_SD42-2-01, AG_SD42-2-02, AG_SD42-2-03, AG_SD42-2-04, AG_S

근거	건수	점검항목 관리코드
		D42-2-05, AG_SD42-2-08, AG_SD42-2-09, AG_SD42-2-11, AG_SD42-2-13, AG_SD42-2-15, AG_SD42-3-01, AG_SD42-3-02, AG_SD42-3-04, AG_SD42-3-06, AG_SD42-3-09, AG_SD42-3-10, AG_SD42-3-11, AG_SD43-2-02, AG_SD43-2-03, AG_SD43-3-01, AG_SD43-3-02, AG_SD43-3-03, AG_SD43-3-04, AG_SD43-3-05, AG_SD43-3-06, AG_SD43-3-07, AG_SD43-3-08, AG_SD43-3-09, AG_SD43-3-10, AG_SD44-1-02, AG_SD44-1-04, AG_SD44-1-13, AG_SD44-1-14, AG_SD44-1-15, AG_SD44-1-16, AG_SD44-1-17, AG_SD44-1-19, AG_SD44-1-21, AG_SD44-2-01, AG_SD44-2-02, AG_SD44-2-03, AG_SD44-2-05, AG_SD44-2-06, AG_SD44-2-08, AG_SD44-3-02, AG_SD44-3-04, AG_SD44-3-09, AG_SD44-3-10, AG_SD44-4-01, AG_SD44-4-02, AG_SD45-1-03
ISO/IEC 27001	22건	AG_PM04-2-06, AG_PM04-3-03, AG_SD41-1-02, AG_SD41-1-08, AG_SD41-1-13, AG_SD42-1-07, AG_SD42-1-08, AG_SD42-1-09, AG_SD42-1-10, AG_SD42-1-11, AG_SD42-1-15, AG_SD43-1-02, AG_SD43-1-03, AG_SD44-1-05, AG_SD44-1-06, AG_SD44-1-07, AG_SD44-1-10, AG_SD44-1-17, AG_SD44-3-05, AG_SD44-3-07, AG_SD44-3-08, AG_SD45-1-02
ISO 19115	1건	AG_SD41-3-04
OWASP LLM Top 10	24건	AG_SD41-1-08, AG_SD41-1-12, AG_SD41-1-17, AG_SD41-2-06, AG_SD41-2-11, AG_SD41-4-03, AG_SD42-1-06, AG_SD42-1-08, AG_SD42-1-13, AG_SD42-1-14, AG_SD42-1-17, AG_SD42-2-12, AG_SD43-1-02, AG_SD43-1-04, AG_SD43-1-05, AG_SD43-2-03, AG_SD44-1-01, AG_SD44-1-05, AG_SD44-1-09, AG_SD44-1-10, AG_SD44-1-14, AG_SD44-1-18, AG_SD44-2-08, AG_SD44-2-11
EU AI Act	9건	AG_PM04-1-05, AG_PM04-2-04, AG_SD41-1-15, AG_SD41-3-05, AG_SD42-2-01, AG_SD44-3-01, AG_SD44-4-02, AG_SD45-1-03, AG_SD45-1-04
GDPR	1건	AG_SD42-2-07
KWCAG	3건	AG_SD41-2-10, AG_SD43-2-05, AG_SD44-2-04
CSAP	3건	AG_SD41-1-18, AG_SD42-1-15, AG_SD44-1-12
KCMVP	7건	AG_SD41-1-18, AG_SD42-1-10, AG_SD42-1-11, AG_SD42-3-05, AG_SD43-1-03, AG_SD44-1-08, AG_SD44-3-08

부록 C. 약어·용어집

본 부록은 본문에서 사용한 주요 약어와 용어를 정리한 것이다. 약어표는 영문 원어와 국문 의미를, 용어표는 개념 정의를 함께 제시한다.

1. 약어표

약어	영문 표기	국문 의미
AES	Advanced Encryption Standard	고급 암호화 표준
AI	Artificial Intelligence	인공지능
AI-BOM	AI Bill of Materials	AI 공급망 구성목록
API	Application Programming Interface	응용프로그램 인터페이스
APM	Application Performance Monitoring	애플리케이션 성능 모니터링
AS-IS/TO-BE	As-Is / To-Be	현행/목표 (업무 분석)
AWS	Amazon Web Services	아마존 웹 서비스(클라우드)
BLEU/ROUGE	Bilingual Evaluation Understudy / Recall-Oriented Understudy for Gisting Evaluation	생성물 평가 지표
BMT	Benchmark Test	벤치마크 시험
CA	Certificate Authority	인증기관
CI/CD	Continuous Integration / Continuous Delivery	지속통합·지속배포
CPU	Central Processing Unit	중앙처리장치
CSAP	Cloud Security Assurance Program	클라우드 서비스 보안 인증제
CSP	Content Security Policy	콘텐츠 보안정책
CUDA	Compute Unified Device Architecture	엔비디아 병렬연산 플랫폼
CVE	Common Vulnerabilities and Exposures	공개 취약점 목록
DAST	Dynamic Application Security Testing	동적 애플리케이션 보안테스트
DB	Database	데이터베이스
DBMS	Database Management System	데이터베이스 관리 시스템
DDL	Data Definition Language	데이터 정의어
DDoS	Distributed Denial of Service	분산 서비스거부 공격
DES	Data Encryption Standard	(구형) 데이터 암호화 표준
DLP	Data Loss Prevention	데이터 유출방지
DML	Data Manipulation Language	데이터 조작어
DR	Disaster Recovery	재해복구
DVC	Data Version Control	데이터 버전관리
E2E	End-to-End	종단간(전 구간)
EDB	Enterprise/Relational Database	정형(관계형) 데이터베이스
ELK	Elasticsearch·Logstash·Kibana	로그 수집·분석 스택

약어	영문 표기	국문 의미
EOL	End of Life	지원 종료(수명 만료)
ERD	Entity-Relationship Diagram	개체관계도
ETL	Extract-Transform-Load	추출·변환·적재
F1	F1 Score	F1 점수(정밀도·재현율 조화평균)
FK	Foreign Key	외래키
FPR	False Positive Rate	오탐률(위양성률)
GCP	Google Cloud Platform	구글 클라우드 플랫폼
GPL/AGPL	(Affero) GNU General Public License	(Affero) GNU 일반 공중 라이선스
GPU	Graphics Processing Unit	그래픽 처리장치
gRPC	gRPC Remote Procedure Call	고성능 원격 프로시저 호출 프레임워크
HITL	Human-in-the-Loop	인간 개입(형)
HSM	Hardware Security Module	하드웨어 보안모듈
HSTS	HTTP Strict Transport Security	HTTPS 강제 적용 웹 보안 정책
HTTP	HyperText Transfer Protocol	하이퍼텍스트 전송 프로토콜
HTTPS	HyperText Transfer Protocol Secure	보안 HTTP
IAA	Inter-Annotator Agreement	라벨러 간 일치도
IP	Internet Protocol (Address)	인터넷 프로토콜(주소)
IQR	Interquartile Range	사분위 범위
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission	국제표준화기구·국제전기기술위원회
JSON	JavaScript Object Notation	자바스크립트 객체 표기법
JWT	JSON Web Token	JSON 웹 토큰
KCMVP	Korea Cryptographic Module Validation Program	국내 암호모듈 검증 제도
KMS	Key Management Service	키 관리 서비스
KOGL	Korea Open Government License	공공누리(공공저작물 자유이용)
KPI	Key Performance Indicator	핵심성과지표
KWCAG	Korean Web Content Accessibility Guidelines	한국형 웹 접근성 지침
L7	Layer 7 (Application Layer)	(OSI) 7계층·응용계층
LIME	Local Interpretable Model-agnostic Explanations	국소 대리모델 설명기법
LLM	Large Language Model	대규모 언어모델
LLM-as-Judge	LLM as a Judge	LLM 기반 자동 평가
LoRA/QLoRA	(Quantized) Low-Rank Adaptation	저랭크 적응(경량 파인튜닝)
MFA	Multi-Factor Authentication	다중요소 인증
MIME	Multipurpose Internet Mail Extensions	파일 형식 식별 표준
ML	Machine Learning	기계학습(머신러닝)

약어	영문 표기	국문 의미
MLflow	Machine Learning flow	ML 실험·모델 관리 도구
MLOps	Machine Learning Operations	머신러닝 운영
MRR	Mean Reciprocal Rank	평균 역순위(검색 품질 평가 지표)
mTLS	mutual Transport Layer Security	상호 TLS 인증
NAT	Network Address Translation	네트워크 주소 변환
NCCL	NVIDIA Collective Communications Library	엔비디아 GPU 집합통신 라이브러리
NDA	Non-Disclosure Agreement	비밀유지계약
nDCG	Normalized Discounted Cumulative Gain	정규화 순위이득(추천 지표)
NER	Named Entity Recognition	개체명 인식
NIST	National Institute of Standards and Technology	미국 국립표준기술연구소
NPU	Neural Processing Unit	신경망 처리장치
OAuth	Open Authorization	개방형 인가 표준
OCR	Optical Character Recognition	광학문자인식
OECD	Organisation for Economic Co-operation and Development	경제협력개발기구
OLAP	Online Analytical Processing	온라인 분석처리
OOM	Out Of Memory	메모리 부족
OS	Operating System	운영체제
OTP	One-Time Password	일회용 비밀번호
OWASP	Open Worldwide Application Security Project	국제 웹 보안 커뮤니티/표준
P95/P99	95th / 99th Percentile	95·99 백분위(지연 지표)
PII	Personally Identifiable Information	개인식별정보
PK	Primary Key	기본키
QA	Quality Assurance	품질보증
RAG	Retrieval-Augmented Generation	검색증강생성
RBAC	Role-Based Access Control	역할기반 접근통제
RDB	Relational Database	관계형 데이터베이스
RDBMS	Relational Database Management System	관계형 데이터베이스 관리 시스템
RFP	Request for Proposal	제안요청서
RLHF	Reinforcement Learning from Human Feedback	인간 피드백 기반 강화학습
RMF	Risk Management Framework	위험관리 프레임워크
ROI	Return on Investment	투자수익(률)
RPO	Recovery Point Objective	목표복구시점
RTM	Requirements Traceability Matrix	요구사항 추적 매트릭스
RTO	Recovery Time Objective	목표복구시간
SaaS	Software as a Service	서비스형 소프트웨어

약어	영문 표기	국문 의미
SAST	Static Application Security Testing	정적 애플리케이션 보안테스트
SBOM	Software Bill of Materials	소프트웨어 구성목록
SCA	Software Composition Analysis	소프트웨어 구성분석
SHA	Secure Hash Algorithm	안전 해시 알고리즘
SHAP	SHapley Additive exPlanations	샤플리 기반 설명기법
SLA	Service Level Agreement	서비스수준협약
SLM	Small Language Model	경량(소형) 언어모델
SMOTE	Synthetic Minority Over-sampling Technique	소수 클래스 합성 오버샘플링 기법
SQL	Structured Query Language	구조화 질의어
SW	Software	소프트웨어
TC	Test Case	테스트 케이스
TCO	Total Cost of Ownership	총소유비용
TEVV	Test, Evaluation, Verification and Validation	시험·평가·검증·확인
TGI	Text Generation Inference	텍스트 생성 추론 서버
TLS	Transport Layer Security	전송계층 보안
TPS	Transactions Per Second	초당 처리건수
TTA	Telecommunications Technology Association	한국정보통신기술협회
TTFT	Time To First Token	첫 토큰 응답 지연
TTL	Time To Live	캐시 유효시간
TTS/STT	Text-to-Speech / Speech-to-Text	음성합성·음성인식
UAT	User Acceptance Test	사용자 인수시험
UI	User Interface	사용자 인터페이스
UI/UX	User Interface / User Experience	사용자 인터페이스·경험
URL	Uniform Resource Locator	자원 위치 지정자
UX	User Experience	사용자 경험
vLLM	vLLM (high-throughput LLM inference)	고처리량 LLM 추론 엔진
VPC	Virtual Private Cloud	가상 사설 클라우드
VPN	Virtual Private Network	가상 사설망
VRAM	Video Random Access Memory	그래픽 메모리
WAF	Web Application Firewall	웹 방화벽
WAS	Web Application Server	웹 애플리케이션 서버
WBS	Work Breakdown Structure	작업분해구조
XAI	eXplainable AI	설명가능한 인공지능
XSS	Cross-Site Scripting	교차 사이트 스크립팅

※ 본 약어표는 약어의 알파벳(A–Z) 순으로 정렬하였다.

2. 용어표

「근거」에는 해당 용어를 규정하고 있는 법령·지침·표준을 기재하였다. 규범 문서에서 정의를 찾을 수 없는 용어는 「기술 관용어」로 표기하였으며, 이는 실무에서 통용되는 표현일 뿐 규범적 정의가 있는 것은 아니라는 뜻이다.

용어	영문	정의	근거
생성형 인공지능	Generative AI	텍스트·이미지·코드 등 새로운 콘텐츠를 생성하는 인공지능. 대규모 데이터로 사전학습한 모델이 입력(프롬프트)에 따라 결과를 만들어내며, 규칙 기반 시스템과 달리 동일 입력에도 결과가 달라질 수 있다.	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조
파운데이션 모델	Foundation Model	대규모 데이터로 사전학습되어 다양한 하위 과제에 적용·미세조정할 수 있는 범용 기반 모델. LLM이 대표적이며, 자체 구축 대신 상용 API로 활용하는 경우가 많다.	기술 관용어
대규모 언어 모델	Large Language Model, LLM	방대한 텍스트로 학습해 자연어를 이해·생성하는 초거대 모델. 문맥에 따라 다음 토큰을 확률적으로 예측하는 방식으로 동작한다.	기술 관용어
경량 언어모델	Small Language Model, SLM	파라미터 규모가 작아 적은 자원으로 온프레미스·엣지 환경에서 운영할 수 있는 언어모델. 도메인 파인튜닝과 결합해 비용·보안 이점을 얻는다.	기술 관용어
검색증강생성	Retrieval-Augmented Generation, RAG	외부 지식베이스에서 관련 문서를 검색해 그 근거를 바탕으로 답변을 생성하는 기법. 모델 내부 지식만 사용하는 방식보다 최신성·정확성을 높이고 환각을 억제한다.	「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)
파인튜닝	Fine-tuning	사전학습 모델을 특정 과제·도메인 데이터로 추가 학습해 성능을 맞추는 과정. 전체 가중치를 갱신하는 방식과 LoRA·QLoRA처럼 일부만 조정하는 경량 방식이 있다.	기술 관용어
프롬프트 엔지니어링	Prompt Engineering	원하는 출력을 얻기 위해 지시문·예시·맥락을 설계·구성하는 기법. 시스템 프롬프트 구조, Few-shot 예시, 컨텍스트 관리 등을 포함한다.	기술 관용어

용어	영문	정의	근거
환각(할루시네이션)	Hallucination	AI가 사실과 다르거나 근거 없는 내용을 그럴듯하게 생성하는 현상. 생성형 AI의 본질적 한계로, 근거 제시·출력 검증·RAG 등으로 완화한다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」
프롬프트 인젝션	Prompt Injection	악의적 입력으로 모델의 지시·정책을 우회·탈취·조작하는 공격. 시스템 프롬프트 노출이나 금지 기능 우회를 유발할 수 있어 입력 필터링·권한 통제가 필요하다.	OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA)
탈옥	Jailbreak	안전장치·정책을 무력화해 금지된 응답을 유도하는 시도. 역할극·우회 표현 등으로 이뤄지며 프롬프트 인젝션의 한 유형이다.	OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA)
가드레일	Guardrail	입력·출력 단계에서 부적절·위험 콘텐츠를 탐지·차단·수정하는 안전 통제 장치. 금칙어 필터, 정책 검증, 출력 후처리 등을 포함한다.	「인공지능(AI) 보안 안내서」(과기정통부·KISA)
설명가능한 인공지능	eXplainable AI, XAI	모델의 예측 근거와 도출 과정을 사람이 이해할 수 있게 설명을 제공하는 기술. SHAP·LIME·Attention 시각화 등이 있으며 고위험 도메인에서 신뢰성 확보에 활용된다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · NIST AI RMF
편향	Bias	학습 데이터·설계의 치우침이 특정 집단·조건에 불리한 결과로 나타나는 현상. 공정성 지표 측정, 금칙어 필터링, 데이터 다양성 확보 등으로 관리한다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · NIST AI RMF MEASURE
드리프트	Drift	운영 중 입력 데이터 분포(데이터 드리프트)나 예측 성능(개념 드리프트)이 시간에 따라 변화·저하하는 현상. 지속 모니터링과 재학습 트리거로 대응한다.	NIST AI RMF MANAGE
신뢰도(신뢰 점수)	Confidence Score	모델이 예측 결과에 대해 산출하는 확신 정도. 임계값 미만이면 안전모드 전환·인간 검토 등 후속 처리의 근거로 활용한다.	기술 관용어

용어	영문	정의	근거
안전모드	Fail-safe	오류·이상·저신뢰 상황에서 시스템을 안전한 사전 정의 상태로 전환하는 설계. 대체 응답, 서비스 보류, 관리자 전환(HITL) 등을 포함한다.	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · 「신뢰할 수 있는 인공지능 개발 안내서」(TTA)
인간 개입	Human-in-the-Loop, HITL	자동화된 AI 판단에 사람이 검토·승인·개입하도록 하는 설계. 고위험 결정의 오류와 책임 문제를 완화한다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · NIST AI RMF
회피 공격	Evasion Attack	입력에 미세한 교란을 가해 모델이 오분류하도록 유도하는 적대적 공격. 적대적 학습·입력 정제 등으로 방어한다.	「국가·공공기관 AI보안 가이드북」(국가정보원) · OWASP LLM Top 10
데이터 중독	Data Poisoning	학습 데이터에 악의적 데이터를 주입해 모델의 성능·판단을 훼손하는 공격. 데이터 무결성 검증과 이상 탐지로 방어한다.	「국가·공공기관 AI보안 가이드북」(국가정보원) · OWASP LLM Top 10
모델 추출	Model Extraction	반복 질의로 모델의 기능·파라미터를 복제·탈취하는 공격. 호출 제한과 출력 정보 최소화로 완화한다.	「국가·공공기관 AI보안 가이드북」(국가정보원) · OWASP LLM Top 10
멤버십 추론	Membership Inference	특정 데이터가 모델 학습에 사용됐는지 추론해 개인정보를 유추하는 공격. 차분 프라이버시·출력 제한으로 대응한다.	「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회)
임베딩	Embedding	텍스트·이미지 등을 의미를 보존한 고차원 벡터로 변환한 표현. 유사도 검색과 RAG의 기반이 된다.	기술 관용어
벡터 데이터베이스	Vector Database	임베딩 벡터를 저장하고 의미 기반 유사도 검색을 제공하는 저장소. RAG의 지식베이스로 활용된다.	기술 관용어
평가셋(골든셋)	Evaluation Set / Golden Set	모델 성능·품질을 측정하기 위해 정답이 확정된 검증 데이터. 학습 데이터와 분리·격리해 누출(Test Leakage)을 방지해야 한다.	기술 관용어
재현성	Reproducibility	동일 조건에서 동일 결과를 재현할 수 있는 성질. 데이터·코드·모델 버전과 파라미터·시드 기록으로 확보한다.	NIST AI RMF MEASURE
데이터 계보	Data Lineage	데이터의 출처·수집·가공·이동 이력을 추적할 수 있는 관리 체계. 품질·저작권·감사 대응의 근거가 된다.	「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · ISO 19115

용어	영문	정의	근거
형상관리	Configuration Management	데이터·모델·코드의 버전과 변경 이력을 통제·추적하는 활동. 재현성과 롤백을 뒷받침한다.	「지능정보기술 감리 실무 가이드」(NIA)
모델 저장소	Model Registry	학습된 모델을 버전·상태(스테이징·운영)와 함께 등록·관리하는 저장소. 접근통제와 승인 워크플로우가 필요하다.	기술 관용어
비식별화	De-identification	개인을 식별할 수 없도록 가명·익명 처리하는 기법. 가명처리, 총계처리, 마스킹, 범주화 등이 있다.	「개인정보 보호법」·「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」
양자화	Quantization	모델 가중치를 낮은 비트로 표현해 크기·연산량을 줄이는 경량화 기법. 추론 속도와 메모리 효율을 높인다.	기술 관용어
오토스케일링	Auto-scaling	부하에 따라 컴퓨팅 자원(인스턴스)을 자동으로 증감하는 기능. AI 추론은 GPU 사용률·요청 큐 깊이 등 특화 메트릭 기반 설계가 필요하다.	기술 관용어
MLOps·LLMOps	ML/LLM Operations	AI 모델의 개발·배포·모니터링·재학습을 자동화·표준화하는 운영 체계. 파이프라인·형상관리·모니터링을 포함한다.	기술 관용어
온도	Temperature	생성 결과의 무작위성(다양성)을 조절하는 파라미터. 값이 낮을수록 결정적, 높을수록 창의적이지만 불안정한 출력이 된다.	기술 관용어
토큰	Token	모델이 처리하는 텍스트의 최소 단위(단어·부분단어 등). 입력·출력 길이와 비용 산정의 기준이 된다.	기술 관용어
컨텍스트 윈도우	Context Window	모델이 한 번에 처리할 수 있는 최대 토큰 범위. 이를 초과한 입력은 잘리거나 요약이 필요하다.	기술 관용어
데이터 다이오드	Data Diode	물리적으로 한 방향 전송만 허용하는 장비. 망분리 폐쇄 환경에서 데이터 반입 통제에 사용된다.	「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원)
재학습	Retraining	성능 저하·드리프트·데이터 갱신에 대응해 모델을 다시 학습·갱신하는 과정. 주기·트리거·검증 절차를 정의해야 한다.	NIST AI RMF MANAGE·「신뢰할 수 있는 인공지능 개발안내서」(TTA)

용어	영문	정의	근거
공정성	Fairness	AI 시스템이 특정 집단·속성에 대해 부당한 차별 없이 일관된 결과를 제공하는 성질. 편향 측정·완화와 함께 신뢰할 수 있는 AI의 핵심 요건으로 다뤄진다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조
투명성	Transparency	AI의 동작 방식·학습 데이터·판단 근거를 이해관계자가 알 수 있도록 공개·설명하는 성질. 모델 명세 공개, 설명가능성(XAI), 로깅·감사추적 등으로 확보한다.	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조 · 「신뢰할 수 있는 인공지능 개발 안내서」(TTA)
강건성	Robustness	입력 잡음·적대적 공격·예외 상황에서도 AI가 성능을 안정적으로 유지하는 성질. 적대적 학습, 입력 검증, 안전모드 설계 등으로 강화한다.	「신뢰할 수 있는 인공지능 개발 안내서」(TTA) · NIST AI RMF
적대적 사례	Adversarial Example	사람은 정상으로 인식하지만 AI 모델은 오분류하도록 미세하게 조작된 입력. 회피 공격의 수단이 되며 적대적 학습으로 방어한다.	「국가·공공기관 AI보안 가이드북」(국가정보원) · OWASP LLM Top 10
앙상블	Ensemble	여러 모델의 예측을 결합해 단일 모델보다 정확도·안정성을 높이는 기법. 적대적 방어와 성능 향상에 활용된다.	기술 관용어
에이전트 AI	Agentic AI	목표를 스스로 계획하고 도구 호출·다단계 실행으로 수행하는 자율형 AI. 자율성이 높은 만큼 안전 정지·권한 통제 등 추가 안전장치가 요구된다.	기술 관용어
피처 엔지니어링	Feature Engineering	원천 데이터에서 모델 학습에 유효한 특성(변수)을 선택·가공·생성하는 과정. 편향 유입 방지와 모델 성능에 큰 영향을 미친다.	기술 관용어
과적합	Overfitting	모델이 학습 데이터에 지나치게 맞춰져 새로운 데이터에서 성능이 떨어지는 현상. 검증셋 분리, 정규화, 데이터 다양성 확보로 완화한다.	기술 관용어
벤치마크	Benchmark	모델·시스템의 성능을 표준화된 데이터·과제로 측정·비교하는 기준. 목표 성능 설정과 상대 평가에 활용된다.	기술 관용어

용어	영문	정의	근거
제로트러스트	Zero Trust	'아무것도 신뢰하지 않고 항상 검증한다'는 보안 원칙. 내부·외부를 구분하지 않고 모든 접근을 인증·인가·최소권한으로 통제한다.	「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원)
망분리	Network Separation	업무망과 외부(인터넷)망을 물리적·논리적으로 분리해 데이터 유출·침입을 차단하는 보안 방식. AI 모델 학습 등 민감 구간을 폐쇄망에 배치할 때 적용한다.	「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원)
차분 프라이버시	Differential Privacy	데이터에 통계적 잡음을 추가해 개별 정보 노출을 막으면서 전체 분석은 가능하게 하는 프라이버시 보호 기법. 멤버십 추론 등 개인정보 유추 공격 방어에 쓰인다.	「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회)
오탐·미탐	False Positive / False Negative	정상을 위협으로 잘못 차단(오탐)하거나 실제 위협을 놓치는(미탐) 판정 오류. 필터링·이상탐지 설계에서 두 오류의 균형(임계값)이 중요하다.	기술 관용어
레드팀	Red Teaming	공격자 관점에서 AI의 취약점·유해 출력·우회 가능성을 의도적으로 시험하는 활동. 탈옥·프롬프트 인젝션 등에 대한 방어력을 점검한다.	「AI 보안 레드티밍 가이드」(과기정통부·KISA)
카나리·블루그린 배포	Canary / Blue-Green Deployment	신규 버전을 일부 트래픽 또는 별도 환경에 먼저 배포해 위험을 줄이는 무중단 배포 전략. 문제 발생 시 즉시 이전 버전으로 롤백할 수 있다.	기술 관용어
롤백	Rollback	오류·성능 저하 발생 시 이전의 안정된 버전(모델·코드·데이터)으로 되돌리는 복구 조치. 형상관리·버전관리가 전제되어야 한다.	기술 관용어
온프레미스	On-premise	클라우드가 아닌 자체 보유 인프라에 시스템·모델을 구축·운영하는 방식. 보안·규제 준수나 민감 데이터 통제가 필요할 때 채택한다.	기술 관용어
메타데이터	Metadata	데이터의 출처·구조·의미·이력 등을 설명하는 '데이터에 대한 데이터'. 데이터 계보·품질관리·검색의 기반이 된다.	ISO 19115

3. 카테고리 정의표

본 절은 각 점검 가이드의 속성 띠 네 번째 항목으로 표기한 「카테고리」를 정의한 것이다. 카테고리는 감리영역·단계와는 별개로 유사한 점검 관점을 묶는 분류 축이며, 감리원이 특정 관점의 점검항목을 한 번에 찾을 때 사용한다.

「개념 근거」에는 해당 카테고리의 개념을 규정하고 있는 법령·지침·표준을 기재하였다. 규범 문서에서 정의를 찾을 수 없는 기술 관용어는 그 사실을 그대로 표기하였으며, 공식 정의가 확립되면 이를 갱신한다.

구분	카테고리	정의	개념 근거	항목 수
사업·거버넌스	**사업타당성/ROI**	AI 도입의 목표 지표와 투자 대비 효과를 착수 전에 정량화하였는지를 보는 관점	「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)	1
	의사소통	이해관계자를 식별하고 그 요구를 정의 단계에서 수렴하는 체계	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP	1
	조직체계	AI 사업 수행 조직의 구성과 역할·책임의 정의	NIST AI RMF GOVERN	1
	AI 위험관리	AI 고유 위험의 식별·분류와 대응 계획 수립	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · NIST AI RMF MANAGE	3
	법령/AI윤리	관련 법령의 준수 여부와 윤리기준의 이행 여부	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조 · 「인공지능기본법 이행 가이드라인」(과기정통부)	2
	컴플라이언스/가드레일	금융 등 분야별 규제가 요구하는 설명의무와 통제 장치	「신용정보의 이용 및 보호에 관한 법률」 · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조	1
	저작권/라이선스	학습·활용 데이터의 권리 처리와 출처·라이선스 기록	「저작권법」 · EU AI Act	2
	형상관리	산출물·데이터·모델 버전의 형상 통제와 이력 유지	「지능정보기술 감리 실무 가이드」(NIA)	1
	교육훈련기술이전	이용자·운영자에 대한 안내와 운영 기술의 이전	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조	1
	시범운영/인수테스트	인수 기준의 충족 여부를 시범운영과 인수시험으로 확인	「지능정보기술 감리 실무 가이드」(NIA)	1
	안정화/하자보수	개통 이후 안정화 기간의 결함 대응 체계	「지능정보기술 감리 실무 가이드」(NIA)	2
데이터	**학습데이터**	학습에 사용할 데이터의 확보 범위와 구성 요건	「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)	3
	라벨링	라벨링 기준의 명확성, 교차검수 절차, 작업자 구성의 다양성	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MAP	6
	데이터 품질	학습·검증 데이터의 품질 기준 정의와 검증 수행	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 학습용 데이터 품질관리 가이드라인」(NIA)	3
	정제(클렌징)/변환	전처리 절차의 적정성과 이상·중복 데이터의 제거	「지능정보기술 감리 실무 가이드」(NIA)	4

구분	카테고리	정의	개념 근거	항목 수
	거버넌스/데이터 계보	데이터의 출처·변경 이력 추적과 의사결정 감사 추적	「인공지능 및 데이터 기반 행정 활성화에 관한 법률」 · ISO 19115	9
	비식별화	개인식별정보의 가명·익명 처리와 마스킹	「개인정보 보호법」 · 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」(개인정보보호위원회)	7
	개인정보보호	수집의 적법 근거, 보유기간 경과 후 파기, 정보주체의 권리 보장	「개인정보 보호법」 · 「AI 프라이버시 리스크 관리 모델」(개인정보보호위원회)	3
	데이터 백업/복구	데이터 백업 주기와 복구 절차의 검증	ISO/IEC 27001 · 「전자정부법」	2
	데이터드리프트	운영 중 데이터 분포의 변화를 감지하고 대응하는 체계	NIST AI RMF MEASURE·MANAGE	1
	벡터 DB/임베딩	임베딩 저장소의 구성과 기존 데이터베이스와의 정합성	기술 관용어 — 규범적 정의 없음	1
모델·품질	**LLM모델**	대규모 언어모델의 선정 기준과 아키텍처 적정성	「신뢰할 수 있는 인공지능 개발안내서」(TTA)	1
	모델경량화	경량 언어모델(SLM) 등 축소 모델의 활용 요건	기술 관용어 — 규범적 정의 없음	1
	모델평가	성능 지표의 정의와 목표 달성 여부의 검증	NIST AI RMF MEASURE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)	5
	모델관리	모델 버전의 저장·백업과 이전 버전으로의 복구	NIST AI RMF MANAGE · ISO/IEC 27001	1
	학습추론 및 저장	학습·추론 과정에서 처리되는 데이터와 그 저장·보관 방식	「지능정보기술 감리 실무 가이드」(NIA) · 「개인정보 보호법」	9
	편향/공정성	학습데이터와 모델 출력의 편향을 식별·측정하고 완화하는 활동	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제34조	12
	할루시네이션	근거 없는 생성 결과를 억제하고 검증하는 장치	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)	2
	AI윤리/xAI	판단 근거의 설명 제공과 모델 정보의 투명한 공개	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조·제34조 · NIST AI RMF	10
	금칙어/가드레일	부적절한 입력과 출력을 사전에 차단하는 필터 체계	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조	3
	프롬프트설계	프롬프트의 설계·관리 체계와 악성 프롬프트에 대한 방어	OWASP LLM Top 10 · 「AI 보안 레드티밍 가이드」(과기정통부·KISA)	1
	RAG파이프라인	검색증강생성 구성의 적정성과 응답 근거의 추적	기술 관용어 — 규범적 정의 없음	1
	멀티모달/에이전트	텍스트 외 형식의 처리와 모델의 자율적 작업 수행	기술 관용어 — 규범적 정의 없음	1
	오케스트레이션	다단계 실행 흐름의 구성과 중단·통제 장치	기술 관용어 — 규범적 정의 없음	1
	AI QA체계	AI 산출물에 대한 품질보증 계획과 신뢰성 시험 체계	「신뢰할 수 있는 인공지능 개발안내서」(TTA) · NIST AI RMF MEASURE	6
	모니터링/재학습	운영 중 성능 저하의 감지와 재학습 착수 기준	NIST AI RMF MANAGE · 「신뢰할 수 있는 인공지능 개발안내서」(TTA)	11

구분	카테고리	정의	개념 근거	항목 수
	테스트시나리오	시험 시나리오의 설계와 결함 처리·재시험 절차	「지능정보기술 감리 실무 가이드」(NIA) · NIST AI RMF MEASURE	5
응용·서비스	**UI/UX, 접근성**	화면 구성의 적정성과 취약계층을 포함한 정보 접근성 보장	「장애인차별금지 및 권리구제 등에 관한 법률」 · KWCAG	4
	챗봇	대화형 서비스에 요구되는 응답 품질과 안내 요건	기술 관용어 — 규범적 정의 없음	1
	외부연계	타 시스템과의 연계 구조와 통합시험의 수행	「지능정보기술 감리 실무 가이드」(NIA)	2
	AI모델 API	모델 호출 API의 연계 프로토콜과 데이터 정합성	「지능정보기술 감리 실무 가이드」(NIA) · 「전자정부법」	1
	성능	응답시간·처리량 등 서비스 수준의 목표 설정과 달성 확인	「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)	2
인프라·운영	**아키텍처**	목표 시스템 구성도의 완전성과 구성요소 간 정합성	「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)	1
	추론인프라	추론 서버의 구성과 부하에 따른 확장 방식	「공공부문 AI 도입·활용 가이드」(행정안전부·NIA)	2
	GPU/NPU연산	연산 자원의 배분과 사용량 한도 관리	기술 관용어 — 규범적 정의 없음	1
	서버스토리지/네트워크	서버·저장장치·회선 등 기반 자원의 용량과 구성	「지능정보기술 감리 실무 가이드」(NIA)	1
	클라우드컨테이너	클라우드 및 컨테이너 환경의 구성과 보안 설정	CSAP · 「인공지능(AI) 보안 안내서」(과기정통부·KISA)	1
	가용성/이중화	장애 시의 전환 방식과 이중화 구성	「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제32조 · ISO/IEC 27001	8
	CI/CD파이프라인	빌드·배포 자동화 파이프라인의 구성과 보안	「인공지능(AI) 보안 안내서」(과기정통부·KISA) · NIST AI RMF GOVERN	3
	MCP/Tool연계	모델이 외부 도구·시스템을 호출하는 연계 구조	기술 관용어 — 규범적 정의 없음	1
보안	**AI특화보안**	적대적 입력, 모델 추출, 학습데이터 중독 등 AI 고유 위협에 대한 방어	「국가·공공기관 AI보안 가이드북」(국가정보원) · OWASP LLM Top 10	10
	인증/인가/접근통제	이용자·관리자·시스템 간 접근 권한의 통제	「국가·공공기관 AI보안 가이드북」(국가정보원) · ISO/IEC 27001	9
	API보안	API 호출의 인증과 유량 제어	「인공지능(AI) 보안 안내서」(과기정통부·KISA) · OWASP LLM Top 10	1
	시큐어코딩	보안약점의 사전 예방과 코드 수준의 검증	「인공지능(AI) 보안 안내서」(과기정통부·KISA) · 「전자정부법」	9
	암호화	전송 구간과 저장 데이터의 암호화 적용	「개인정보 보호법」 · KCMVP	7
	망분리/폐쇄망	학습·추론 환경의 망 구성과 분리 수준	「국가 망 보안체계(N2SF) 보안 가이드라인」(국가정보원)	4
	보안관리	보안 정책의 수립과 점검 이행 여부	「국가·공공기관 AI보안 가이드북」(국가정보원) · ISO/IEC 27001	2

※「항목 수」는 해당 카테고리로 분류된 점검 가이드의 건수이다. 항목 수가 적은 카테고리는 그 관점의 점검이 아직 제한적이라는 의미이므로, 부록 F의 커버리지 현황과 함께 확인한다.

부록 D. 사업특성별 점검항목 조정 매트릭스

본 부록은 현행 감리 점검항목과 본 가이드의 점검항목을 하나의 점검표로 통합하는 절차와 서식을 제시한다. 개별 사업의 특성에 따라 어떤 점검항목을 그대로 두고, 무엇을 보완하며, 무엇을 더할지를 감리계획 수립 단계에서 명시적으로 판정하도록 하는 것이 목적이다.

본 가이드의 점검항목은 현행 점검항목을 대체하지 않는다. 따라서 두 점검표를 단순히 합치면 중복 점검과 점검 누락이 동시에 발생한다. 조정 매트릭스는 이를 방지하기 위한 판정 도구이며, 작성 결과는 감리계획서에 첨부하여 점검 범위에 대한 발주기관·사업자와의 합의 근거로 사용한다.

D.1 조정 절차

▼ 점검항목 조정 절차

STEP 1 현행 점검표와 본 가이드를 나란히 놓는다

현행 감리 점검항목	본 가이드 점검항목 195건

↓

STEP 2 사업 특성을 판별한다

AI 적용 범위 (전면 / 부분)	AI 기술유형 (LLM·RAG·에이전트)	데이터 민감도	서비스 위험도	개발방법론

↓

STEP 3 감리영역별로 네 가지 구분에 따라 판정한다

유지	현행 점검항목을 그대로 적용
수정	현행 점검항목에 AI 특성을 반영하여 보완
삭제	해당 사업에 적용되지 않아 제외
AI 추가	본 가이드 195건에서 선별하여 추가

사업관리·품질보증 / 응용시스템 / 데이터베이스 / 시스템구조·보안 — 4개 영역 각각 판정

↓

STEP 4. 해당 사업의 최종 점검표 확정

감리계획서에 첨부하고, 각 판정의 사유를 함께 기록한다

절차의 핵심은 STEP 2와 STEP 3이다. 사업 특성을 판별하지 않고 매트릭스를 작성하면 판정이 형식적으로 흐르며, 판정 근거를 남기지 않으면 검사 단계에서 점검 범위를 두고 다툼이 생긴다. 감리원은 각 판정의 사유를 한 줄이라도 기록하여야 한다.

D.2 조정 매트릭스 서식

감리영역별로 아래 네 가지 구분에 따라 점검항목을 판정한다. 「대상 점검항목」에는 현행 점검항목의 번호 또는 본 가이드의 관리코드를 기재하고, 「판정 사유」에는 그렇게 판정한 근거를 간략히 적는다.

감리영역	구분	조정 내용	대상 점검항목	판정 사유
사업관리 및 품질 보증	유지	현행 점검항목을 그대로 적용		
	수정	현행 점검항목에 AI 특성을 반영하여 보완		
	삭제	해당 사업에 적용되지 않아 제외		
	AI 추가	본 가이드에서 선별하여 추가		
응용시스템	유지	현행 점검항목을 그대로 적용		
	수정	현행 점검항목에 AI 특성을 반영하여 보완		
	삭제	해당 사업에 적용되지 않아 제외		
	AI 추가	본 가이드에서 선별하여 추가		
데이터베이스	유지	현행 점검항목을 그대로 적용		
	수정	현행 점검항목에 AI 특성을 반영하여 보완		
	삭제	해당 사업에 적용되지 않아 제외		
	AI 추가	본 가이드에서 선별하여 추가		
시스템구조·보안	유지	현행 점검항목을 그대로 적용		
	수정	현행 점검항목에 AI 특성을 반영하여 보완		
	삭제	해당 사업에 적용되지 않아 제외		
	AI 추가	본 가이드에서 선별하여 추가		

※ 「삭제」는 현행 점검항목을 감리 대상에서 제외한다는 뜻이 아니라, 해당 사업의 특성상 점검 실익이 없어 이번 감리에서 다루지 않는다는 의미이다. 제외 사유는 반드시 기재한다.

D.3 작성 예시

대표적인 두 가지 사업 유형에 대하여 매트릭스를 기입한 예시를 제시한다. 아래 예시는 판정의 방식을 보이기 위한 것이므로, 실제 사업에서는 요구사항과 산출물 범위를 확인한 뒤 다시 판정하여야 한다.

예시 1. 대민 민원상담 챗봇 (RAG 기반, 개인정보 처리, 대외 공개)

사업 특성: AI가 서비스 전면에 적용되며 일반 국민이 직접 사용한다. 민원인의 개인정보가 프롬프트와 응답 경로에 유입될 수 있고, 잘못된 답변이 곧바로 민원으로 이어지므로 서비스 위험도가 높다. 자체 학습데이터 구축과 라벨링을 포함한다.

감리영역	구분	대상 점검항목	판정 사유
사업관리 및 품질보증	유지	일정·위험·이슈·변경·품질·산출물·요구사항 관리 등 현행 사업관리 점검항목 전체	AI 사업에도 전통적 사업관리 통제는 동일하게 필요
	수정	현행 요구사항관리 점검 + AG_PM04-1-01 KPI 점검가이드라인	검수 분쟁 방지를 위해 AI 성능 목표의 정량화 여부를 함께 확인
	삭제	해당 없음	—
	AI 추가	AG_PM04-1-03 AI 제약사항점검 AG_PM04-1-05 AI 리스크 분류 체계 점검요건 AG_PM04-2-05 사용자매뉴얼 목차 설계 AG_PM04-3-02 윤리기준 최종검토 AG_SD41-4-01 편향성 정량지표 점검	대민 서비스이므로 리스크 분류·윤리 검토·이용자 안내가 필수
응용시스템	유지	화면 설계·기능 명세·사용성 관련 현행 점검항목	기존 웹 서비스 구조를 그대로 사용
	수정	현행 예외처리 점검 + AG_SD41-2-11 예외처리 정책 요건, AG_SD41-2-13 시스템 오작동 시 기본값 요건	AI 오작동 시의 기본 응답과 대체 경로를 예외처리 범위에 포함
	삭제	해당 없음	—
	AI 추가	AG_SD41-2-05 환각 점검 요건 AG_SD42-2-04 환각발언 예방 파이프라인설계 AG_SD44-2-03 환각(Hallucination) 억제테스트 통과점검 AG_SD41-2-06 부적절 프롬프트 차단 요건 AG_SD44-2-08 프롬프트 인젝션 방어필터 동작점검 AG_SD43-2-06 AI 민원·상담 서비스 특화 요건 점검	환각과 부적절 입력이 곧바로 민원으로 이어지는 서비스 특성
데이터베이스	유지	데이터 표준·정합성·백업 관련 현행 점검항목	관계형 DB 구조는 기존과 동일
	수정	현행 암호화 설계 점검 + AG_SD42-3-05 개인정보 암호화 컬럼 설계점검	프롬프트·응답 로그에 개인정보가 포함될 수 있어 암호화 대상에 추가
	삭제	AG_SD41-3-07 기존데이터 마이그레이션 및 레거시 데이터 이관 검증 요건 AG_SD43-3-04 레거시데이터 이행 ETL 검증 점검	신규 구축 사업으로 이관 대상 레거시 데이터가 없음

감리영역	구분	대상 점검항목	판정 사유
	AI 추가	AG_SD41-3-10 벡터 DB 및 RAG 파이프라인 요구정의 적정성 점검 AG_SD41-3-06 민감정보 및 개인 식별번호 비식별화 마스킹 및 가명-익명처리요건 AG_SD44-3-03 개인정보 비식별화 및 파기 점검	RAG 파이프라인과 개인정보 비식별화가 서비스의 핵심 위험
시스템구조·보안	유지	인프라 구성·용량·백업·DR 관련 현행 점검항목	기존 인프라 위에 추론 서버를 추가하는 구조
	수정	현행 접근통제 점검 + AG_SD41-1-17 API 접근통제 인증요건 점검	추론 API 인증·인가를 접근통제 범위에 포함
	삭제	해당 없음	—
	AI 추가	AG_SD41-1-12 SecurityGuideline모델 공격 방어 요건 AG_SD44-1-13 모델 신뢰성 테스트 점검 AG_SD45-1-03 모델 드리프트 감지 및 재학습 트리거 기준 요건	외부 노출 서비스이므로 모델 공격 방어와 드리프트 관리가 필요

예시 2. 내부 문서검색 서비스 (RAG 기반, 직원 전용, 대외 비공개)

사업 특성: 기관 내부 직원만 사용하며 대외에 공개되지 않는다. 자체 모델 학습 없이 상용 모델과 사내 문서 임베딩을 결합한 검색 서비스로, 라벨링 작업이 없다. 다만 사내 비공개 문서를 다루므로 망 구성과 유출 통제의 중요도가 높다.

감리영역	구분	대상 점검항목	판정 사유
사업관리 및 품질보증	유지	현행 사업관리 점검항목 전체	사업 규모와 무관하게 동일하게 적용
	수정	해당 없음	—
	삭제	AG_PM04-1-06 데이터 라벨링 도급관리 AG_PM04-1-07 데이터 라벨링 요구정의 AG_PM04-1-08 라벨링데이터 교차검수 요건 및 다수 검수자 검증 프로세스 반영 AG_PM04-1-09 라벨링 편향방지 및 주관적 판단 배제 라벨링 기준 요건 AG_SD44-4-01 편향-차별훈련 이행 점검	자체 학습과 라벨링이 없어 라벨링 통제 항목의 점검 실익이 없음
	AI 추가	AG_PM04-1-05 AI 리스크 분류 체계 점검요건	내부 서비스라도 AI 위험 수준의 사전 판정은 필요
응용시스템	유지	검색 화면·권한 관련 현행 점검항목	기존 내부 포털 구조를 그대로 사용
	수정	해당 없음	—
	삭제	해당 없음	—
	AI 추가	AG_SD41-2-05 환각 점검 요건 AG_SD44-2-03 환각(Hallucination) 억제테스트 통과점검 AG_SD42-2-06 AI한계점 안내 프론트엔드 설계 AG_SD43-2-05 비정형 문서 처리 AI 요건 점검	근거 문서 없는 답변과 비정형 문서 처리 정확도가 주된 위험
데이터베이스	유지	데이터 표준·백업 관련 현행 점검항목	기존 문서관리 DB를 연계
	수정	해당 없음	—
	삭제	AG_SD41-3-06 민감정보 및 개인 식별번호 비식별화 마스킹 및 가명-익명처리요건	학습데이터를 별도로 구축하지 않아 비식별화 대상이 없음
	AI 추가	AG_SD41-3-10 벡터 DB 및 RAG 파이프라인 요구정의 적정성 점검 AG_SD42-3-09 EDB-벡터 DB-하이브리드 데이터모델 정합성	벡터 DB와 기존 DB의 정합성이 검색 품질을 좌우
시스템구조·보안	유지	인프라·백업·보안 관련 현행 점검항목	내부망 기존 인프라 활용
	수정	현행 자료 유출통제 점검 + AG_SD41-1-01 민감정보 유출차단 DLP 연동 요건	사내 문서가 외부 모델 API로 전송되지 않는지 확인 범위에 추가
	삭제	해당 없음	—
	AI 추가	AG_SD41-1-20 망분리 환경 모델 훈련 보안요건 점검 AG_SD41-1-15 AI 공급망 보안(AI-BOM) 요건 점검 AG_SD44-1-01 민감정보 유출 차단점검	비공개 문서를 다루므로 망 구성과 공급망 보안이 핵심

※ 두 예시를 비교하면, 같은 RAG 기반 서비스라도 대외 공개 여부와 자체 학습 유무에 따라 「AI 추가」와 「삭제」의 대상이 달라진다. 매트릭스 작성의 실익은 이 차이를 감리계획 단계에서 명시적으로 드러내는 데 있다.

부록 E. 현행 감리 점검항목 연계표 (97건)

본 부록은 현행 정보시스템 감리에서 사용하는 표준 점검항목이 AI 서비스 구축 사업에서 어떻게 적용되는지, 그리고 본 가이드의 어느 점검항목과 맞물리는지를 정리한 것이다. 대상은 「감리기준 점검항목」 87건과 「감리영역별 기본 점검표」의 전문가 점검분야 10건을 합한 97건이다.

연계의 목적은 현행 점검항목을 대체하는 데 있지 않다. AI 서비스 구축 사업도 정보시스템 구축 사업이므로 현행 점검항목은 그대로 유효하며, 본 가이드는 그 위에 AI 특유의 점검 관점을 얹는 관계에 있다. 따라서 감리인은 현행 점검항목으로 감리를 수행하되, 아래 표에서 「수정」으로 표시된 항목에 대해서는 본 가이드의 대응 점검항목을 함께 확인하여야 한다.

1. 판정 구분

구분	의미	감리인의 조치
유지	현행 점검항목을 그대로 적용한다. AI 사업이라 하여 점검의 내용이나 기준이 달라지지 않는 항목이다.	현행 점검항목만으로 점검한다.
수정	현행 점검항목은 유효하나, AI 사업에서는 확인하여야 할 사항이 추가된다.	현행 점검항목으로 점검하되, 대응란의 본 가이드 항목을 함께 확인한다.
대체	현행 점검항목이 AI 사업에서는 실질적 의미를 갖지 못하여 본 가이드 항목으로 갈음한다.	본 가이드 대응 항목으로 점검한다.
해당없음	AI 서비스 구축 사업에는 적용되지 않는다.	점검 대상에서 제외한다.

※ 97건에 대한 판정 결과는 유지 23건 · 수정 74건이다. 「대체」와 「해당없음」이 없는 것은, 현행 점검항목 가운데 AI 사업에서 효력을 잃는 항목이 발견되지 않았기 때문이다.

2. 감리기준 점검항목 연계 (87건)

「감리기준 점검항목」의 5개 감리영역 87건을 대상으로 한다. 「본 가이드 대응」란의 코드는 부록 A의 점검항목 코드이며, 코드 옆의 항목명은 지면 관계상 생략하였다.

사업관리 (14건)

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석/설계	계약문서에서 정한 과업내용이 대비표(과업대비표, 요구사항추적표) 등에 반영되었는지 여부	**유지**	—	과업대비표·요구사항추적표 작성체계는 AI 사업도 동일하다.
분석/설계	범위관리체계를 적정하게 수립하여 관리하고 있는지 여부	**수정**	AG_PM04-1-03 AG_PM04-1-06	AI 기술 제약과 라벨링 도급이 범위·WBS에 반영돼야 한다.
분석/설계	변경관리체계를 적정하게 수립하고 이행하고 있는지 여부	**유지**	—	변경관리 절차와 파급효과 분석은 AI 사업도 동일하다.
분석/설계	일정관리체계를 적정하게 수립하고 진척관리하고 있는지 여부	**유지**	—	일정계획 수립과 진척통제 방식은 AI 사업도 동일하다.
분석/설계	자원관리체계를 적정하게 수립하여 관리하고 있는지 여부	**수정**	AG_PM04-1-06 AG_PM04-2-03	라벨링 도급인력과 라벨러 다양성이 자원관리에 포함된다.
분석/설계	의사소통관리체계를 적정하게 수립하여 관리하고 있는지 여부	**수정**	AG_PM04-1-02	AI 사업은 이해관계자 범위가 넓어 별도 매핑이 필요하다.

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석/설계	위험관리체계를 적정하게 수립하고 이행하고 있는지 여부	수정	AG_PM04-1-04 AG_PM04-1-05	AI 위험은 별도 분류체계로 식별기준을 보강해야 한다.
종료	세부 검사항목별과업이행 여부	수정	AG_SD41-1-03 AG_SD44-1-02	AI 성능목표는 정량 검사기준으로 명시해야 검증 가능하다.
종료	사용자요구사항을 충분히 반영하였는지 여부	수정	AG_PM04-3-04	AI 사업 사용자승인은 이해관계자 인수테스트로 확인한다.
종료	범위 변경관리를 적정하게 수행하였는지 여부	유지	—	범위 변경절차 준수와 추적성 점검은 AI 사업도 동일하다.
종료	진척관리를 적정하게 수행하였는지 여부	유지	—	정기 진척관리와 도구 활용 점검은 AI 사업도 동일하다.
종료	인력관리를 적정하게 수행하였는지 여부	유지	—	투입인력 계획 대비 실적 관리는 AI 사업도 동일하다.
종료	의사소통관리를 적정하게 수행하였는지 여부	유지	—	보고 주기와 피드백 점검은 AI 사업도 동일하다.
종료	위험관리를 적정하게 수행하였는지 여부	수정	AG_PM04-1-05 AG_SD44-1-21	AI 위험 분류별 대응매뉴얼 이행까지 확인해야 한다.

품질보증활동 (13건)

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석/설계	사업 목표 수립 여부	수정	AG_PM04-1-01 AG_SD41-1-03	AI 사업목표는 KPI와 모델 정확도 목표로 구체화된다.
분석/설계	방법론 및 절차/표준의 수립 여부	수정	AG_SD41-4-02 AG_SD41-4-05	AI는 데이터·모델 버전관리와 산출물 표준을 추가해야 한다.
분석/설계	품질보증활동 계획을 적정하게 수립하였는지 여부	수정	AG_SD41-4-04 AG_SD41-4-01	AI 산출물 품질보증과 편향성 정량지표를 계획에 담아야 한다.
분석/설계	총괄시험 계획을 적정하게 수립하였는지 여부	수정	AG_SD41-1-21 AG_SD42-1-16 AG_SD44-4-02	AI 신뢰성 시험과 자동화 평가가 총괄시험에 포함돼야 한다.
분석/설계	방법론 및 절차/표준의 준수 여부	유지	—	요구분석 단계 절차·표준 준수 점검은 AI 사업도 동일하다.
분석/설계	품질보증활동을 적정하게 수행하였는지 여부	유지	—	계획 대비 품질보증활동 수행 점검은 AI 사업도 동일하다.
분석/설계	사용자 요구사항 및 관련 산출물 간의 추적성, 일관성	유지	—	요구사항 추적·변경 관리체계는 AI 사업도 동일하다.
종료	방법론 및 절차/표준의 준수 여부	유지	—	방법론과 표준 준수 여부 확인은 AI 사업도 동일하다.
종료	품질보증활동을 적정하게 수행하였는지 여부	유지	—	종료단계 품질보증활동 수행 점검은 AI 사업도 동일하다.
종료	사용자 요구사항 및 관련 산출물 간의 추적성, 일관성	수정	AG_SD44-1-02 AG_SD44-4-02	AI 비기능 요구는 성능·신뢰성 시험 결과로 추적해야 한다.
종료	시스템 전환 전략을 적정하게 수립하였는지 여부	수정	AG_SD41-3-07 AG_SD45-1-03	데이터 이관 검증과 재학습 체계가 전환전략에 필요하다.
종료	5 교육계획을 적정하게 수립 및 실시하였는지 여부	수정	AG_PM04-2-05 AG_SD44-1-19	AI 사용자·운영 매뉴얼이 교육 교재의 기준이 된다.

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
종료	시스템을 인수하여 운영할 조직이 구성되고, 인수 준비가 이루어졌는가	유지	—	인수조직 구성과 인수준비 점검은 AI 사업도 동일하다.

응용시스템 (23건)

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석	현행 업무를 충분히 분석하였는지 여부	유지	—	현행업무 분석은 AI 여부와 무관한 일반 항목이다.
분석	현행 시스템의 업무 및 기능을 충분히 분석하였는지 여부	유지	—	현행 시스템 기능 분석은 사업 성격과 무관하다.
분석	사용자 요구사항 도출 및 분석의 충분성, 적정성	수정	AG_SD41-1-03 AG_SD41-2-02 AG_SD41-2-03	AI는 모델 정확도 목표·모델명세·설명가능성 요구가 추가된다.
분석	갭(GAP) 분석 활동의 적절성	유지	—	AS-IS/TO-BE 갭분석은 일반 분석기법이다.
분석	프로세스/이벤트 모델링과 기능 도출의 충분성, 적정성	유지	—	프로세스·이벤트 모델링 자체는 AI에서도 동일하다.
분석	사용자 접근통제 및 보안 분석을 적정하게 수행하였는지 여부	수정	AG_SD41-1-01 AG_SD41-1-17 AG_SD41-2-12	AI는 DLP·API 인증·접속로깅 요건을 함께 분석해야 한다.
설계	프로그램 분할을 적정하게 수행하였는지 여부	유지	—	프로그램 분할·공통모듈 도출은 일반 설계활동이다.
설계	업무 기능 설계를 누락 없이 상세하게 설계하였는지	수정	AG_SD42-2-04 AG_SD42-2-11 AG_SD42-2-15	AI는 추론 아키텍처·환각예방·안전모드 로직 설계가 필요하다.
설계	사용자 인터페이스가 편의성을 확보할 수 있도록 설계되었는지 여부	수정	AG_SD42-2-06 AG_SD42-2-07 AG_SD42-2-08	AI 화면은 한계 안내·설명가능성·신뢰점수 노출을 설계해야 한다.
설계	내/외부 인터페이스 설계의 충분성, 적정성	수정	AG_SD41-1-10 AG_SD42-2-14	AI 연계는 목적기반 API·추론 연계 프로토콜 설계가 추가된다.
설계	접근권한 및 통제 설계를 적정하게 수행하였는지 여부	수정	AG_SD42-1-07 AG_SD42-1-08 AG_SD42-1-09	AI는 관리자페이지·API·모델저장소 접근통제 설계가 필요하다.
설계	단위시험 계획을 적정하게 수립하였는지 여부	수정	AG_SD42-1-16	AI 단위시험은 별도 모델 테스트환경 구축계획이 필요하다.
구현	업무 기능 구현의 충분성, 완전성	수정	AG_SD43-2-03 AG_SD43-2-05 AG_SD43-2-06	AI 기능 구현은 프롬프트 관리체계 등 특화 점검이 필요하다.
구현	사용자 인터페이스가 편의성을 확보할 수 있도록 구현되었는지 여부	수정	AG_SD43-2-01 AG_SD44-2-04	AI 화면은 가이드 작동과 비동기 추론 진행표시를 확인해야 한다.
구현	내/외부 시스템 인터페이스 구현의 충분성, 완전성	유지	—	인터페이스 구현 확인은 AI에서도 동일하게 적용된다.
구현	접근 권한 및 통제기능을 정확하게 구현하였는지 여부	수정	AG_SD43-1-02 AG_SD43-1-04	AI는 모델저장소 비인가 접근차단 구현을 함께 봐야 한다.

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
구현	단위 시험을 실시하였는지 여부	유지	—	단위시험 수행·결과관리는 일반 구현활동이다.
구현	통합 시험 계획을 적정하게 수립하였는지 여부	수정	AG_SD41-1-21 AG_SD42-1-17	AI 통합시험 계획에 신뢰성·자동화 평가가 포함돼야 한다.
시험	통합시험 실시 및 검증을 적정하게 수행하였는지 여부	수정	AG_SD44-2-11 AG_SD44-2-12 AG_SD44-4-02	AI 통합시험은 연계 정합성과 신뢰성 평가를 함께 검증한다.
시험	시험결과의 관리 및 개선을 적정하게 수행하였는지 여부	유지	—	시험결과 관리·오류개선 절차는 일반 항목이다.
시험	사용자 지침서를 적정하게 작성하였는지 여부	수정	AG_PM04-2-05 AG_SD42-2-06 AG_SD43-2-01	AI 지침서는 한계·유의사항 안내와 매뉴얼 목차 요건이 있다.
시험	사용자 인수시험 계획 및 수행을 적정하게 수행하였는지 여부	수정	AG_PM04-3-01 AG_PM04-3-04	AI 인수시험은 평가단 XAI 평가와 이해관계자 검증을 포함한다.
전개	사용자 인수시험을 수행하였는지 여부	수정	AG_PM04-3-04 AG_PM04-3-02	전개 시 인수승인과 윤리기준 최종 검토를 함께 확인한다.

데이터베이스 (18건)

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석	현행 업무 관련 데이터에 대한 분석 충분성	수정	AG_SD41-3-01 AG_SD41-3-09	AI는 학습데이터 특성·선정기준과 품질기준 분석이 추가된다.
분석	현행 시스템의 데이터 현황을 충분히 분석하였는지 여부	수정	AG_SD41-3-04	현행 데이터 현황에 AI용 메타데이터 확보 점검이 필요하다.
분석	사용자 요구사항 도출 충분성	수정	AG_SD41-3-10	AI는 벡터DB·RAG 파이프라인 요구정의가 추가된다.
분석	데이터 흐름 정의가 명확한지 여부	수정	AG_SD41-3-03 AG_SD41-3-05	AI 데이터흐름은 계보 추적과 의사결정 감사추적이 필요하다.
분석	엔티티 정의, 관계설정 등 데이터 모델링의 명확성 및 엔티티/프로세스간 연관관계 설정의 적정성	수정	AG_SD42-3-09	AI는 정형DB와 벡터DB 하이브리드 모델 정합성을 봐야 한다.
분석	데이터에 대한 접근통제 및 보안 분석을 적정하게 수행하였는지 여부	수정	AG_SD41-1-01 AG_SD41-3-06	AI는 비식별화·가명처리와 민감정보 유출차단 분석이 필요하다.
설계	데이터베이스 테이블을 적정하게 설계하였는지 여부	수정	AG_SD42-3-05 AG_SD42-3-11	AI는 암호화 컬럼과 정제 전후 적재 공간 분리 설계가 필요하다.
설계	공통 코드를 적정하게 설계하였는지 여부	수정	AG_SD41-3-02	AI 모델 공통분류 코드체계 명세요건을 함께 봐야 한다.
설계	데이터베이스 성능을 고려하여 설계하였는지 여부	수정	AG_SD42-3-08 AG_SD42-3-10	AI는 성능 최적화 실행계획과 성능지표 마트 설계가 추가된다.
설계	접근권한 및 통제 설계를 적정하게 수행하였는지 여부	수정	AG_SD42-3-03	AI는 비식별화·가명화 뷰 기반 권한 설계를 함께 본다.
설계	백업 및 복구대책을 적정하게 수립하였는지 여부	수정	AG_SD41-1-02 AG_SD44-1-17	AI는 모델 버전 백업·롤백 대책을 함께 수립해야 한다.
설계	초기데이터 구축에 대한 계획을 적정하게 수립하였는지 여부	수정	AG_SD41-3-08 AG_SD41-3-11	AI 초기데이터는 학습·검증·시험 분할과 홀드아웃 확보가 핵심이다.

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
설계	기존 데이터에 대한 전환 계획을 적정하게 수립하였는지 여부	**수정**	AG_SD41-3-07	AI는 레거시 데이터 이관 검증 요건을 함께 계획해야 한다.
구현	데이터베이스 테이블 구현을 적정하게 수행하였는지 여부	**수정**	AG_SD43-3-09	AI는 정제 전후 DB 분리 환경 구현을 함께 확인한다.
구현	데이터베이스 성능을 충분히 고려하여 구현하였는지 여부	**수정**	AG_SD43-2-04 AG_SD44-3-13	AI는 캐싱 히트율과 인덱스·파티셔닝 효율성을 함께 본다.
구현	데이터 접근권한 및 통제를 정확하게 구현하였는지 여부	**수정**	AG_SD43-2-02 AG_SD44-3-07	AI는 마스킹 뷰 조회와 DB 권한 분리 구현을 함께 확인한다.
구현	단위 시험을 실시하였는지 여부	**수정**	AG_SD43-3-06 AG_SD44-3-11	AI DB시험은 정합성 배치검증과 ERD·물리DB 일치를 본다.
전개	데이터 이관계획 및 데이터 전환/검증을 적합하게 수행하였는지 여부	**수정**	AG_SD43-3-04 AG_SD43-3-05	AI 이관은 ETL 검증과 이상데이터 제거율 확인이 추가된다.

시스템구조 및 보안 (19건)

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
분석	현행 및 신규 시스템의 운영환경을 충분히 분석하였는지 여부	**수정**	AG_SD41-1-06 AG_SD41-1-07	AI 인프라 도입환경·GPU 용량산정 분석을 함께 점검
분석	시스템 관련 사용자요구사항 도출 및 분석의 충분성, 적정성	**수정**	AG_SD41-1-03 AG_SD41-1-02 AG_SD41-1-22	모델 정확도 목표와 성능저하 임계값 요구가 추가 필요
분석	시스템 보안 요건을 분석하였는지 여부	**수정**	AG_SD41-1-12 AG_SD41-1-15 AG_SD41-1-01	모델 공격·AI 공급망 위협을 보안요건에 포함해야 함
분석	시스템 아키텍처에 대한 검증을 적정하게 수행하였는지 여부	**수정**	AG_SD41-1-04 AG_SD41-1-08 AG_SD41-1-19	AI 구성도·컨테이너·오픈소스 구성요소 검증이 추가됨
분석	최종 아키텍처를 적정하게 수립하였는지 여부	**수정**	AG_SD41-1-04 AG_SD41-1-07 AG_SD42-1-18	AI 시스템 구성도와 SLM 등 모델 선정 타당성 병행
분석	표준준수를 위한 요구사항 식별 여부	**수정**	AG_SD41-2-10 AG_SD41-4-05	다국어·접근성 지원과 AI 산출물 작성표준 식별 필요
설계	시스템 구성요소에 대한 상세설계를 충분하게 수행하였는지 여부	**수정**	AG_SD42-2-15 AG_SD42-1-19	AI 모델 아키텍처·추론서버 스케일링 설계를 함께 봄
설계	시스템 보안에 대한 상세설계를 적정하게 수행하였는지 여부	**수정**	AG_SD42-1-06 AG_SD42-1-08 AG_SD42-1-15	모델 공격방어·API 인가·클라우드 보안그룹 설계 추가
설계	시스템 설치 및 검증 계획을 적정하게 수립하였는지 여부	**유지**	—	장비 설치·검증 계획은 AI 사업에서도 달라지지 않음
설계	시스템 전환 계획을 적정하게 수립하였는지 여부	**수정**	AG_SD41-3-07 AG_SD41-1-02	레거시 데이터 이관 검증과 백업·DR 체계를 함께 점검
구현	시스템 도입/설치 및 보안환경의 구축을 충분하게 수행하였는지 여부	**수정**	AG_SD43-1-02 AG_SD43-1-03 AG_SD45-1-02	모델저장소 접근차단·스토리지 암호화 구축 확인 필요

단계	현행 점검항목	구분	본 가이드 대응	판정 사유
구현	시스템 구성요소에 대한 검증을 적정하게 수행하였는지 여부	**수정**	AG_SD43-1-01 AG_SD44-1-10 AG_SD44-1-20	AI 프레임워크 CVE·라이선스·모델 방어 검증이 추가됨
구현	3 시스템 시험 계획을 적정하게 수립하였는지 여부	**수정**	AG_SD42-1-16 AG_SD42-1-17 AG_SD41-1-21	AI 신뢰성 테스트와 자동화 시험 계획을 함께 수립해야 함
시험	시험환경을 충분하게 구축하였는지 여부	**수정**	AG_SD42-1-16 AG_SD42-3-06 AG_SD41-3-11	학습·검증 데이터 분할과 테스트 데이터 확보를 함께 봄
시험	시스템 시험 실시 및 검증을 적정하게 수행하였는지 여부	**수정**	AG_SD44-1-02 AG_SD44-1-14 AG_SD44-1-04	모델 성능지표 달성과 안전모드 전환 시험이 추가됨
시험	시스템 최적화 활동을 적정하게 수행하였는지 여부	**수정**	AG_SD43-2-04 AG_SD42-1-19	추론 캐싱 히트율·오토스케일링 최적화를 함께 점검
시험	운영자 지침서를 적정하게 작성하였는지 여부	**수정**	AG_SD44-1-19 AG_SD44-1-21	AI 운영매뉴얼 목차와 위험대응 매뉴얼을 함께 확인
전개	운영환경의 설치 및 배포를 완전하게 수행하였는지 여부	**수정**	AG_SD45-1-01 AG_SD45-1-02 AG_SD45-1-05	CI/CD 자동배포·컨테이너 이미지·리소스 가용성 점검
전개	시스템 및 업무 전환을 적정하게 수행하였는지 여부	**수정**	AG_SD43-3-04 AG_SD41-3-07	레거시 ETL 이행 검증으로 전환 데이터 일관성 확인

3. 전문가 점검분야 연계 (10건)

「감리영역별 기본 점검표」에서 전문가가 별도로 점검하는 10개 분야를 대상으로 한다. 이 분야는 AI 서비스에서 점검의 대상 자체가 달라지는 경우가 많으므로, 대응란의 항목을 반드시 함께 확인하여야 한다.

점검분야	단계	현행 점검 내용	구분	본 가이드 대응	판정 사유
SW개발보안	설계단계	소스코드 보안약점 기준 중 분석/설계에 반영하여야 하는 항목은 적정하게 수행되었는가?	**수정**	AG_SD42-1-13 AG_SD42-1-14 AG_SD42-2-12	보안약점 예방 아키텍처와 악성 프롬프트 필터 설계 포함
SW개발보안	종료단계	소스코드 보안약점 점검 도구를 이용한 49개 보안약점 점검	**수정**	AG_SD43-1-04 AG_SD43-1-05 AG_SD44-2-08	보안약점 필터·프롬프트 인젝션 방어 동작을 함께 점검
성능/가용성	전 단계	시스템 성능 및 자원의 사용 등은 적정하게 확보되었는가?	**수정**	AG_SD44-1-14 AG_SD44-1-18 AG_SD44-2-07	추론 지연시간과 GPU 등 리소스 쿼터 점검이 추가됨
기능점수(FP)	전 단계	사업비 산출내역서의 기능점수는 적정하게 산정되어 작성되었는가?	**유지**	—	기능점수 산정 기준은 AI 사업에서도 동일하게 적용
UI/UX	전 단계	웹표준 및 UI/UX가 적정하게 구현되었는가?	**수정**	AG_SD42-2-01 AG_SD42-2-06 AG_SD44-2-04	AI 한계 안내·다크패턴 배제 등 AI 특화 UI 요소 추가

점검분야	단계	현행 점검 내용	구분	본 가이드 대응	판정 사유
웹접근성 및 호환성	전 단계	웹접근성 진단 4원칙이 적정하게 준수되었는가?	**수정**	AG_SD41-2-10	AI 서비스의 다국어·접근성 지원 요건을 함께 확인
보안취약점	전 단계	대상 목표시스템에 대한 보안 점검과 안전 조치가 적정하게 수행되었는가?	**수정**	AG_SD44-1-10 AG_SD44-1-09 AG_SD43-1-01	CVE·WAF 외 적대적 사례 공격 방어 점검이 추가됨
개인정보보호	전 단계	개인정보에 대한 취급 및 보호 조치는 적절하게 되었는가?	**수정**	AG_SD41-3-06 AG_SD44-3-03 AG_SD44-3-04	학습데이터 비식별화·파기·옵트아웃 처리 점검 필요
기능테스트	전 단계	단위업무 시스템의 세부 기능에 대한 구현이 적정하게 이루어 졌는가?	**수정**	AG_SD44-2-02 AG_SD44-2-03 AG_SD44-2-05	XAI 동작·환각 억제·오작동 기본값 응답 시험 추가
데이터 품질	전 단계	목표시스템의 데이터 품질은 확보되었는가?	**수정**	AG_SD41-3-09 AG_SD43-3-06 AG_SD44-3-11	AI 학습데이터 품질기준과 정합성 배치검증을 함께 봄

※「감리기준 점검항목 V3.0」(2024.1) 및「감리영역별 기본 점검표」를 기준으로 작성하였다. 해당 문서가 개정되는 경우 본 연계표도 함께 갱신되어야 한다.

부록 F. 기술 유형별 커버리지 현황

본 부록은 본 가이드의 점검항목이 어떤 AI 기술 유형을 다루고 있으며, 어떤 유형을 아직 다루지 않는지를 정리한 것이다. 감리원은 착수 전에 해당 사업이 적용하는 기술이 본 가이드의 반영 범위 안에 있는지를 확인하고, 반영 범위 밖의 기술에 대해서는 별도의 점검 방법을 계획하여야 한다.

공공 부문에서 감리 목적으로 그대로 적용할 수 있는 인공지능 기술 분류체계가 아직 단일하게 확립되어 있지 않으므로, 본 부록은 본 가이드가 실제로 다루고 있는 기술을 모델·데이터/검색·응용/서비스·인프라/운영의 네 계층으로 나누어 제시한다. 공식 분류체계가 확정되면 그 체계에 맞추어 재작성한다.

F.1 반영 수준 구분

표기	반영 수준	감리원의 대응
●	전면 반영	**본 가이드의 점검항목을 그대로 적용한다**
◐	부분 반영	**해당 점검항목을 적용하되, 비고에 적힌 미포함 범위는 별도 점검 방법을 계획한다**
○	미반영	**본 가이드에 점검항목이 없다. F.2의 대응 지침에 따른다**

F.2 기술 유형별 커버리지

계층	기술 유형	반영	해당 점검항목	비고
모델	대규모 언어모델(LLM) 활용	●	AG_SD41-1-03, AG_SD41-1-21, AG_SD42-1-02 AG_SD42-2-15, AG_SD44-1-02, AG_SD44-1-13	모델 선정·성능목표·신뢰성 검증을 전 단계에 걸쳐 점검
	경량 언어모델(SLM)	◐	AG_SD42-1-18	활용 요건만 다루며, 경량화 기법별 성능 저하 검증은 미포함
	파인튜닝·추가학습	◐	AG_SD41-3-08, AG_SD41-3-11, AG_SD42-3-02 AG_SD44-3-02	학습·검증 데이터 분할과 버전 매핑은 다루나, 학습 절차·하이퍼파라미터 관리 점검은 미포함
	자체 사전학습(파운데이션 모델 구축)	○	—	상용·공개 모델 활용을 전제로 작성하였으므로 미반영
	인간 피드백 기반 정렬(RLHF)	○	—	정렬 데이터 수집·보상모델 검증 절차 미반영
데이터·검색	학습데이터 구축·라벨링	●	AG_PM04-1-06, AG_PM04-1-07, AG_PM04-1-08 AG_PM04-1-09, AG_PM04-1-10, AG_PM04-2-01 AG_PM04-2-02, AG_PM04-2-03, AG_SD41-3-09 AG_SD43-3-01	도급관리·요구정의·교차검수·편향배제까지 전 과정 점검
	검색증강생성(RAG)·벡터DB·임베딩	◐	AG_SD41-3-10, AG_SD42-3-09, AG_SD43-2-04	요구정의·데이터모델 정합성·캐싱은 다루나, 검색 품질 지표(nDCG·MRR 등) 검증 점검은 미포함

계층	기술 유형	반영	해당 점검항목	비고
	데이터 품질·정합성	●	AG_SD41-3-09, AG_SD42-3-04, AG_SD43-3-06 AG_SD44-3-10	품질기준 명시부터 배치검증·불균형 해소까지 점검
	데이터 드리프트·재학습	●	AG_SD41-1-22, AG_SD41-1-23, AG_SD43-3-10 AG_SD44-2-09, AG_SD45-1-03	임계값 설정·감지·재학습 트리거 기준을 점검
응용·서비스	프롬프트 설계·관리	●	AG_SD41-2-06, AG_SD42-2-12, AG_SD43-2-03 AG_SD44-2-08	설계 관리체계와 악성 프롬프트 방어를 함께 점검
	환각(할루시네이션) 통제	●	AG_SD41-2-05, AG_SD42-2-04, AG_SD44-2-03	요구정의·설계·시험 3단계에 걸쳐 점검
	설명가능성(XAI)·투명성	●	AG_PM04-3-01, AG_SD41-2-02, AG_SD41-2-03 AG_SD41-2-04, AG_SD42-2-07, AG_SD44-2-02	모델명세 공개·설명 제공·신뢰점수 노출을 점검
	편향·공정성	●	AG_PM04-1-10, AG_SD41-2-01, AG_SD41-3-01 AG_SD41-4-01, AG_SD42-2-01, AG_SD42-2-02 AG_SD42-2-03, AG_SD42-3-01, AG_SD44-2-01 AG_SD44-3-01, AG_SD44-4-01	수집 단계 편향배제부터 지표 설계·모니터링·훈련 이행 점검까지 전 단계에 대응
	대화형 서비스(챗봇·상담)	◐	AG_SD43-2-06	민원·상담 특화 요건만 별도 점검. 다중턴 대화 품질 검증은 미포함
	비정형 문서 처리	◐	AG_SD43-2-05	문서 처리 AI 요건만 점검. 문서 유형별 추출 정확도 검증 기준은 미포함
	멀티모달(이미지·음성 생성·인식)	○	—	텍스트 기반 생성형 AI를 전제로 작성. 생성 이미지 표시 의무 등 향후 반영 대상
	AI 에이전트·자율 실행	◐	AG_SD42-2-16	안전정지 루프 통제만 점검. 다단계 실행 경로·권한 상승 통제는 미포함
	MCP·외부 도구 연계	◐	AG_SD41-1-10, AG_SD42-2-14	레거시 API 설계와 연계 프로토콜 관점에서만 접근. 도구 연계 인증·권한 범위 점검은 미포함
인프라·운영	추론 서버·스케일링	●	AG_SD41-1-07, AG_SD42-1-19, AG_SD44-1-03 AG_SD45-1-05	용량 산정부터 자동 확장·가용성까지 점검
	GPU/NPU 자원 관리	◐	AG_SD44-1-18	리소스 쿼터만 점검. 연산 자원 공동활용·배분 정책은 미포함
	AI 특화 보안	●	AG_SD41-1-12, AG_SD41-1-14, AG_SD41-1-15 AG_SD41-1-16, AG_SD42-1-06, AG_SD43-1-01 AG_SD44-3-06	적대적 공격·모델 추출·데이터 중독·공급망까지 대응
	망분리·폐쇄망 학습	●	AG_SD41-1-20, AG_SD42-1-15, AG_SD44-1-11 AG_SD44-1-12	요구정의·설계·시험 전 단계 점검

계층	기술 유형	반영	해당 점검항목	비고
	MLOps·LLMOps 자동화	◐	AG_SD41-1-09, AG_SD42-1-05, AG_SD45-1-01	CI/CD 파이프라인 보안·자동화는 점검하나, 실험 추적·모델 레지스트리 운영 체계는 미포함
	온디바이스·엣지 AI	○	—	중앙 추론 인프라를 전제로 작성. 단말 배포·갱신 관리 미반영
	AI 데이터센터 인프라	○	—	정보시스템 구축사업의 감리 범위를 벗어나므로 미반영

F.3 미반영 기술이 포함된 사업의 대응

반영 수준이 ○인 기술이 사업에 포함된 경우, 감리원은 본 가이드에 점검항목이 없다는 사실만으로 해당 기술을 점검 대상에서 제외하여서는 안 된다. 다음 순서로 점검 방법을 마련한다.

❑ 첫째, 해당 기술이 유발하는 위험을 감리계획서에 별도로 식별하여 기재한다. 위험이 식별되지 않으면 점검 자체가 계획되지 않는다.

❑ 둘째, 부록 B에 등재된 법령·가이드라인·표준 중 해당 기술에 적용되는 요건을 찾아 점검 기준으로 삼는다. 예를 들어 이미지·음성 생성 기능이 포함된 사업은 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제31조의 생성물 표시 의무가 그대로 적용된다.

❑ 셋째, 반영 수준이 ●인 인접 기술 유형의 점검항목 중 성격이 유사한 항목을 조정하여 적용한다. 조정 내역은 부록 D의 조정 매트릭스에 「수정」으로 기재한다.

❑ 넷째, 위 방법으로도 점검할 수 없는 사항은 감리계획서에 점검 제약사항으로 명시하고, 발주기관과 점검 범위를 사전에 합의한다.

감리 현장에서 ○ 유형의 기술이 반복하여 확인되는 경우에는 1.5절의 개정 착수 기준에 따라 해당 기술의 점검항목 신설을 검토한다.

※ 본 표의 반영 수준은 본 가이드 발간 시점의 점검항목 구성을 기준으로 한 것이므로, 점검항목이 추가되면 함께 갱신한다.

정보시스템감리협회
KOREA INFORMATION SYSTEM SUPERVISION ASSOCIATION

AI SERVICE AUDIT GUIDE · 2026

AI 서비스 구축 사업

감리 점검 가이드

생성형 AI 정보화사업을 위한 단계별·영역별 감리 점검체계