01 | 생성형 AI 감리 수행 사례

김일수 | 이사
㈜씨에이에스
㈔정보시스템감리협회

2025년도 정보시스템감리 공동 심포지엄

(생성형) AI감리 수행 사례 -감리원 관점-

김일수 | 이사
㈜씨에이에스
(사)정보시스템감리협회

CONTENTS

1 정보화 사업의 감리역할
2 정보화 사업의 감리 가이드 구성체계
3 AI 관련 감리 참고 자료
4 정보시스템 감리와 AI 감리의 차이점
5 (생성형) AI 감리 수행 사례
6 (생성형) AI 감리 수행시 고려사항

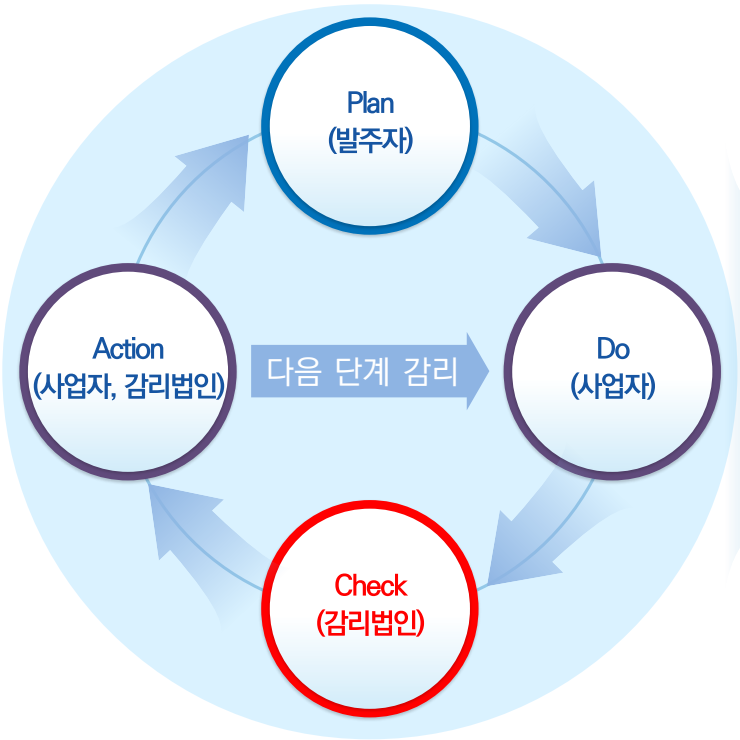


CHAPTER

1. 정보화 사업의 감리역할

1) PDCA 사이클로 구성한 발주자, 사업자, 감리 역할





"감리"란 **Plan**에서 정의된 요구사항이
Do에서 Plan의 요구사항이 계획에 반영되어 실행되고 있는지
Check하여 감리수행결과보고서를 제출하면
사업자는 지적된 사항에 대해 시정조치계획서/결과서를 통한
Action을 수행하면 감리는 점검 후
시정조치결과보고서를 제출한다.

이해관계자의 주요 산출물

1 Plan
- 제안요청서(RFP)

2 Do
- 계획 : 제안서, 사업수행계획서, 요구사항정의서, WBS 등
- 실행 : 프로그램 소스/실행 파일, 단계별 산출물 등

3 Check
- (단계별) 감리계획서, (단계별) 감리수행결과보고서

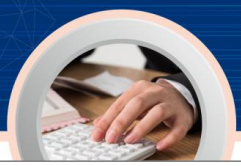
4 Action
- 수행사 : (단계별) 시정조치계획서/결과서, 조치 산출물 등
- 감리법인 : (단계별) 시정조치결과보고서

- 정보시스템 감리기준: "감리"란 발주자와 사업자 등의 이해관계로부터 독립된 자가 정보시스템의 효율성을 향상시키고 안전성을 확보하기 위하여 제3자의 관점에서 정보시스템의 구축 및 운영 등에 관한 사항을 종합적으로 점검하고 문제점을 개선하도록 하는 것을 말한다.

CHAPTER

2. 정보화 사업의 감리 가이드 구성체계

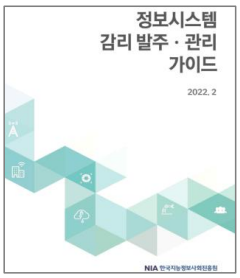
1) 수행 주체별 감리 가이드





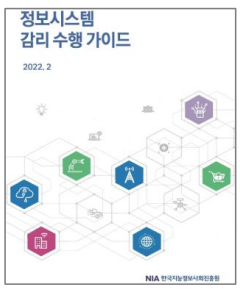
[그림 1-1] 감리점검 가이드 근거

- 출처: SW사업자를 위한 정보시스템 감리 준비 및 대응 가이드



1. 목적

- 본 가이드는 행정기관 등에서 추진하는 정보시스템 감리용역을 효율적으로 발주하고 관리할 수 있도록 하기 위해 작성되었으며 이를 통해 발주자가 감리를 활용하여 정보화 사업을 성공적으로 수행하도록 하기 위함이다.
- 이를 위해 감리용역의 기획단계부터 종료단계까지 각 단계별로 활동(Activity)을 정의하고 활동별 수행내용과 관련근거 및 서식, 결과물, 세부내용, 기준, 사례 등을 제공하여 감리 발주·관리를 효율적으로 할 수 있도록 하였다. 또한, 감리 발주·관리에서 필요한 팁과 FAQ를 제공하여 도움이 될 수 있는 가이드가 되도록 하였다.
- 본 가이드를 통해 발주자는 감리시행여부 판단, 감리대가 산정, 제안요청서 작성, 제안서 평가, 계약체결, 감리시행 관리, 감리결과 조치, 감리용역 검사, 감리용역결과 평가를 등 감리 발주·관리를 효율적으로 수행할 수 있을 것으로 기대된다.



1.1 목적

이 가이드는 감리법인 및 감리원이 전자정부법 및 '정보시스템 감리기준'에 근거하여 정보시스템 감리를 수행할 때 참조할 수 있도록 다음 각 사항 등에 관한 실무지침을 제시할 목적으로 작성되었다.

1) 감리 절차, 감리 수행방법 및 참조사항
2) 감리점검체계에 따른 표준 점검항목
3) 감리계획서 및 보고서 등의 서식
4) 그 밖에 감리 수행을 위해 필요한 사항



1.1 목적

본 가이드는 정보시스템 감리를 실시함에 있어서 정보화 사업을 수행하는 SW사업자가 사전에 자료를 준비하고 감리 과정에서의 대응 및 감리 개선권고사항에 따른 시정조치를 수행하는데 참조할 수 있는 실무 지침으로 작성하였다.

1) 단계별 감리에 대한 사전 준비
2) 현장 감리 절차별 감리 대응
3) 감리수행결과에 따른 시정조치계획 및 결과 작성

2) 정보시스템 감리 점검 가이드



4 정보화사업 유형별 표준 점검항목

정보화사업 유형별 표준 점검항목은 「정보시스템 감리 점검 프레임워크 V3.0」에 따라 정보기술 아키텍처(EA) 구축, 정보화전략계획(ISP) 수립, 정보시스템 개발, 데이터베이스 구축, 정보시스템 운영 및 유지보수 등 감리 수행시 권장하는 사업 유형별 점검항목을 제시한 것이다.

따라서 해당 사업의 규모나 특성, 감리의 형태, 발주자의 요구사항 등을 고려하여 제시된 점검항목을 선택적으로 적용할 수 있다.

정보시스템 감리 점검 프레임워크 V3.0



• 출처: 정보시스템 감리 수행 가이드

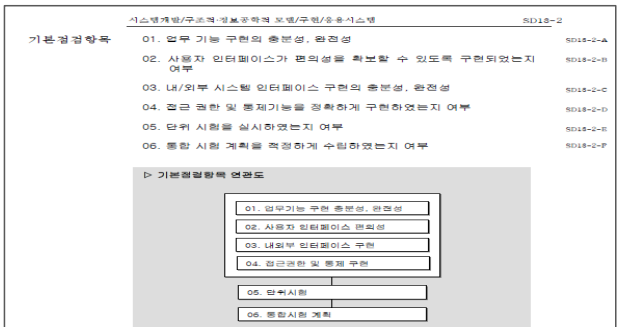
4.3. 시스템 개발 사업

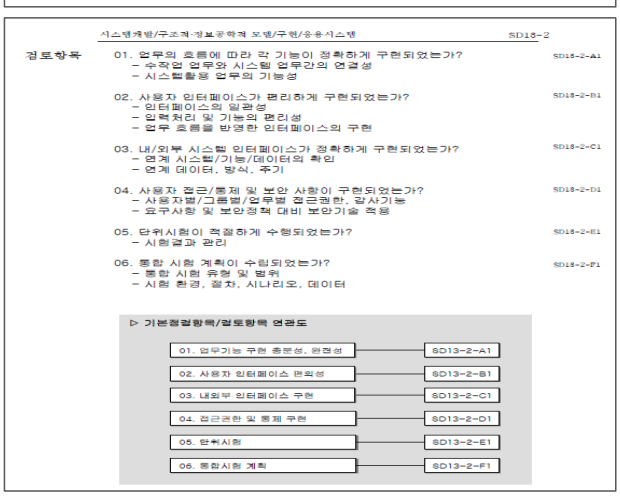
4.3.1. 구조적/정보공학적 개발 모델

분석	응용 시스템	1. 현행 업무를 충분히 분석하였는지 여부 2. 현행 시스템의 업무 및 기능을 충분히 분석하였는지 여부 3. 사용자 요구사항 도출 및 분석의 충분성, 적정성 4. 프로세스, 이벤트 모델링 및 기능 도출의 충분성, 적정성 5. 사용자 접근통제 및 보안 분석을 적정하게 수행하였는지 여부
	데이터 베이스	1. 현행 업무관련 데이터를 충분히 식별하였는지 여부 2. 현행 시스템의 데이터 현황을 충분히 분석하였는지 여부 3. 사용자 요구사항 도출 및 분석의 충분성, 적정성 4. 데이터 흐름 정의가 명확한지 여부 5. 엔티티 정의, 관계설정 등 데이터 모델링의 명확성 및 엔티티/프로세스간 연관관계 설정의 적정성 6. 데이터에 대한 접근통제 및 보안 분석을 적정하게 수행하였는지 여부
설계	응용 시스템	1. 프로그램 분할을 적정하게 수행하였는지 여부 2. 업무 기능 설계를 누락 없이 상세하게 설계하였는지 여부 3. 사용자 인터페이스가 편의성을 확보할 수 있도록 설계되었는지 여부 4. 내/외부 인터페이스 설계의 충분성, 적정성 5. 접근권한 및 통제 설계를 적정하게 수행하였는지 여부 6. 단위시험 계획을 적정하게 수립하였는지 여부
	데이터 베이스	1. 데이터베이스 테이블을 적정하게 설계하였는지 여부 2. 공통 코드를 적정하게 설계하였는지 여부 3. 데이터베이스 성능을 고려하여 설계하였는지 여부 4. 접근권한 및 통제 설계를 적정하게 수행하였는지 여부 5. 백업 및 복구대책을 적정하게 수립하였는지 여부 6. 초기데이터 구축에 대한 계획을 적정하게 수립하였는지 여부 7. 기존 데이터에 대한 전환 계획을 적정하게 수립하였는지 여부
구현	응용 시스템	1. 업무 기능 구현의 충분성, 완전성 2. 사용자 인터페이스가 편의성을 확보할 수 있도록 구현되었는지 여부 3. 내/외부 시스템 인터페이스 구현의 충분성, 완전성 4. 접근 권한 및 통제기능을 정확하게 구현하였는지 여부 5. 단위 시험을 실시하였는지 여부 6. 통합 시험 계획을 적정하게 수립하였는지 여부
	데이터 베이스	1. 데이터베이스 테이블 구현을 적정하게 수행하였는지 여부 2. 데이터베이스 성능을 충분히 고려하여 구현하였는지 여부 3. 데이터 접근권한 및 통제를 정확하게 구현하였는지 여부 4. 단위 시험을 실시하였는지 여부

3) 감리보고서의 점검항목 적용 사례







• 출처: 정보시스템 감리 점검 해설서

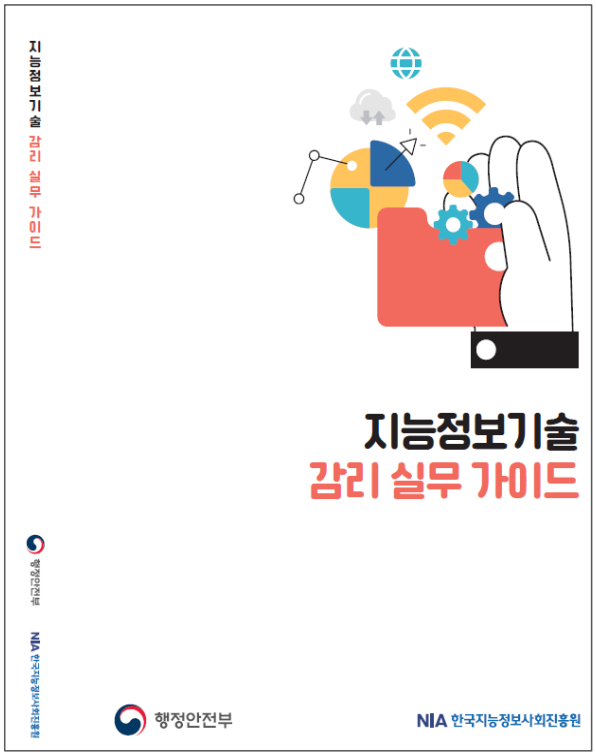
2. 응용시스템

가. 점검항목별 점검결과

점검항목		점검결과
1. 단위업무의 세부점사항목별 과업이행 여부	01. 단위업무의 세부점사항목별 과업이 적정하게 이행되었는가? – 세부점사항목별 완료 여부 – 완료된 과업이 적합한지, 또는 부적합한지 여부	나.(1)(2)번 참조
2. 업무 기능 구현의 충분성, 완전성	01. 업무의 흐름에 따라 각 기능이 정확하게 구현되었는가? – 수작업 업무와 시스템 업무간의 연결성 – 시스템활용 업무의 기능성	나.(1)(2)번 참조
3. 사용자 인터페이스의 편의성 확보 여부	01. 사용자 인터페이스가 편리하게 구현되었는가? – 인터페이스의 일관성 – 입력처리 및 기능의 편리성 – 업무 흐름을 반영한 인터페이스 – 웹 접근성/웹 표준 준수 여부	나.(1)(2)번 참조
4. 내/외부 시스템 인터페이스 구현의 충분성, 완전성	01. 내/외부 시스템 인터페이스가 정확하게 구현되었는가? – 연계 시스템/기능/데이터의 확인 – 연계 데이터, 방식, 주기	나.(1)(2)번 참조
5. 접근 권한 및 통제기능 등 보안성을 고려하여 구현하였는지 여부	01. 사용자 접근/통제 및 보안 사항이 구현되었는가? – 사용자별/그룹별/업무별 접근권한, 감사기능 – 요구사항 및 보안정책 대비 보안기술 적용 02. 소스코드의 품질(Secure Coding)을 확보하였는가?	나.(1)번 참조
6. 단위 테스트를 실시하였는지 여부	01. 단위테스트가 적절하게 수행되었는가? – 테스트결과 관리	나.(1)(2)번 참조
7. 통합 테스트 계획, 실시 및 검증을 적정하게 수행하였는지 여부	01. 통합 테스트 계획이 수립되었는가? – 통합 테스트 유형 및 범위, 절차, 시나리오 등 02. 통합테스트가 계획에 따라 적정하게 실시되었는가? – 테스트 환경, 절차, 시나리오, 데이터의 적용 03. 통합테스트를 통해 내외부 시스템 간의 연계 및 구현된 시스템 간의 기능 완전성과 데이터 무결성이 확보되었는가? – 내부 시스템 간의 연계 및 통합 – 외부 시스템과의 연계 기능의 완전성 – 연계 데이터의 무결성	나.(1)(2)번 참조
점검항목		**점검결과**
8. 테스트결과의 관리 및 개선을 적정하게 수행하였는지 여부	01. 테스트 결과에 대한 관리 및 개선이 이루어졌는가? – 테스트 유형별 시나리오, 데이터, 결과 관리의 적정성 – 테스트 결과에 따른 오류 개선	나.(1)(2)번 참조
9. 사용자 지침서를 적정하게 작성하였는지 여부	01. 사용자 지침서가 적정하게 작성되었는가? – 구현 기능에 대한 사용방법 설명 충분성 – 지침서 내용의 현행화 여부 – 사용자의 교육, 이해 및 사용 편의성 확보 여부	나.(1)(2)번 참조
10. 사용자 인수테스트 계획 및 수행을 적정하게 수행하였는지 여부	01. 사용자 인수테스트에 대한 계획이 수립되었는가? – 사용자 테스트 환경, 절차, 시나리오, 데이터 02. 사용자 인수테스트를 통하여 최종 사용자의 승인을 획득하였는가? – 사용자 요구사항 및 사업목표의 달성 여부 – 테스트방법, 승인기준 및 절차, 테스트결과	나.(1)(2)번 참조

4) 지능정보기술 감리 실무 가이드(2023.2)







5) 빅데이터 감리 점검가이드 (1/2)



2.1 빅데이터 감리 점검가이드의 구성

정보시스템 감리는 준거성과 실증적 실사방법에 의해 위험을 도출하고 통제를 제시하는 분야로 기본이 기 설정된 작업절차와 활동에 따라 수행되었는지를 점검하는 준거성 검사이다. 이에 본 점검가이드에서는 최소한의 준거성을 점검할 수 있는 가이드라인 존재, 공공정보화 사업 발주현황, 그리고 위험을 인식하고 통제가능성을 기준으로 빅데이터 정보시스템 개발의 단계별/영역별로 점검가이드를 작성하였다.

빅데이터 감리 점검가이드는 기존 감리점검 프레임워크의 사업유형을 토대로 새롭게 기본점검항목과 검토항목을 적용하였다.

[표 2-1] 빅데이터 기본점검항목 및 검토항목

단계	영역 구분	빅데이터 감리가이드		기존 감리가이드
		기본점검항목	검토항목	기본 점검항목
분석	시스템 구조	-	-	5
	응용시스템	1	6	5
	데이터베이스	2	4	6
설계	시스템 구조	1	4	4
	응용시스템	3	7	6
	데이터베이스	1	3	7
구현	시스템 구조	-	-	3
	응용시스템	1	4	6
	데이터베이스	-	-	4
시험	시험활동	-	-	7
전개	운영준비	1	1	4
합계		10	26	57

3.2 설계단계 점검

3.2.1 설계단계 감리 점검 개요

빅데이터 설계단계에서는 분석단계에서 정의된 요구사항이 설계에 잘 반영이 되었는지를 점검한다. 응용시스템 측면에서는 데이터의 수집, 저장, 분석, 시각화에 이르는 과정이 처리되고 있는지를 점검한다. 데이터 측면에서는 데이터가 표준화가 정의되어 있는지와 데이터가 표준을 잘 준수하고 있는지를 점검한다. 시스템 측면에서는 데이터의 수집, 저장, 분석, 시각화를 위한 구성과 아키텍처를 점검한다. 외부 시스템과 연계를 위한 인터페이스와 접근제어, 보안 등의 설계를 점검한다.

[표 3-2] 빅데이터 설계단계 점검항목

분야	점검항목		분석	구축	비고
응용시스템	01. 데이터 수집·정제 체계의 적정 설계 여부	01. 데이터 수집·정제 체계가 마련 되었는가? • 수집 대상 데이터의 목록의 작성 • 수집 대상 데이터 현황 정의 (보유기관,담당자, 데이터 제공형식 등) • 데이터 수집 및 정제 체계	○	○	• 가이드 v1.2 • '21 빅데이터
		02. 수집 대상 데이터에 대한 적절한 처리방안이 마련되었나? • 데이터에 포함된 개인정보의 유출 문제를 처리하기 위한 방안 수립 여부 • 오류 데이터 제거, 길이 체크, 날짜 포맷 점검 등 데이터 정제 과정 점검	○	○	• 가이드 v1.2 • '21 빅데이터
	02. 데이터 저장 체계의 적정 설계 여부	01. 데이터 추출(Extract), 변환(Transformation), 탑재(Load) 기능이 정상 동작하나? • 데이터 품질관리 절차 및 기준 유무 • 개인정보 탐색· 및 처리방법의 구체화 확인 (비식별화, 암호화 등)	○	○	• 가이드 v1.2 • '21 빅데이터
		02. 데이터 보유기관에서 주기적으로 수집 가능한 연계가 적용되었는가? • 데이터 보유기관별 연계방법 및 주기 확인	○	○	• 가이드 v1.2 • '21 빅데이터
	03. 데이터 분석모델 설계	01. 분석대상 및 범위가 요구사항과 일치하는가? • 요구사항별 분석대상과 범위 지정	○	○	• 가이드 v1.2 • '21 빅데이터
		02. 분석대상별 분석모델이 지정되었는가? • 분석대상별로 분석모델의 지정 여부 확인 (분류모델, 지수모델, 예측모델 등) • 분석모델의 설계 타당성 검증 (도메인별 데이터 분석 전문가의 검증) • 분석 알고리즘 모델의 설계 여부	○	○	• 가이드 v1.2 • '21 인공지능

5) 빅데이터 감리 점검가이드 (2/2)



분야	점검항목		분석	구축	비고
응용시스템	04. 데이터 시각화 설계	01. 분석 결과가 잘 반영되었나? • 분석 화면 구성 설계 • 데이터 분석 시각화 설계 • 서비스를 위한 기능 설계		○	• 가이드 v1.2 • '21 빅데이터
데이터	01. 데이터 표준화 및 융합	01. 수집된 데이터의 속성(칼럼명, 문자형, 숫자형, 크기 및 길이, 필수 여부 등) 정의가 되어 있는가? • 코드 표준화 관련 명명규칙 • 사전관리(단어, 용어, 도메인 등), 통합코드 등 • 표준 항목 정의/적용 및 관리 체계	○	○	• 가이드 v1.2 • '21 빅데이터
		02. 분석에 필요한 데이터 목록과 유형이 정의되고 변환 기준이 정의되었는가? • 분석에 필요한 데이터 목록 • 데이터별 유형 설계 • 데이터 수집 시의 규칙 정의 (유효성 검증방안)	○	○	• 가이드 v1.2 • '21 빅데이터
		03. 수집된 데이터는 구조를 통일하고 병합하는 표준화 과정을 거쳤는가? • 데이터 정제 및 변환 기준 • 데이터 변환 기준 및 변환 절차	○	○	• 가이드 v1.2 • '21 빅데이터
시스템	01. 시스템간 연계 설계	01. 외부데이터 수집을 위한 연계가 명확히 정의되어 있는가? • 데이터 소스 연결 관리 • 데이터 송/수신 연계 • 실시간/배치 연계	○	○	• 가이드 v1.2 • '21 빅데이터
		02. 시스템 구성요소에 대한 상세설계가 충분하게 이루어졌는가? • HW, SW(분리발주 SW 포함) 등 구성 • 시스템 구성요소 검증 방법 및 절차 • 문제 발생 시 복구 및 업무처리 계획		○	• 가이드 v1.2 • '21 빅데이터
		03. 시스템 구성 및 아키텍처에 사용자 요구사항이 적용되고 명세화 되었는가? • 시스템 성능/가용성 • 동시 사용자 접속 처리 용량 • 명세화된 항목별 검증 및 확인 방법		○	• 가이드 v1.2 • '21 빅데이터
		04. 시스템의 보안, 접근 권한 정책이 설계되었는가? • 데이터에 대한 접근/통제 설계 • 시스템간 접근/통제 설계 • 암호화 전송 등		○	• 가이드 v1.2 • '21 빅데이터

[SD32-2-B1]

02 분석대상별 분석모델이 지정되었는가?

- 분석대상별로 분석모델의 지정 여부 확인(분류모델, 지수모델, 예측모델 등)
- 분석모델의 설계 타당성 검증 (도메인별 데이터 분석 전문가의 검증)
  분석 알고리즘 모델의 설계 여부

● **목적**

빅데이터에서 정의된 분석대상을 조사하고 각 분석항목에 분석모델이 지정되어 빅데이터 구축 목표를 달성할 수 있을지를 검토하고, 빅데이터의 적정성을 판단하는 근거로서 확인하는데 목적이 있다.

● **필요성**

빅데이터로 분석하고자 하는 대상이 정의되고, 정의된 분석대상의 결과를 도출해내기 위해 필요한 분석모델이 사전에 정의되어 작업방법과 결과를 예측하여 빅데이터의 구축 목표를 달성할 수 있도록 점검하고 확인하는 과정으로 필요하다. 따라서, 분석 대상별로 사전에 분석결과를 예측하고 분석모델을 지정하는 등의 다각도의 분석을 하였는지 확인하는 것이 필요하다.

● **감리관점/ 점검기준**

절차			산출물								사업성과	
계획 적정성	절차 적정성	준수성	기능성	무결성	편의성	안정성	보안성	효율성	준거성	일관성	실현성	충족성
○	○		○								○	○

CHAPTER

3. AI 관련 감리 참고자료

1) 공공부문 초거대 AI 도입 • 활용 가이드라인2.0 (25.4)





0 가이드라인 목적 및 구성

● 목 적

- 본 가이드라인은 중앙부처·지자체·공공기관·국회·법원 등 전 공공부문에서 디지털플랫폼정부의 철학과 원칙에 따라 초거대 AI를 도입·활용하는데 필요한 관련 기준과 절차, 고려 사항을 안내하는 것을 목적으로 함
- 또한 공공부문의 정보시스템 담당자에게는 초거대 AI 도입과 활용에 필요한 실무적 지침과 공공의 다양한 AI 서비스 사례들을 제공하고자 함

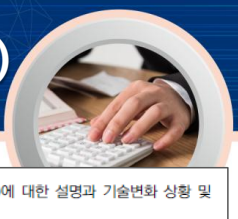
● 구 성

- 본 가이드라인은 초거대 AI 개요, 공공부문 초거대 AI 추진방향과 활용사례, 초거대 AI 도입 절차, 공공부문 AI 성과관리 등 4개의 장과 부록으로 구성함

표 1 초거대 AI 도입·활용 가이드라인 구성

1. 초거대 AI 개요	2. 공공부문 초거대 AI 추진방향과 활용사례	3. 초거대 AI 도입 절차	4. 공공부문 AI 성과관리
1.1 초거대 AI의 개념과 구성 요소	2.1 공공부문 초거대 AI 추진 방향	3.1 도입 원칙 및 고려사항	4.1 성과 관리 필요성
1.2 초거대 AI 발전 경과 및 최근 기술 동향	2.1.1 공공AI 3대 전략 목표	3.2 도입 절차	4.2 AI 성과지표 프레임워크
1.3 국내 초거대 AI 시장 현황	2.1.2 범정부 초거대 AI 공통기반 구현	3.2.1 데이터 보안 등급	4.3 AI 성과지표 Pool
1.4 해외 주요국 AI 정책 동향	2.2 공공부문 초거대 AI 활용 사례	3.2.2 클라우드 서비스 구성 방안	
	2.2.1 초거대 AI 적용 서비스 분류	3.2.3 데이터 학습 방식	
	2.2.2 서비스 유형별 활용 사례	3.2.4 서비스 도입 방식	
	2.2.3 업무 분야별 활용 사례	3.2.5 유지보수 및 운영	
	2.2.4 해외 활용 사례	3.3 초거대 AI 도입 체크리스트	

1) 공공부문 초거대 AI 도입 • 활용 가이드라인2.0 (25.4)



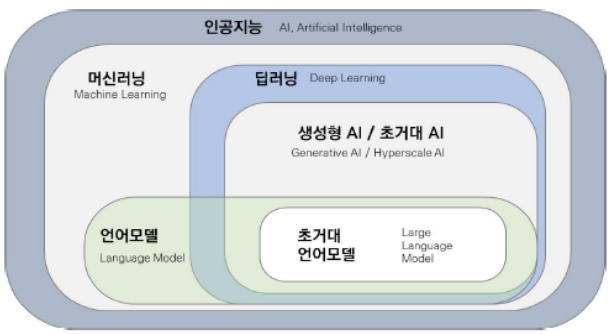
1.1 초거대 AI의 개념과 구성요소

● 초거대 AI

- 초거대 AI는 기존 인공지능(AI)보다 훨씬 방대한 양의 데이터를 이용해 수천억에서 수조 개 이상의 매개변수(Parameter)를 가진 대규모 모델을 의미. 대규모 인공신경망 알고리즘을 활용하며 높은 수준의 추론과 이해 능력을 보여줌
- 특정 작업만 하는 기존 AI와 달리 자연어 처리, 이미지 인식, 코드 작성 등 여러 분야에 폭넓게 활용될 수 있는 범용성이 특징이며, 새로운 작업에 대해 별도 학습 없이 간단한 설명이나 예시만으로도 신속히 대응할 수 있다는 장점이 있음

초거대 언어 모델(LLM, Large Language Model)이란 초거대 AI로 구현된 언어 모델(LM, Language Model)로 기존 언어 모델보다 훨씬 큰 규모의 모델과 데이터를 활용하여 뛰어난 언어 이해력과 생성 능력을 가짐. LLM은 초거대 AI의 하위 개념으로, 초거대 AI에는 LLM뿐만 아니라 이미지 생성, 음성 인식, 로봇 AI 등 다양한 모델이 포함될 수 있음

그림 1 인공지능 종류와 초거대 AI 관계



- 1장 '초거대 AI 개요'에서는 초거대 AI와 언어모델(LLM)에 대한 설명과 기술변화 상황 및 이에 따른 국내외 현황에 대하여 안내함
- 2장 '공공부문 초거대 AI 추진방향과 활용사례'에서는 공공부문에서의 초거대 AI 활용을 통한 3대 전략목표와 범정부 공통기반 구현사항, 공공부문에서 도입 가능한 서비스 활용 사례(예시 등)를 행정·공공기관 및 해외의 주요국에 대해 제시함
- 3장 '초거대 AI 도입 절차'에서는 공공부문에서 실제 도입시 준수해야 할 원칙과 관련 기술 적용에 필요한 검토사항 등을 절차별로 안내함
- 4장 '공공부문 AI 성과관리'에서는 공공부문에서 성과관리 필요성과 방법, 지표 Pool 등을 제시하여 중장기 성과관리 및 환류 체계에 대하여 안내함

관련용어

용 어	정 의
초거대 AI	대규모 데이터셋을 기반으로 훈련된 딥러닝 모델을 사용하여 문서를 요약, 대조하거나 새로운 콘텐츠를 생성하는 등의 인공지능
초거대 언어 모델 (LLM, Large Language Model)	초거대 규모로 자연어를 학습시킨 인공지능 언어 모델
파운데이션 모델 (Foundation Model)	초거대 AI의 서비스를 제공하기 위해 기초가 되는 모델
파인튜닝 (Fine-tuning)	파운데이션 모델을 특정 작업이나 도메인에 최적화하기 위해 특화된 데이터로 모델의 가중치를 미세하게 조정하여 추가로 학습시키는 방법
사후학습 (Post-training)	파운데이션 모델에 데이터의 최신성과 전문성을 위해 자체 보유데이터로 추가 학습하여, 파운데이션 모델 자체를 고도화하는 방법
검색 증강 생성 (RAG, Retrieval-Augmented Generation)	답변 생성에 있어 외부 리소스를 추가하는 방식의 기술
소규모 초거대 언어 모델 (sLLM, Small Large Language Model)	초거대 언어 모델에 비해 상대적으로 적은 파라미터를 사용하여 학습 시간이나 비용을 절감한 모델
멀티모달 (multimodal)	사람이 기계와 상호 작용할 때 입출력에 텍스트, 음향, 이미지 등 다양한 정보 유형을 통합하여 사용하는 것
AI 에이전트 (AI Agent)	사람이 직접 프롬프트를 입력하는 거대언어모델(LLM) 기반 AI 챗봇과 달리 자율성을 바탕으로 스스로 업무를 수행하는 인공지능
클라우드컴퓨팅서비스	클라우드컴퓨팅을 활용하여 상용(商用)으로 타인에게 정보통신자원을 제공하는 서비스
클라우드컴퓨팅서비스 보안인증 (CSAP, Cloud Security Assurance Program)	「클라우드컴퓨팅 발전 및 이용자 보호에 관한 법률」 제23조의2에 따라 한국인터넷진흥원의 장이 클라우드컴퓨팅서비스의 보안성에 대하여 실시하는 인증
민간 클라우드	「클라우드컴퓨팅법」 제2조제3호에 따른 '클라우드컴퓨팅서비스'로 민간 기업 또는 단체가 제공하는 클라우드컴퓨팅서비스
멀티 클라우드	두 개 이상의 독립적인 클라우드 서비스 제공자가 제공하는 복수의 클라우드컴퓨팅서비스를 연계하여 사용하는 클라우드 배포 모델
디지털서비스 이용자원시스템	「클라우드컴퓨팅법 시행령」 제15조의2제3항에 따라 디지털서비스를 등록 및 관리하는 시스템

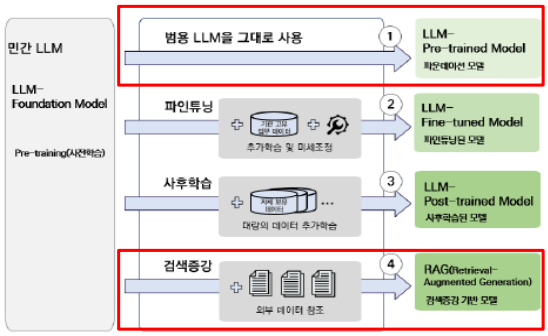
1) 공공부문 초거대 AI 도입 • 활용 가이드라인2.0 (25.4)



3.2.3 데이터 학습 방식

● 데이터 학습 방식에 따른 LLM 유형은 크게 파운데이션 모델, 파인튜닝된 모델, 사후학습된 모델, RAG(Retrieval-Augmented Generation, 검색 증강 생성) 기반 모델로 구분할 수 있음

그림 10 학습방식에 따른 LLM 유형



① 파운데이션 모델
- 추가학습을 하지 않고 범용 LLM을 그대로 사용하는 모델
- GPT, BERT 등 LLM의 사전학습(Pre-trained) 모델로 추론 속도, 리소스 측면의 제약이 없을 때 사용할 수 있음
- 대규모 말뭉치로 사전학습 되어 일반적인 작업에 사용 가능하며, 추가학습없이 다양한 자연어 생성 작업을 할 수 있음
- (장점) API로 연결하면 서비스(텍스트 생성, 번역, 요약, 질의 응답 등)를 바로 사용할 수 있고, 별도의 학습 과정이 필요하지 않아 추가 비용이 필요 없음
- (단점) 특정 용도에 최적화되어 있지 않아 전문성을 갖기에는 성능적 한계가 있음

② 파인튜닝된 모델
- 문장 요약, 분류, 감정 분석 등 구체적인 작업 수행을 위해서 작업과 관련된 데이터로 파인튜닝 작업을 수행한 모델
- 모델의 전문성 확보를 위해 특별히 선별된 데이터로 파라미터의 미세조정을 수행함
- (장점) 특정 데이터를 추가로 학습시켜 사용자의 요구에 최적화된 성능을 기대할 수 있음
- (단점) 파운데이션 모델을 파인튜닝을 하는데 필요한 컴퓨팅 리소스, 인력, 시간이 투입되어야 하므로 추가 비용이 필요함

③ 사후학습된 모델
- 기술, 법률, 특허 등 최신 정보를 활용하는 모델로 사용하기 위해 사후학습을 수행한 모델
- 모델의 최신성과 전문성 확보를 위해 대량의 데이터로 추가 학습하며, 파운데이션 모델에서 학습된 지식을 보존하면서 파운데이션 모델 자체를 고도화하는 방식임
- (장점) 대량의 데이터로 추가 학습된 모델은 보다 전문성 있는 성능을 기대할 수 있음
- (단점) 파운데이션 모델을 대량의 데이터로 추가 학습시키는데 상당한 시간이 필요하고, 파인튜닝보다 많은 컴퓨팅 리소스가 필요함

④ RAG(Retrieval-Augmented Generation, 검색 증강 생성) 기반 모델
- 기존 모델을 추가 학습하지 않고, 보유 데이터 기반 검색을 활용하여 최신 정보 반영
- 사전 학습된 LLM이 정적인 데이터만 활용하는 것과 달리, 기업 및 기관의 내부 데이터베이스, 공공데이터 등을 연계하여 보다 정확하고 신뢰할 수 있는 답변 제공
- (장점) 모델을 직접 업데이트하지 않아도 최신 정보로 업데이트 가능
- (단점) 검색된 정보의 신뢰성이 보장되지 않을 경우, 부정확한 응답 가능성 존재

고려사항

- 파인튜닝된 모델이나 사후학습된 모델은 추가학습을 위해 추가적인 데이터 준비와 학습을 위한 데이터 전처리 과정이 필요
- 추가학습된 모델은 목표한 성능을 갖추었는지 검증과 최신성이 중요한 도메인의 경우 주기적인 모델 업데이트가 필요함
- RAG 기반 모델은 모델 자체를 재학습하지 않으므로, 최신 정보 반영이 가능하나 활용 데이터의 신뢰성을 보장해야 하므로 벡터DB 등 관리가 중요

2) 챗GPT 등 생성형 AI활용 보안 가이드 라인 (2023.6)





국가정보원 NATIONAL INTELLIGENCE SERVICE NSR 국가보안기술연구소

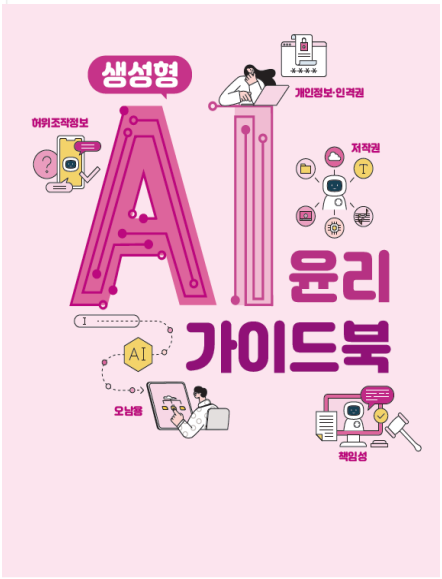
챗GPT 등 생성형 AI 활용 보안 가이드라인

목차

1. 개요 1
1.1. 생성형 인공지능 기술 소개 1
1.2. 생성형 인공지능 기술의 기본 개념 – GPT를 사례로 3
1.3. 해외 주요국 동향 8
2. 생성형 인공지능 기술의 대표적인 보안 위협 10
2.1. 개요 10
2.2. 잘못된 정보 11
2.3. 생성형 AI 모델 악용 13
2.4. 유사 AI 모델 서비스 빙자 15
2.5. 데이터 유출 16
2.6. 플러그인 취약점 19
2.7. 확장 프로그램 취약점 21
2.8. API 취약점 23
3. 안전한 생성형 인공지능 기술 사용 가이드라인 25
3.1. 개요 25
3.2. 서비스 사용 주의사항 26
3.3. 서비스와의 대화 시 주의사항 28
3.4. 서비스 플러그인 사용 주의사항 33
3.5. 서비스 확장 프로그램 사용 주의사항 35
3.6. AI 모델 생성기반 공격 대처 방안 37
4. 생성형 인공지능 기반 정보화사업 구축 방안 및 보안 대책 39
4.1. 개요 39
4.2. 구축 유형에 따른 생성형 AI 기술 도입시 고려사항 40
4.3. AI 모델 API 활용 및 민간 AI 모델 도입 방안 42
4.4. 자체 데이터 세트 및 AI 모델 구축 단계별 보안 위협 대응 방안 47
5. 부록 50
5.1. 약어 51
※ 챗GPT 등 생성형 AI 활용 보안수칙 52

3) 생성형 AI 윤리 가이드북 (2023.12)





방송통신위원회 NIA 한국지능정보사회진흥원

PART 01 생성형 AI란?

생성형 AI가 무엇일까요? 08
생성형 AI가 기존은 변화는 무엇일까요? 08
생성형 AI를 왜 윤리적으로 활용해야 하나요? 10
생성형 AI의 역기능은 무엇인가요? 11
생성형 AI의 윤리적 활용을 위해 어떤 노력이 필요 할까요? 12

PART 02 저작권

챗GPT가 쏘아 올린 생성형 AI 열풍 16
AI 생성물의 저작권은 누구에게? 16
1 생성형 AI를 활용해 새로운 콘텐츠를 만들었어요! 그럼 저는 저작권을 가질 수 없는 건가요? 17
2 생성형 AI가 작성한 거라길래 재가공해도 괜찮을 것 같아서 다운받아 사용했어요. 그런데 콘텐츠 게시자가 저작권 위반이라고 항의를 해왔네요. 생성형 AI가 만든 콘텐츠의 재사용도 법적으로 문제가 되나요? 18
3 이미지(영상) 생성형 AI를 활용해서 연예인과 유명인의 얼굴이 나오는 콘텐츠를 만들었어요. 친구들과 함께 보려고 개인 블로그, SNS에 올릴까 하는데 이무래도 초상권이 문제되겠죠? 20
4 음성 생성형 AI를 활용해 인기 있는 가수의 음성으로 AI 커버곡을 만들었어요. 온라인에 공유하면 문제가 될까요? 22
5 내가 만들어서 온라인에 공유한 콘텐츠(글, 이미지, 음원 등)를 생성형 AI가 무단으로 학습에 활용하고 있음을 발견했어요. 내 저작권을 지킬 수 있을까요? 23
6 생성형 AI에 별도로 출처 표기가 없다라고요. 상업적 이용은 아니고 개인 SNS에 올릴 건데, 그래도 출처를 표기하지 않으면 문제되나요? 25
7 유료로 웹툰이나 웹소설을 결제해서 보다가 생성형 AI로 만든 것처럼 보이는 콘텐츠를 발견했어요. 웹툰, 웹소설 플랫폼에 문의해서 환불받을 수 있을까요? 26

PART 03 책임성

생성형 AI가 만든 작품이 미술대회 1등? 30
AI가 다 했으니 사람은 책임 없다? 30

PART 04 허위조작정보

나도 햇갈리는데 남들은? 42
AI 챗봇으로 보이스 피싱까지? 42

PART 05 개인정보·인격권

생성형 AI는 당신에 지난 여름에 한 일을 알고 있다? 52
유익한 것은 취하고 유해한 것은 피아라! 52

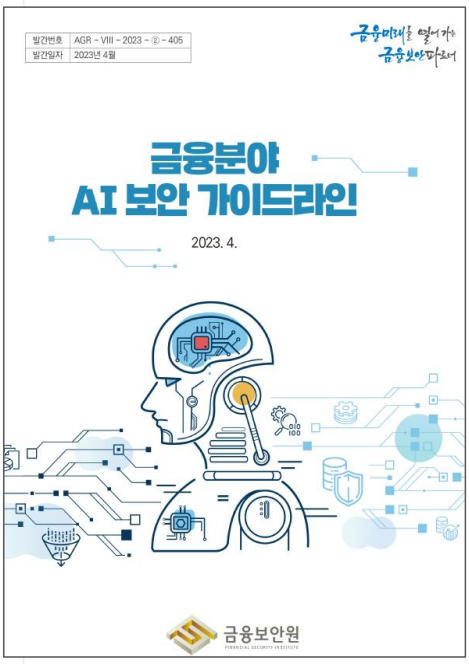
PART 06 오남용

생성형 AI는 나만의 일타 강사? 60
생성형 AI는 도구일뿐! 60
1 생성형 AI를 써보니 결과도 빠르게 알려주고, 검색보다 쉽고 편하게 원하는 정보를 얻을 수 있더라고요. 이제는 아빤 일이든 맨 처음 생성형 AI에 물고 있는데요, 제가 너무 생성형 AI에 의존하는 것일까요? 61
2 생성형 AI와 계속 대화하다 보니 정말 인간처럼 느껴져요. 때로는 친한 친구 같고요. 이런 감정이 문제가 있는 것일까요? 63
3 우주과학 연구에 관심이 많은데, 정보 검색에 시간이 많이 걸려서 생성형 AI에 질문을 해봤어요. 빠르게 답변을 제시해 주던데 이 정보를 그대로 믿어도 되나요? 64
4 생성형 AI가 전문적인 정보까지 알려주는 것 같더라고요. 심리, 의료, 재무 등 전문 분야에 대한 상담도 제대로 해줄까요? 66

별첨 생성형 AI를 현명하게 활용하기 위한 체크리스트 69

4) 금융분야 AI 보안 가이드라인 (2023.4)





목 차

제1장 가이드라인 개요 1
1. 추진 배경 및 목적 2
2. AI 개요 4
3. 가이드라인 구성 5
4. 용어 정리 6

제2장 AI 서비스 구성 7
1. AI 서비스 구성 8
2. 운영계 9
3. 개발계 15

제3장 AI 학습 데이터 및 모델 보안 관리 21
1. 학습 데이터 수집 23
2. 학습 데이터 전처리 24
3. AI 모델 설계·학습 28
4. AI 모델 검증·평가 35

참고
1. 데이터 오염 공격 40
2. 모델 오염 공격 42
3. 모델 추출 공격 43
4. 모델 인버전 공격 45
5. 회피 공격 47

별첨 AI챗봇서비스 보안성 체크리스트 49

별첨 AI챗봇서비스 보안성 체크리스트

분류	소분류	연번	보안성 체크리스트
공통	계정관리	1	챗봇 서비스 사용자별 역할(e.g., 관리자, 이용자)에 따라 계정을 생성하고 관리하는가?
	접근통제	2	챗봇 시스템에 대한 해킹 등을 방지하기 위해 정보보호시스템이 설치 및 운영되고 있는가?
		3	챗봇 시스템 관련 파일(e.g., AI 모델 파일, 소스코드, 학습 데이터 등)에 대한 접근통제 정책 등록/변경/삭제 시 그 내역을 기록하고 정기적으로 적정성 및 이상 유무를 검토하는가?
	보안점검	4	챗봇 서비스 개발 시 보안 가이드라인 등을 참고하여 개발하고 오픈 전 시스템에 대한 보안점검을 실시하는가?
		5	챗봇 시스템에 대해 주기적으로 취약점을 점검하는가?
	입력 제한	6	개인정보를 처리하지 않는 챗봇의 경우 입력창에 개인정보를 입력하지 않도록 이용자에게 사전에 안내하는가?
		7	개인정보를 처리하지 않는 챗봇의 경우 챗봇 서비스 내 사용자 입력 시, 개인정보가 입력되거나 개인정보가 포함되는 파일이 업로드되는 것을 제한하는가?
	중요정보 보안	8	챗봇 서비스 관련 중요정보*에 대해 적절한 보호대책(계정/권한 관리, 접근통제, 암호화 등)이 적용되어 있는가? * 개인정보, AI 모델파일, 학습데이터(학습용, 검증용 등), 챗봇관리시스템, 챗봇인프라 관련 로그, 발췌문 원문 저장 DB 등
		9	챗봇 서비스 내 중요정보* 암호화 시, 안전한 암호화 알고리즘을 사용하여 암호화되고 관리되는가? * 개인정보, AI 모델파일, 학습데이터(학습용, 검증용 등), 챗봇관리시스템, 챗봇인프라 관련 로그, 발췌문 원문 저장 DB 등
		10	암호화에 사용되는 암호키에 대해 적절한 암호키 관리방안(e.g., 안전한 난수발생기를 통한 암호키 생성, 암호키 교체 주기 적용 등)을 수립하고 이행하는가?
	원격접속 금지	11	내부망 이외의 곳*에 위치한 챗봇서비스 관리자 시스템에 원격 접속을 원칙적으로 금지하고 있는가? 원격으로 접속이 불가피한 경우,강화된 보안통제(e.g.,추가인증, 접속IP제한, 중요정보 노출 최소화 등)를 적용하는가? * 클라우드나 DMZ 구간 등

5) AI 데이터 품질관리 가이드 라인 v3.5 (2025.2)





AI 데이터 품질관리 가이드라인 v3.5 요약

AI 데이터 품질관리 가이드라인 v3.5 구성

□ "인공지능 학습용 데이터 품질관리 가이드라인 v3.1"을 기반으로 2024년 초거대AI 확산 생태계 조성사업의 데이터 특성(LLM, LMM, 합성)을 반영하여 고도화한 "AI 데이터 품질관리 가이드라인 v3.5" 구성(2025년)

- "1권. AI 데이터 품질관리 가이드 v3.5 (부제: 품질관리 프레임워크 중심)"
  "2권. AI 데이터 구축 가이드 v3.5 (부제: 구축사례 중심)"
  "3권. 생성형AI 데이터 품질관리 가이드 v2.0 (부제: LLM, LMM, 합성데이터 중심)" 분리 구성

[제1권] AI 데이터 품질관리 가이드

- 인공지능 학습용 데이터 구축 시 구축계획 수립단계부터 데이터 획득/수집, 정제, 가공 등에 대한 절차, 산출물 및 품질관리 활동에 대한 기준 및 절차를 제시한 기준서 제공
- 품질 자가점검 프로세스, 품질관리 기준 및 지표 예시를 제시하여 데이터 구축 기준 및 계획 수립 시 참고자료로 활용할 수 있도록 구성

[제2권] AI 데이터 구축 가이드

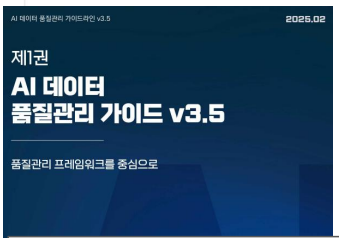
- 2024년 초거대AI 데이터 확산 생태계 조성사업의 데이터 특성(LLM, LMM, 합성)별 데이터 획득/수집, 정제, 가공, 학습 단계의 데이터 구축과정 사례를 수록
- 데이터 특성별 1-Cycle 자가점검 결과를 수록하고 발생 오류에 대한 분석결과를 제공
- 품질검증 결과서를 기반으로 주요 오류 유형 분류 및 구체적인 사례 수록

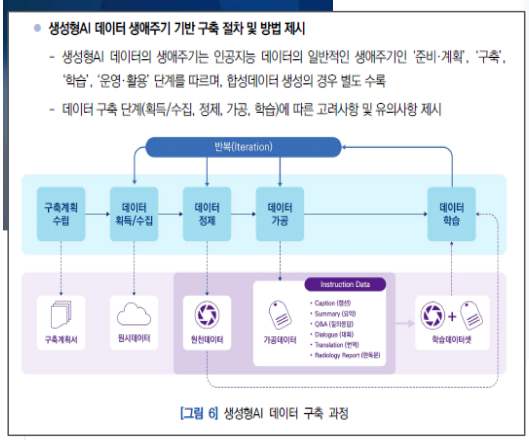
[제3권] 생성형AI 데이터 품질관리 가이드

- 생성형AI 데이터 구축 확대에 따라 데이터 특성(LLM, LMM, 합성)을 반영하여 작성
  ※ 단, 공통 사항 및 정의의 경우 '제1권 AI 데이터 품질관리 가이드'에서 기술
- 양질의 생성형AI 데이터 구축을 위한 기초자료로 활용할 수 있도록 학습유형에 따른 특성, 품질관리 지표와 설정 가이드를 제시한 기준서 제공

5) AI 데이터 품질관리 가이드 라인 v3.5 (2025.2)







[그림 6] 생성형AI 데이터 구축 과정

2.1 데이터 획득/수집(210)

- 인공지능의 학습에 필요한 데이터를 현실 세계에서 직접 획득/수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 목적에 따른 원시데이터를 확보하는 활동
- 인공지능 학습용 데이터의 전체 품질 수준은 구축단계에서 80-90%가 좌우될 만큼 중요
  - 이유 1 : 인공지능 학습용 데이터가 인공지능 학습에 필요한 데이터 요구사항을 기반으로 구축되는 목적성을 갖기 때문에 데이터 품질 수준이 중요
  - 이유 2 : 한번 구축된 인공지능 학습용 데이터는 자체적인 품질 이슈가 발견되지 않는 한 인간에 데이터가 그대로 개방되어 배포됨에 따라 품질 수준 중요
- 이 단계에서는 정제 및 가공단계에서의 요구사항을 우선 파악 후 데이터 가공의 목적에 맞는 데이터 환경 구축 중요

2.1.2 데이터 획득/수집(210) 프로세스 품질관리 활동 및 산출물

프로세스	품질관리 활동	주요 검토 내용	구축 및 품질 산출물
210. 데이터 획득/수집	2101. 획득/수집 방법 및 기준 현행화	• 데이터 획득 및 확보 방법, 절차, 법 제도적 적절성 • 획득항목 및 구축량 획득/수집 단계 품질기준 및 조직, 절차, 도구 • 데이터 획득 방법 및 기준, 교육, 검사에 대한 체계 수립 • 데이터 획득 기준 변경 관리	구축공정 가이드라인 (획득/수집)
	2102. 데이터 법적 근거 검토	• 법 제도 준수(개인정보, 초상권, 저작권) 근거자료 확보 • 데이터 획득/수집의 공평성 위배 방지 방안	개인정보 자율점검표
	2103. 수집 도구 및 저장환경 검토	• 획득/수집에 사용될 도구 확보 및 사용 방법 • 획득/수집 데이터 저장환경 구축	수집 도구 수집 도구 사용설명서
	2104. 원시데이터 품질검사	• 원시데이터가 인공지능을 활용하여 문제를 해결하는데 요구되는 분류 다양성, 신뢰성, 충분성, 균일성, 사실성, 공평성, 형식성 등 기준에 적합하게 구성되어 있는지 확인 • 데이터에 대한 클래스 분포, 인스턴스 분포, 분석 가능 항목의 통계적 분포 확인	원시데이터 품질관리보고서 품질검사 도구

〈표 II-12〉 데이터 획득/수집 단계 품질관리 활동 및 산출물

- 원시데이터를 수집하거나 직접 원시데이터를 생성하는 경우, 검사 항목을 고려해서 원시데이터의 품질을 확보해야 함

검사 항목	내 용
분류 다양성	• 학습 목적에 부합하도록 실제 세상의 데이터와 유사한 특성을 가진 데이터 확보해야 함 – 사물, 사람, 장소, 시간, 환경, 언어 특성 등 학습에 유용한 모든 특성 정보를 포함할 수 있도록 고려 – 사물, 사람, 장소, 시간, 환경, 언어 특성 등의 특성 정보가 학습에 유용한 범위에서 다양하게 수집
신뢰성	• 데이터는 반드시 신뢰할 수 있는 출처로부터 수집해야 함
충분성	• 데이터에 포함된 카테고리와 인스턴스 등 특성 정보는 학습에 유용한 수량이어야 함
균일성	• 분류/탐지/인식/이해/예측 등의 카테고리별 인스턴스 수량의 균일성과 적정 비율을 확보해야 함
사실성	• 원시데이터를 인위적인 환경과 조건하에 수집해야 하는 경우, 반드시 실제 환경과 상황 특성을 반영해야 함
공평성	• 원시데이터는 지역, 사회 및 인종적 편견 등 활용 의도와 무관한 편향적인 특성이 제거되고, 윤리적으로 공평해야 함

〈표 II-15〉 원시데이터의 검사 항목

2.2 데이터 정제(220)

- 획득한 원시데이터를 학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동
- 원시데이터 획득 시 정제기준을 명확하게 정의하고, 구축목적에 따라 적절한 데이터 선별을 위한 명확한 정제기준을 수립하며, 기준 미달 또는 활용 불가능한 데이터의 효과적인 제거 방법 수립

2.2.2 데이터 정제(220) 프로세스 품질관리 활동 및 산출물

프로세스	품질관리 활동	주요 검토 내용	구축 및 품질 산출물
220. 데이터 정제	2201. 정제방법 및 기준 현행화	• 구축목적에 적절한 데이터 선별을 위한 정제기준 수립 • 정제단계 품질기준 및 조직, 절차, 도구 • 데이터 정제방법 및 정제 결과에 대한 교육 및 훈련 • 데이터 정제 결과에 대한 검사절차와 기준 수립 • 데이터 정제기준 변경 관리	구축공정 가이드라인 (정제)
	2202. 개인정보/민감정보 비식별화 등 법적 준거 확보	• 개인정보보호 등 비식별화 기준 및 방안 등 법적 준거 확보	
	2203. 정제 도구 및 저장환경 검토	• 정제단계에 사용될 저작도구 확보 및 사용 방법 • 정제 데이터 저장환경 구축	저작도구(정제) 사용설명서
	2204. 원천데이터 품질검사	• 사용하기 적합한 데이터 여부 품질검사 • 원천데이터 품질검사를 위한 검사 도구 및 검사 결과 • 품질검사 결과에 따른 보완조치	원천데이터 품질관리보고서 품질검사 도구

〈표 II-16〉 데이터 정제단계 품질관리 활동 및 산출물

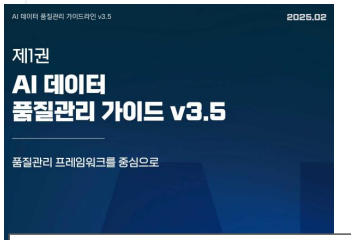
- 정제단계 품질관리를 위한 고려사항

항 목	고려사항
정제기준의 명확성	• 데이터 사용 목적에 적합한 정제기준 수립 여부 검사
데이터 중복성 방지	• 데이터 정제 후 정보 비교 후 중복도 여부 검사
정제 가이드라인	• 정제 작업을 위한 정제 가이드라인 작성 및 관리 여부 검사
정제 도구	• 정제 작업에 사용될 소프트웨어(S/W) 도구 확보 및 사용 방법 숙지
정제 작업 방식	• 데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부 검사

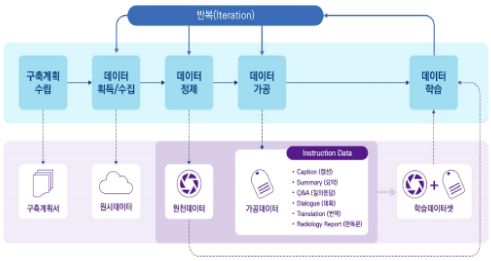
〈표 II-18〉 데이터 정제단계 고려사항

5) AI 데이터 품질관리 가이드 라인 v3.5 (2025.2)





- **생성형AI 데이터 생애주기 기반 구축 절차 및 방법 제시**
  - 생성형AI 데이터의 생애주기는 인공지능 데이터의 일반적인 생애주기인 '준비·계획', '구축', '학습', '운영·활용' 단계를 따르며, 합성데이터 생성의 경우 별도 수록
  - 데이터 구축 단계(획득/수집, 정제, 가공, 학습)에 따른 고려사항 및 유의사항 제시



[그림 6] 생성형AI 데이터 구축 과정

2.3 데이터 가공(230)

- AI모델의 기능이나 목적에 맞게 원천데이터에 적절한 라벨을 추가하는 과정
- 어노테이션 정보는 데이터 확장을 고려하여 항목명, 타입, 필수 구분, 항목 설명을 포함해야 하며 라벨링 용어 및 분류체계는 표준을 준수해야 함
- 가공 방법에 따른 도구 사용 및 작업 방식에 의한 모든 산출 데이터는 CSV, JSON 등 표준화된 포맷으로 제공하며 이에 대한 품질검사 진행

2.3.2 데이터 가공(230) 프로세스 품질관리 활동 및 산출물

프로세스	품질관리 활동	주요 검토 내용	구축 및 품질 산출물
230. 데이터 가공	2301. 가공 방법 및 기준 현행화	• 가공 방법 및 기준 정의 • 학습에 필요한 가공작업 및 어노테이션 설정 • 데이터 형식 및 데이터값 범위 기준 마련 • 데이터 유형별 객체 구분 작업 • 데이터 가공 방법 및 기준, 교육, 검사체계 수립 • 데이터 가공 기준 변경 관리	구축공정 가이드라인 (가공)
	2302. 가공 도구 및 저장환경 검토	• 가공에 사용될 저작도구 확보 및 사용 방법 • 데이터 가공 기준 변경 관리 • 가공데이터 저장환경 구축	저작도구(가공) 사용설명서
	2303. 가공데이터 품질검사	• 정확성(의미 정확성, 구문 정확성)을 중점으로 검사 • 가공데이터 품질검사를 위한 검사 도구 및 검사 결과 • 품질검사 결과에 따른 보완 조치	가공데이터 품질관리보고서 품질검사 도구

〈표 II-19〉 데이터 가공단계 품질관리 활동 및 산출물

2.3.4 고려사항

- 가공 단계의 품질관리를 위해서는 가공작업을 위한 매뉴얼을 확보하고, 가공 도구를 확보하여, 해당 데이터의 특성을 고려해서 가공작업 방식을 정해야 함
- 데이터 가공작업에 필요한 규격은 다음의 사항을 반드시 포함하도록 고려해야 함

항 목	고려사항
가공 정의	• 가공 유형, 가공 구성 항목, 속성값, 타입 등을 정의해야 함
구축공정 가이드라인 (가공)	• 목적에 맞게 작성된 구축공정 가이드라인(가공)에 대한 타당성 고려 • 참값(Ground Truth) 라벨링, 가공 오류, 모호한 가공 처리, 예외처리 가공작업 등을 사전에 정의하여 통과 및 반려 기준을 마련해야 함 • 목적에 맞는 어노테이션 항목 구성 여부 검사 후 작업자 전달 • 작업자가 효율적으로 작업할 수 있는 환경 구성
가공 도구	• 가공작업에 사용하는 플랫폼 및 SW 도구의 기능 및 사용 방법을 기술해야 하며 플랫폼 또는 SW 도구 선택 시 데이터 보안 준수 여부를 확인해야 함

〈표 II-21〉 데이터 가공단계 고려사항

2.4 데이터 학습(240)

- 데이터 학습 단계에서는 원천데이터와 가공데이터의 묶음인 '학습 데이터셋'을 이용하여 사전에 정의된 인공지능 모델을 학습시키고, 학습된 AI모델의 성능을 향상시키거나 보정하는 활동 수행

2.4.1 데이터 학습(240) 프로세스



[그림 II-14] 데이터 학습 프로세스

2.4.2 데이터 학습(240) 프로세스 품질관리 활동 및 산출물

프로세스	품질관리 활동	주요 검토 내용	구축 및 품질 산출물
240. 데이터 학습	2401. 구축목적-AI모델 합치성 확인	• 데이터를 인공지능 모델로 학습 시, 영역별 품질검사 기준 확인 • 구축목적에 부합하는 AI모델 선정 확인 • AI모델 구축/운영 및 테스트를 위한 환경 구축 정보 • AI모델 정보/라이센스 가이드	AI모델 설명서
	2402. 학습 결과 확인 및 최적화	• 학습용 데이터가 AI모델로 학습시켰을 때 목표로 했던 수준의 성능 달성이 가능한지 검사 • AI모델(알고리즘) 성능 검사 • 모델 단계 유효성 검사 • 학습모델, 활용 서비스 개발에 가능한 학습 데이터셋 품질오류 및 이슈 내용을 분석하여 보완조치 반영	AI모델 소스코드 학습모델 파일 모델 구축 도커 이미지 AI모델 테스트 결과서 품질검사 도구
	2403. 품질검증결과 보완조치	• 최종데이터 제3자 품질검증결과가 미달성인 경우, 보완조치하여 결과 보고 ※ 제3자 품질검증 → 제3자 품질검증 세부 활동 정의 → 최종품질검증 프로세스 부분 내용 참고	보완조치결과서

〈표 II-23〉 데이터 학습 단계 품질관리 활동 및 산출물

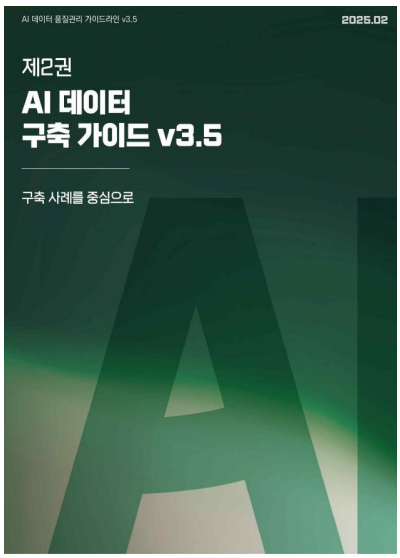
2.4.4 고려사항

항 목	고려사항
데이터 학습 유효성	• '학습 데이터셋'을 이용하여 사전에 정의된 인공지능 모델을 학습시키고, 학습된 AI모델의 성능을 향상시키거나 보정하는 검사를 수행 • 구축한 데이터셋을 인공지능 학습모델로 학습시킨 후 목표로 했던 수준의 성능 달성이 가능한지 확인 • 학습모델 평가를 위한 정확도, 정밀도, 재현율 등 성능지표 확인
AI모델 성능	• 인공지능 모델 성능 검사
품질오류 분석	• 학습모델, 활용 서비스 개발에 가능한 학습 데이터셋 품질오류 및 이슈 내용을 분석

〈표 II-25〉 데이터 학습 단계 시 고려사항

5) AI 데이터 품질관리 가이드 라인 v3.5 (2025.2)





6.2.3 학습모델 품질지표 선정

- Task1 질의응답 (서술형)
  - BERT Score

구분	내 용	지표 설정 근거
측정지표	BERT Score	• 모델 요약 태스크 사례의 지표별 최대 성능 대비 15% 낮게 설정 (BERT Score: 0.82, ROUGE: 0.22) ※ 실행 로그에는 ROUGE 값도 출력되도록 하여 유효성 증빙 자료에 표시 예정 [http://kr-it.com/xml/35065/35065.pdf]
지표설명	생성 문장과 정답 문장 간의 내재하고 있는 의미와 동사구조의 다양성 평가	
계산식	$R_{BERT}=\frac{1}{\lvert x\rvert}\sum_{x_i\in x}\max_{\hat{x}_j\in\hat{x}}x_i^T\hat{x}_j$, $P_{BERT}=\frac{1}{\lvert\hat{x}\rvert}\sum_{\hat{x}_j\in\hat{x}}\max_{x_i\in x}x_i^T\hat{x}_j$, $F_{BERT}=2\frac{P_{BERT}\cdot R_{BERT}}{P_{BERT}+R_{BERT}}$ ※ $x=<x_1,\ldots,x_k>$: 정답 문장 토큰 ※ $\hat{x}=<\hat{x}_1,\ldots,\hat{x}_l>$: 모델 추론 문장 토큰, ※ F_{BERT}: 최종 BERT Score 수치	
정량목표	0.7 이상	

- Task1 질의응답 (객관식)
  - 정확도

구 분	내 용	지표 설정 근거
측정지표	정확도(ACC)	• 유사사례) KMMLU 한국어 객관식 데이터셋 - 45개의 주제에 대한 객관식 (다지선다) 문제를 통한 평가 - 총 35,030개의 객관식 문제 - 객관식 총 문항에 대해 올바르게 추론한 비율 • 평가지표 - 정확도 ACC(0.4 이상) • KMMLU 평가 순위의 성능에서 3위(Gemini-pro)의 0.42를 고려
지표설명	• 객관식 문제에 대한 질의응답 성능을 평가하는데 사용되는 지표로 전체 예측 중 정답의 비율	
계산식	정확도 = $\frac{Correct\ predictions}{All\ predictions}$	
정량목표	0.40 이상	

- Task2: 텍스트 분류
  - F1-score

구 분	내 용	지표설정근거
측정지표	• F1-score	• "행정문서 대상 기계독해 데이터" 활용 AI모델의 평가 기준인 0.78(F1-score)을 참조하여 설정
지표설명	• precision (객관식) 모델이 참값이라고 예측한 데이터 중 정답이 참값인 데이터의 비율 (주관식) 생성 문장의 단어가 정답 문장에 포함 정도 • recall (객관식) 실제로 참값인 데이터를 모델이 참값이라고 예측한 데이터 비율 (주관식) 정답 문장의 단어가 생성 문장에 포함 정도	
계산식	$F_1=\frac{2}{\mathrm{recall}^{-1}+\mathrm{precision}^{-1}}=2\frac{\mathrm{precision}\cdot\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$ ※ precision: 모델이 참값이라고 예측한 데이터 중 정답이 참값인 데이터의 비율 ※ recall: 실제로 참값인 데이터를 모델이 참값이라고 예측한 데이터 비율	
정량목표	• 0.83 이상	

6.2.4 학습모델 선정과 근거

- 학습모델 선정: Llama-3 70B
- 선정 사유: Task1(질의응답), Task2(텍스트 분류)

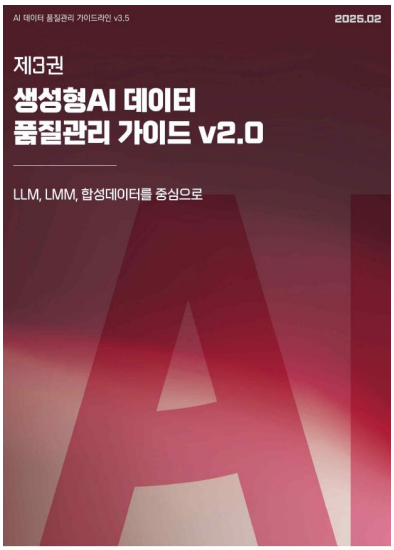
Task	모델 후보	선정 사유
질의 응답	LLaMA [70B]	• 높은 일반화 성능 - 다양한 분야 사전학습으로 질의응답 유형에 관계없이 우수한 성능 기대 • 상용 가능 라이센스 - 활용 제약 없음
	maywell_kiqu-1 [70B] (Alpaca 기반 한국어 튜닝)	• 특정 리더보드 1위이나 양자화 시 성능 저하 가능성 및 일반적인 활용성 고려 시 LLaMA가 더 적합
텍스트 분류	LLaMA [70B]	• 기본 성능 및 다양한 분야 사전학습으로 인한 높은 일반화 성능 기대 • 상용 가능한 라이센스로 활용 용이
	EEVE-Korean [10B] (Alpaca 기반 한국어 튜닝)	• 특정 분야(추론+독해, 여행 관련)에 특화, LLaMA의 더 큰 규모와 일반화 성능 고려 시 LLaMA가 더 적합

- 근거: 유효성 점검 결과

점검내용	학습용 데이터의 유효성					
점검항목	• 유효성 검사 항목 - 데이터셋의 활용목적에 맞는 구체적인 학습 Task(모델의 탐지, 추론, 예측 등) 설정 여부 학습 Task에 부합하는 AI모델 선정 및 개별 AI모델 구현 - AI모델에 따른 성능지표 설정 및 학습 성능 측정					
	구분	TASK 명	모델명	지표		
	LLM	질의응답 (객관식)	LLaMa 7B	Accuracy		
		질의응답 (주관식)	Mistral 7B	리커트 척도		
점검결과	• 유효성 점검 결과 - BERT Score, ACC: 목표로 하는 성능 목표치를 충족함 - F1- Score: 목표로 하는 성능 목표치를 충족함					
	품질특성	TASK 명	모델명	지표	목표치	결과값
	유효성	서술형 질의응답 정확도	Llama-3 70B	BERT Score	0.70 이상	0.85
		객관식 질의응답 정확도	Llama-3 70B	정확도	0.40 이상	0.83
		텍스트 분류 정확도	Llama-3 70B	F1-점수	0.83 이상	0.84

5) AI 데이터 품질관리 가이드 라인 v3.5 (2025.2)





과학기술정보통신부 NIA 한국지능정보사회진흥원 TTA 한국정보통신기술협회

5.2.3.1 거대 언어모델(LLM) 유효성

- LLM(Large Language Model)은 자연어처리 분야에서 사용되는 모델로, 대규모의 텍스트 데이터를 학습하여 언어 이해와 생성 능력을 갖춘 AI모델을 의미함
  - 많은 양의 데이터를 기반으로 학습되기 때문에 다양한 주제와 문맥에서 자연스러운 언어 생성이 가능하며, 이는 대화 시스템, 기계번역, 요약, 질의응답 등 다양한 자연어처리 작업에 유용하게 사용되고 있음
  - 초거대 규모로 자연어를 학습시킨 인공지능 언어모델로, 학습에 필요한 매개변수(파라미터)의 규모가 수십억(1Billion, 10억 개) 정도
  - 대규모 텍스트 데이터로 사전 훈련되지만, 실제 작업에 특화되지 않아서 특정 도메인의 언어 이해가 부족하거나 새로운 데이터에 대해서 잘못된 정보를 제공하는 등 일반화 성능이 낮고, 편향적 결과를 생성하는 문제가 발생할 수 있음

※ 제시한 정량목표범주(예시)는 과제 특성에 따라 다르게 적용될 수 있으므로 참고용으로 활용

Task 분야	측정지표	학습모델 (알고리즘 예시)	정량목표범주 (예시)
질의응답 (Question Answering)	ACC	LLaMA2-7B (JCommonSenseQA)	55.52% 이상
	BLEU/ROUGE	polyglot-ko-1.3b	추가 사전학습모델의 성능 기존 사전학습모델의 성능 비교
	F1 점수	BERT(KoBERT, KcBERT), RoBERTa, ELECTRA	0.7
		GPT-2	0.73
	ROUGE-L	Polyglot-Ko, LLaMA	60% 이상
	평가순위	LLaMA, BLOOMZ	3위 이상
	Rouge-1	LLaMA3	43.1 이상
		openchat-3.5-0106 (mistral)	50% 이상
	Accuracy	Llama-3 70B	0.4 이상
텍스트 분류 (Text Classification)	F1-Score	Llama-3 70B	0.83% 이상
	Recall@1	openchat-3.5-0106(mistral)	50% 이상
		gemma-2-9b-it	50% 이상
		Meta-Llama-3-8B-Instruct	50% 이상
	Rouge-1/Rouge-L	openchat-3.5-0106(mistral)	50% 이상
		gemma-2-9b-it	50% 이상
		Meta-Llama-3-8B-Instruct	50% 이상
텍스트 질의응답 (Text Question Answering)	Accuracy	HCX-003	62% 이상
	Likert 척도 5점	HCX-003	평균 3.5 이상
	context precision	LLaMa 7B	55 이상
텍스트 생성 (Text Generation)	Likert 척도	LLaMA-13B(PPO)	학습 전후 33% 이상 증가
	perplexity	LLaMA-13B	학습 전후 10% 이상 감소
		LLaMa3.1 8B	학습 전후 10% 이상 감소
		LLaMa 7B	학습 전후 10% 이상 감소
	PPO 모델 자연어 추론	LLaMA-13B(PPO)	76.32 이상
	SSA	GPT-NeoX 3.8b, LLaMa	60% 이상
	F1-Score	Klue roberta Large	0.7 이상
	kappa	Klue roberta Large	0.791 이상
	ROUGE-L	KoLLaVa 7B	0.6
	G-Eval	davidkim205/ komt-mistral-7b-instruct	3.5 이상
기계 번역 (Machine Translation)	BLEU*100	Meta-Llama-3.1-8B-Instruct	8.19 이상
텍스트 자동완성 (Text Autocompletion)	Accuracy	Llama3.1	80% 이상
	ROUGE-1	Llama3.1	0.5 이상

5.2.3.3 이미지(합성데이터) 유효성

- 합성데이터(Synthetic data)란 실제데이터(Real data)와 통계적 특성이 유사하여, 실제데이터 분석결과와 유사한 결과를 얻을 수 있도록 새롭게 생성해낸 가상의 데이터
  - 개인정보보호 등의 사유로 실제 데이터의 활용이 어렵거나, 활용할 수 있는 데이터의 양이 부족한 경우에 주로 활용되며, 실제데이터와 유사할수록 활용
  - 합성데이터는 실제데이터와 같은 분포를 가지고 있기 때문에 실제 자료와 비슷한 결과를 얻을 수 있어 AI모델 학습 시 사용
  - AI모델 고도화를 위한 실제 데이터가 부족한 경우, 합성데이터를 통해 가치 있는 학습 데이터셋을 대량으로 구축 가능

- 데이터를 합성한 정도에 따라 완전 합성데이터, 부분 합성데이터로 분류

분류	설명
완전 합성데이터 (Fully synthetic data)	• 공개하려고 하는 데이터에 측정된 실제 데이터가 하나도 없이 모두 가상으로 생성된 데이터
부분 합성데이터 (Partially synthetic data)	• 공개하려는 변수 중 일부만을 선택하여 합성데이터로 대체한 데이터

〈표 Ⅳ-54〉 이미지(합성데이터) 분류

※ 제시한 정량목표범주(예시)는 과제 특성에 따라 다르게 적용될 수 있으므로 참고용으로 활용

Task 분야	측정지표	학습모델 (알고리즘 예시)	정량목표범주 (예시)
이미지 생성 (Image Generation)	FID	Denoising Diffusion Probabilistic Model	87 이하
		Stable Diffusion XL	70 이하
		ViT Adapter	50 이하
	Inception Score	Denoising Diffusion Probabilistic Model	10 이상
	mAP	ViT Adapter	0.1 이하
	민감도	ViT Adapter	0.5 이상
	정확도	SwinViT	0.1 이하
이미지 분류 (Image Classification)	F1-Score	efficientNet V2 XL	85% 이상
	F1-Score	Segmenter	80% 이상
	Top-1 Accuracy	ResNet EfficientNet	65% 이상
	Top-3 Accuracy	ResNet EfficientNet	80% 이상
	AUROC	DenseNet EfficientNet	0.83
	Accuracy	VisionTransformer	80 이상
	AUC-PR(Area under the precision-Recall Curve)	Vision Transformer	0.8 이상
	Precision/Recall	Vision Transformer	0.8 이상
이미지 분할 (Image Segmentation)	DSC (Dice Similarity Coefficie)	HoverNet	80% 이상
이미지 탐지(검출) (Image Detection)	mAP	Faster R-CNN	0.7 이상
	IoU	DeepLabv3+ U-Net	0.7 이상
	Sensitivity	Faster R-CNN	80%
이미지 합성 (Image Synthesis)	FID	YOLOv8	70 이하
	mAP	YOLOv8	0.25 이하
	민감도	YOLOv8	0.6 이하
	PSNR	Stable Diffusion-v2 Inpainting	25 이상
	PSNR	SytleGAN2 StarGAN2	25 이상
	FID	SytleGAN2 StarGAN2	100 이하
	blue@1	MaxVit-LSTM	80% 이상
	BLEU-1	MaxVit-LSTM	80% 이상
	Accuracy	ViT ConvNeXt ResNet	85% 이상
이미지 캡셔닝 (Image Captioning)	CIDEr	ExpansionNet-v2	120

〈표 Ⅳ-55〉 이미지(합성데이터) Task 분야별 유효성 측정지표

CHAPTER

4. 정보시스템 감리와 AI 감리의 차이점

1) 기존 감리와 AI 감리를 위한 차별화된 감리 접근 필요



정보시스템 감리와 AI 감리의 차이점은?

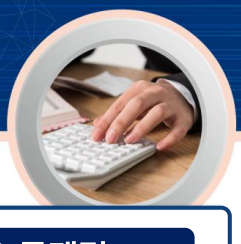
“정보시스템 감리는 “명세대로 잘 만들었는가?” 결과의 예측 가능, AI감리는 “확률 기반 추론” 결과의 불확실성 존재”

기존 정보시스템 개발 vs AI 모델 개발 방법론 차이

항목	기존 정보시스템 개발	AI 모델 개발
개발목적	명확한 기능 구현(예: 결재 기능, 조회 기능 등)	데이터 기반 예측 또는 분류 모델 생성
요구사항 정의	사용자 요구사항을 기능 단위로 정의	문제정의 + 데이터를 통한 해결 방식 도출
설계 방식	화면 설계, DB, 시스템 구조 등 상세 명세 기반	알고리즘 설계 + 데이터 구조 중심
개발 방식	코드 작성과 UI 구현 중심	모델 학습 및 성능 튜닝 중심
테스트 방식	기능 단위 테스트, 통합 테스트 등 명확	정밀도, 재현율 등 통계적 성능 평가 필요
변경관리	**명세 변경 시 제한적 변경 허용**	**반복적 실험과 모델 개선(Agile)**
주요 방법론	**폭포수 모델, 구조적/정보공학적 모델, CBD 등**	**CRISP-DM, Agile, ML Ops 등**
결과물	기능이 명확히 작동하는 시스템	일정 수준이상 성능을 보장하는 AI 모델

• 출처: chatgpt.com

2) 생성형 AI 프로젝트 주요 현황 및 문제점



생성형 AI 도입 주요 현황

도입 확산	• 공공기관, 기업, 교육기관 등 다양한 분야에서 생성형 AI 도입이 급증하고 있음 • 주로 PoC(개념 증명) 형태로 시작
활용 영역	• 고객 응대 챗봇, 보고서 초안 작성, 문서 요약/변역, 마케팅 콘텐츠 제작, 교육 콘텐츠 보조 등 다양한 영역에서 활용 시도
도입 방식	• OpenAI(GPT-4), 네이버(하이퍼클로바X), Google(Gemini), Meta(LLaMA) 등 다양한 모델을 API 또는 오픈 소스로 활용
기대 효과	• 생산성 향상, 시간 단축, 창의적 업무 지원 등 긍정적 기대감은 높음, 그러나 실제 성과는 뚜렷하지 않음

생성형 AI 프로젝트의 주요 문제점

방법론 부재	☑ 생성형 AI는 기존 시스템 개발과는 전혀 다른 특성을 가짐 (명확한 입/출력 구조가 아닌 확률 기반 생성) ☑ Prompt Engineering, RAG(Retrieval-Augmented Generation, Fine-Tunning, Evaluation 등 특화된 접근 방식이 필요하나, 대부분 경험과 감에 의존
절차 부족	☑ 데이터 정제, 프롬프트 검증, 사용자 피드백 반영, 보안•윤리 검토 등의 단계가 빠지거나 임시로 대응 ☑ 검증되지 않은 결과물이 모델에 반영되어 신뢰성과 일관성 부족
산출물 부족	☑ 프로젝트 결과물이 대부분 내부 문서 또는 코드로 남아 재사용성과 확장성 미흡 ☑ 생성형 AI 프로젝트의 핵심인 Prompt 설계 기록, 실패 사례, 사용자 피드백 등이 체계적으로 정리되지 않음

• 출처: chatgpt.com

3) AI 감리를 위한 점검 항목 필요



AI 감리를 위한 라이프 사이클별 점검 항목 이란?

“AI는 데이터 기반 학습과 반복적 실험을 통해 결과를 도출합니다. 감리 시 AI 개발 라이프 사이클별 고유한 점검항목 필요”

CRISP-DM 기반 라이프 사이클별 점검 항목 (예시)

Step 1 문제 정의 (Business Understanding)
- AI 적용 목적과 문제 정의가 명확한가?
- 비즈니스 문제와 데이터 과제가 잘연계 되어 있는가?
- AI 사용의 필요성과 타당성 검토가 이루어졌는가?

Step 2 데이터 이해 (Data Understanding)
- 수집 데이터 출처, 정합성, 품질 확인
- 데이터 편향(Bias) 이나, 왜곡 가능성 여부 점검
- 개인정보 포함 여부 및 비식별화 여부 확인

Step 3 데이터 준비 (Data Preparation)
- 결측치, 이상치, 중복 등 정제 절차의 적법성
- 학습용, 테스트용, 검증용 분리 기준과 비율의 적정성
- 전처리 과정(정규화, 인코딩)의 적정성

Step 4 모델링 (Modeling)
- 알고리즘 선택의 이유와 타당성 검토
- 하이퍼파라미터 설정 근거와 실험 반복성 여부
- 모델 과적합 여부 확인(검증 데이터 사용 여부)

Step 5 평가 (Evaluation)
- 평가지표(정확도, 정밀도, F1 Score 등) 선정의 적절성
- 예측 결과의 설명 가능성 (Explainability) 확보
- 모델 공정성(Fairness), 차별 여부 점검

Step 6 배포 및 운영 (Deployment)
- 운영 중 지속적 성능 모니터링 체계 구축 여부
- 재학습 및 모델 업데이트 정책 수립 여부
- 사용자에게 AI 판단 결과 설명 제공 여부

Step 7 윤리 · 법적 점검 (전주기)
- AI 윤리 원칙(책임성, 공정성, 투명성 등) 반영 여부
- 개인정보보호법, AI 관련 가이드라인 준수 여부
- 데이터 처리 및 결과에 대한 책임 소재 명확화

• 출처: chatgpt.com

4) 생성형 AI 핵심 접근 사항



생성형 AI 라이프 사이클별 핵심 사항 (예시)

Step 1 서비스 식별
- 조직내 어떤 문제를 AI로 해결할 것인가?
  예) 문서 요약, 고객 응대 자동화, 내부 지식 질의 응답 등

Step 2 데이터 식별
- AI가 학습/활용할 수 있도록 조직 내 주요문서, 콘텐츠, 지식 정보 등 선별
  예) 고객 매뉴얼, 업무 문서, 보고서, 이메일, 로그

Step 3 데이터 정제
- 문서 형식 통일, 오류 제거, 민감 정보 마스킹
  불필요한 콘텐츠 제거 -> 품질 높은 데이터 확보

Step 4 정제 된 데이터 활용 및 AI 구축 방식 결정

항목	Fine-tunning	RAG
방식	모델의 가중치 자체를 재학습	검색 -> 응답 생성 구조 (검색 + 생성)
데이터 요구	레이블된 학습 데이터 필수 (대량 필요)	정제된 문서
학습비용	높음 (GPU 필요, 반복 학습)	비교적 낮음 (벡터 DB만 구축)
유지보수	데이터 변경 시 재학습 필요	문서만 바꾸면 자동 반영 가능
개발기간	길고 복잡함	빠르고 프로토타입 가능
설명력	설명 어려움 (블랙박스)	검색 결과와 출처 제시 가능
활용 예	특정 포맷 응답, 고정 업무 대화	지식 질의 응답, FAQ 자동화 등
확장성	업그레이드 어려움	문서 추가로 유연하게 확장 가능

최근 추세

RAG 선호 증가 : 비용 절감, 유지보수 용이, 빠른 서비스화

대부분의 조직이 **RAG 중심**으로 시작하고,
특정 영역에서만 **Fine-tunning을 병행**하는 방향으로 진행

CHAPTER

5. (생성형) AI 감리 수행 사례

1) (생성형) AI 감리 수행 경험



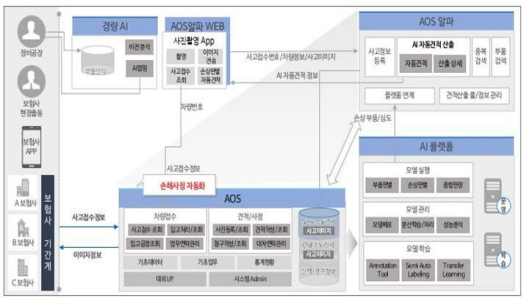
인공지능 기반 통계분류 자동화 시스템 구축 (22~23)

인공지능 통계 분류 모델 (코드 분류)

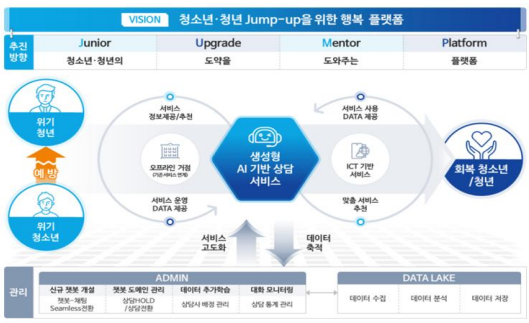
	기존 통계분류 시스템	인공지능 기반 통계분류 시스템
텍스트 전처리	없거나 별도의 전처리 모듈을 적용 (띄어쓰기, 오탈자 등에 대한 전처리 어려움)	최신의 자연어처리 기술에 의한 전처리 진행 (어휘 분절 알고리즘, 공개 언어모델 등 적용)
자동 코딩	사례사전 및 색인정보DB 활용 - 다수의 사례사전 구축 (수용조건 및 배제조건 색인어로 구성) - 사례사전의 지속적 관리 어려움 - 색인DB 구축의 한계	기계학습 알고리즘 기반 분류모델 학습 - 대량 데이터 전체를 학습에 활용 가능 - 다양한 분류모델 구축 가능 (분류 항목 별 예측정확도 확률을 계산 가능)
수동 코딩	내검원에 의한 수동내검 진행 - 기존 자동연계 결과가 없는 경우 위주 (일부 자동연계에 대해서도 검증 수행) - 예산제약 등으로 선별 내검 실시	선택적 에디팅 기준 마련 및 강화학습 실시 - 분류 영역 및 항목 별 예측 확률 정보 분석 - 내검원 수동내검 진행 기준 모색 - 추가학습 및 재학습으로 수동코딩 축소

카히스토리 사진정보 관리 AI시스템 구축 (24)

사고차량 분석 (이미지 분석)

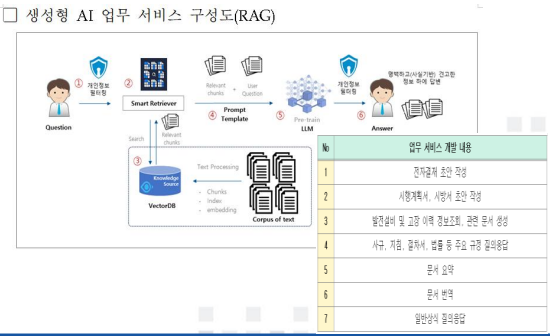


청소년 상담 챗봇 (심리 상담)



생성형 AI기반의 고위험 고립위기 청(소)년 사전예방 시스템 (24~25)

내부 업무서비스 온 프레미스 (LLM)



No	업무 서비스 개발 내용
1	전자결재 초안 작성
2	사업계획서, 사업서 초안 작성
3	발전설비 및 고장 이력 정보조회, 관련 문서 생성
4	사규, 지침, 결재서, 법률 등 주요 규정 질의응답
5	문서 요약
6	문서 번역
7	일반상식 질의응답

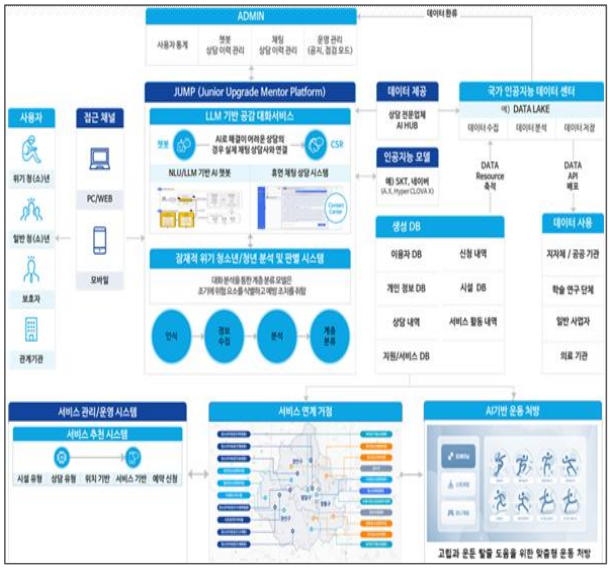
생성형 AI 업무서비스 구축 (25)

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



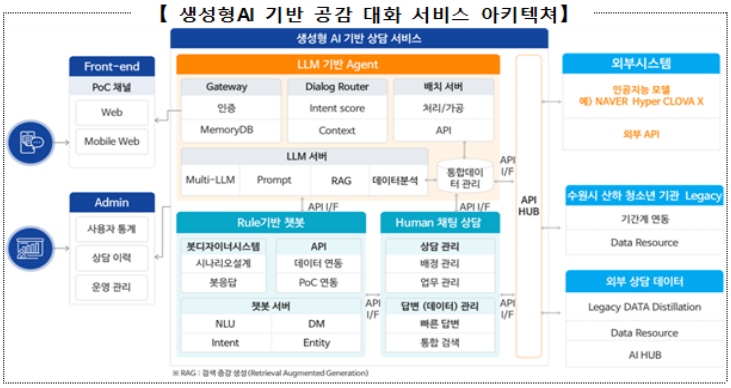
사업 이해

시스템 구성도

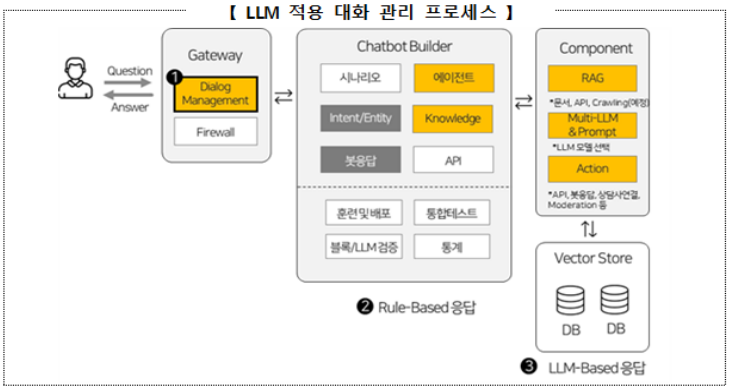


주요 기술

【 생성형AI 기반 공감 대화 서비스 아키텍쳐】



【 LLM 적용 대화 관리 프로세스 】



• 출처: 제안요청서

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



설계단계 감리 AI 관련 점검 항목

감리 점검 항목 (인공지능)

2. AI(인공지능)

가. 점검항목별 점검결과

점검항목	검토항목	점검결과
1. 데이터 수집체계를 적정하게 수행하였는지 여부.	- 추관기관 및 자료 제공기관과 사업자간 학습데이터 수집을 위한 의사소통 및 협조체계 수립 여부. - 학습 데이터 수집을 위한 업무적, 기술적 요건에 대한 준비 여부. • 수집 가능 여부, 보안, 세부 데이터 항목(품질) 등을 검토하여 데이터 수집 목록을 작성하였는지 여부. - AI 학습을 통해 모델 정확도 보증하기 위한 적정규모 이상 학습 데이터 확보 여부. - 특정 분류항목에 편향되지 않은 데이터 확보 여부. - 학습 데이터 보족 시, 학습 정확도를 높이기 위한 방안 수립여부. • 학습 데이터 생성 및 확보, 기존 사례 DB 적용 등.	나.(1)번 참조.
2. 데이터 전처리를 적정하게 수행하였는지 여부.	- 데이터 처리방식 및 수행 적정성. • 분석 효율을 높일 수 있도록 데이터에서 의미있는 정보만 추출하였는지 여부. • 의미 파악이 어려운 비정형 데이터에 대한 형태 변환(분석 가능한 수준) 여부. • 다양한 형식으로 수집된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환(오타 교정, 띄어쓰기 보정)하였는지 여부. • 수집된 데이터의 불일치성을 교정하기 위해 결측치 처리, Noise data 처리 기술 등 적정한 정제 처리 여부. - 비지도 학습에 대한 자연어 전처리의 충분성, 완전성 확보 여부 점검.	나.(1)번 참조.
3. 데이터 저장처리를 적정하게 수행하였는지 여부.	- 저장할 데이터 크기 및 최대 저장기간 등을 고려한 저장 공간 용량 설계 여부. - 데이터 유형에 따른 수집 주기, 저장 방식, 보관 주기 및 백업 방식, 저장 공간 확대 방안 등을 포함한 저장 계획서를 작성하였는지 여부.	나.(1)번 참조.

점검항목	검토항목	점검결과
	- 추기적으로 데이터 운영 관련 에러 및 여유 공간 등을 실시간 모니터링하고 문제 발생시 신속한 대응을 위한 체계 마련 여부. - 데이터를 수집하여 활용하는 단계에서 발생할 수 있는 데이터 보안 침해 가능성을 검토하고 보안 관리 필수 사항을 도출하였는지 여부. - 개인정보 비식별화 조치 등 보안 관리 조치를 수행하였는지 여부.	
4. 데이터 품질 확보 활동을 적정하게 수행하였는지 여부.	- 수집 데이터의 검증 및 정제, 품질검사를 통한 정확성 검증된 데이터 활용 여부. - AI 학습품질 확보를 위해 세부단위 기준 적정규모 이상의 학습 데이터를 확보하였는지 점검. - 평가데이터의 학습 또는 검증데이터의 중복 사용방지 여부. - 각 모델 학습에 이용하는 학습데이터 품질 확보 방안 및 확보 여부 확인. - 각 모델 평가를 위한 평가데이터 확보. - AI 학습을 위한 데이터 품질확보를 위한 전처리 과정 및 학습데이터, 검증데이터, 평가데이터의 충분한 확보 여부.	나.(1)번 참조.
5. 구축 모델 적정성 확보 여부.	- AI 학습에 적용되는 모델 선정 및 적용 타당성, 운영 효율성을 위한 선정 모델 적정성 여부. - 특정 표본에 대해 잘못된 예측을 반환하는지 점검 . - 시간이 지남에 따른 모델 정확도나 공정성이 떨어지는 리스크가 있는지 점검. - 가명화 또는 익명화 해야 하는 개인 식별정보가 있는지 확인. - 각 모델이 소수 분류를 충분히 잘 배분하여 모델이 각 그룹에서 동일한 성능을 보이는지 점검. - 구축된 모델이 윤리적 문제가 없는지 검토(ex. 성별,나이,지역 등). - 인공지능 모델 정확도 검증을 위한 측정 모델 선정의 적정성 점검 (지표: 정확도, 민감도, 정밀도). - 알고리즘이 특정한 데이터 셋에만 최적화 되었는지 등 범용성 확보 여부 점검. - 모델을 배포한 후 데이터 갱신 방안 수립 점검.	나.(1)번 참조.

점검항목	검토항목	점검결과
6. 질의 응답 프로그램 개발 설계.	- 대화관리 모듈 및 화면 설계가 적정하게 이루어졌는가?. • 워크 플로우에 따른 화면 정의 및 UX디자인 컨셉. • 대내/외 연계, AI 대화 시스템 한계 및 대안.	나.(1)번 참조.

[붙임 1] 전문가 실증결과 의견

가. 점검항목별 점검결과

점검항목	검토항목	점검결과
1. LLM(언어모델).	- 정형/비정형 텍스트 입력처리 공통화 적정성 검토. • 생성형 AI기반 자연어 처리 방안 수립 여부.↓ (사전학습 언어모델). - 생성형AI 기반의 기계학습 알고리즘 검토. • 모델 적용 알고리즘 선정의 적정성 확보 여부. • 딥러닝 알고리즘 및 특징추출 기술요소의 적합성 확보 여부. - 객관적인 모델 평가와 모델 검증을 위한 계획이 수립되었는가?. • 요구되는 품질목표에 미달하는 경우 반복적인 모델링을 통한 튜닝 및 학습 수행 계획 점검. • 모델 평가기준과 평가절차에 따른 평가실시 계획 점검. • 성능평가를 위한 테스트커버리지 계획의 적정성 점검. • 운영성 확보를 위해 실 운영 데이터를 통한 운영성확보 계획 점검.	나.(1)번 참조.

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



설계단계 감리 주요 내용

AI 개발 라이프 사이클별 점검 결과 요약 제시

[도표 2-3] 설계 감리 시점 상담 서비스 구현 AI 모델 구축 현황 요약

구분	일반 상담	고립분류모델	자살 키워드 탐지모델	비고
모델 선정	네이버 Hyperclova X (LLM)	BERT (문장 분류 모델 기반 Rule Based 처리)		(다) 참조
학습 데이터 수집	- RAG 시스템 용 데이터 XXXX건 구축 완료(캐릭터 11종→8종 변경에 따라 추가 작업 필요) - 수행사 보유/활용 데이터(가공 완료) • 상담데이터:XXXX건 • 음성상담: XXX 회기	- 학습용: XXXX건, 평가용: XXX 건 확보 목표로 구축중	- 학습용: XXXX건, 평가용: XXX 건 확보 목표로 구축중	(라) 참조
전처리	- 상담 이력 데이터에 대해 1턴 대화 단위별로 대화 의미 및 발화 의도 등 분석을 위해 topic, keyword 라벨링 수행			(마) 참조
데이터 저장	- 저장 용량을 산정하였으나, AI 구축 각 단계 발생 데이터 구분과 운영 단계 데이터 증가분 추이(예: 5년 등)가 반영되지 않음			(바) 참조
데이터 품질	- RAG 데이터: 샘플링 데이터(10%)에 대해 심리상담 전문가 검증 진행 완료함. 최종 데이터 셋 확정 시 전수 검증 예정임 - 고립 분류 모델과 자살 키워드 탐지 모델: 데이터 셋 구축 완료 후 심리상담 전문가에 의한 전수 검사 계획됨			(사) 참조
구축 모델 검증	- AI 모델 별 품질 및 성능 평가 지표를 정의하였으나, 제안요청서에 명시된 내용을 반영한 보완이 필요하며, 성능 지표 및 달성 목표에 대한 주관기관의 검토/승인 진행 예정 - LLM(언어모델)의 성능 향상 및 최적화를 위해 RAG 데이터 구축 적용 외에 상담 도메인 데이터 학습 및 미세조정 (Fine Tuning) 필요			(아) 참조
질의, 응답, UI 설계	- 생성형 AI 기반 사용자 질의/응답 화면은 AI 캐릭터 및 사용자 페이지 형태로 현재 설계 Mobile 화면 설계 진행 중임(제안요청서에 PC/Mobile 환경 모두가 개발 대상임) - 사용자 질의/응답 화면과 연계한 사용자 대화 흐름 제어, NLU, LLM 분석 모델과 연계 등 개발 API 설계 중, API 설계서에 누락됨			(자) 참조

AI 모델 별 성능 평가 지표 요약

[도표 2-16] AI 모델 별 성능 평가 지표 요약

구분		일반대화모드	RAG 시스템		고립위험분류/자살위기탐지모델
평가요소		챗봇의 대화 성능	검색 정확성	검색 랭킹 품질	분류성능
평가 지표	지표명	SSA	Precision@K	nDCG	F1 Score
	정의	합리성(Sensibleness)과 구체성(Specificity) 값의 평균	상위 k개의 검색결과 중 관련성이 높은 결과의 비율(k=5)	검색 결과의 랭킹 품질을 평가하는 지표	정밀도(Precision)와 재현율(Recall)의 조화평균
평가 방법	평가 주체	심리상담전문가	AI 개발팀	AI 개발팀	AI 개발팀
	방법	챗봇 질의/응답결과 정성평가 후 평균	시스템 측정	시스템 측정	시스템 측정
	평가 데이터	A/B 테스트를 위한 대화 세션	멀티턴 대화 기반 벤치마크 데이터 셋 50건 (시스템 평가 가능하도록 정답 라벨링 포함 구축)		일반문장:200건, 고립/자살문장:10건으로 20:1 비율로 혼성 평가
목표 수준	예시	- Google 79% - 사람: 86%	Precision@5 ≥ 0.8	nDCG@10 ≥ 0.7	F1 Score ≥ 0.85
		주관기관 협의 필요			

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



설계단계 감리 주요 내용

설계단계 감리 시점 모듈 별 개발 현황 요약 제시

[도표 3-1-1] 생성형 AI 기반 상담 서비스 플랫폼 구축 공정현황

구분	항목	내용	추진현황	수행사
프론트엔드	사용자 인터페이스	- PC 와 모바일 호환 웹기반 인터페이스 - 사용자 친화적 디자인	- 중간보고 시현 가능 - (다) 프론트엔드 참조	XX
백엔드	대화 관리 시스템	- 사용자와의 대화 흐름제어 - NLU, LLM 분석 모델과 연계 - 상황에 맞는 응답 생성 및 관리	- 중간보고 시현 어려움 - (라) 백엔드 참조	XX
AI 구성 요소	LNU 모듈	- 사용자 입력 의도 및 감정 분석 - 고위험 상황 감지 - 문맥 이해 및 해석	- 중간보고 시현 가능 - (마) AI 구성 요소	XX
	LLM기반 대화 생성	- 상담 도메인 지식 기반 응답 생성 - 맥락을 고려한 연속 대화 유지 - 개인화된 상담 내용 제공		
	대화 분석 모듈	- 실시간 대화 내용 분석 - 위험 수준 평가 - 개입 필요성 판단		
시스템 연동	외부연계 모듈	- 청소년안전망시스템 연동 - 오프라인 상담사 연계 - 가족 상담 연계 기능	- 중간보고 시현 어려움 - (바) 시스템연동 참조	XX
안전관리	위험관리시스템	- 자살 위험 등 극단적 상황 감지 - 긴급 개입 프로토콜 실행 - 관련 기관 즉시 알림	- 중간보고 시현 가능 (단방향) - (사) 안전관리 참조	XX / XX
연계관리	온•오프라인 프로그램 연계	- AI 대화 상담을 통한 프로그램 연계 - 마음돌봄 어플리케이션 연계 - 메타버스 상담 플랫폼 연계 - 청(소)년 거점센터 예약 서비스 - 맞춤형 개입	- 중간보고 시현 어려움 - (아) 연계관리 참조	XX

데이터베이스 관련 요구사항

[도표 4-1-1] 데이터 요구사항 현황표

제안요청서		요구사항 제분화 수	제안요청서 요구산출물
요구사항 ID	요구사항명		
DAR-001	DB 표준화 및 상위표준 준수	1	표준(용어,단어,도메인,코드)정의서
DAR-002	데이터 표준 관리	1	데이터 표준 관리 방안(매뉴얼, 가이드 등)
DAR-003	데이터 구조 설계	1	데이터베이스 정의서, 엔터티(정의서, 애트리뷰트 정의서, 물리데이터 모델, 다이어그램 정의서,테이블정의서, 컬럼정의서
DAR-004	데이터 구조 검증	1	데이터 모델 검증 결과서
DAR-005	데이터 구조 관리	1	데이터 구조관리 방안
DAR-006	데이터 값 검증	1	데이터 값 진단 계획결과
DAR-007	데이터 품질체계 수립	1	데이터(표준, 구조, 값) 관리 매뉴얼, 연계 데이터 관리 매뉴얼
DAR-008	데이터 개방 및 메타데이터 구축	2	개방가능 데이터 목록, 메타데이터 표준관리 항목
DAR-009	초기 데이터 구축	1	
DAR-010	학습용 데이터셋 구축	1	학습용 데이터 셋
DAR-011	데이터 백업 요구사항	1	
DAR-012	데이터 보호 및 개인정보 보호	1	
계		13	

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



종료단계 감리 AI 관련 점검 항목

감리 점검 항목(인공지능)

2. AI(인공지능)

가. 점검항목별 점검결과

점검항목	검토항목	점검결과
1. 데이터 수집체계를 적정하게 수행하였는지 여부.	- 학습자료 수집을 위한 업무적, 기술적 요건에 대한 확보 여부. • 수집 가능 여부, 보안, 세부 데이터 항목(품질) 등을 검토하여 데이터 수집 목록을 작성하였는지 여부. - AI 학습을 통해 모델 정확도를 보증하기 위한 적정 규모 이상 학습 데이터 확보 여부. - 특정 분류항목에 편향되지 않은 데이터 확보 여부. - 학습 데이터 부족 시, 학습 정확도를 높이기 위한 방안 수립여부. • 학습데이터 생성 및 확보, 기존 사례 DB 적용.	나.(1)번 참조.
2. 데이터 전처리를 적정하게 수행하였는지 여부.	- 데이터 처리방식 및 수행 적정성 여부. • 다양한 형식으로 수집된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환(오타 교정, 띄어쓰기 보정)하였는지 여부. • 수집된 데이터의 불일치성을 교정하기 위해 결측치 처리, Noise data 처리 기술 등 적정한 정제 처리 여부.	나.(1)번 참조.
3. 데이터 품질 확보 활동을 적정하게 수행하였는지 여부.	- 수집 데이터의 검증 및 정제, 품질검수를 통한 정확성 검증된 데이터 활용 여부. - 평가데이터의 학습 또는 검증데이터의 중복 사용방지 여부. - 각 모델 학습에 이용하는 학습데이터 품질 확보 방안 수립 여부. - AI 학습을 위한 데이터 품질확보를 위한 전처리 과정 및 학습데이터, 검증데이터, 평가데이터의 충분한 확보 여부.	나.(1)번 참조.
4. 구축 모델 적정성 확보 여부.	- 각 모델의 분류 클래스에 따른 데이터를 배분하여 모델이 각 그룹에서 동일한 성능 확보 여부. - 구축된 모델이 윤리적 문제가 없는지 검토(ex. 성별,나이,지역 등). - 구축 모델 정확도 검증을 위한 측정 모델 선정 적정성 여부(지표: 정확도, 민감도, 정밀도).	나.(1)번 참조.
점검항목	검토항목	점검결과
5. 질의 응답 프로그램 구현 적정성 여부.	- 대화관리 모듈 및 화면 구현이 적정하게 이루어졌는가? • 워크 플로우에 따른 화면 구현 및 사용자 인터렉션의 적정성.	나.(1)번 참조.

점검항목	검토항목	점검결과
7. 개인정보보호	- 개인정보의 안전성 확보조치 기준 준수성. • 홈페이지의 개인정보 처리방침 및 회원 가입 절차 . • 접근권한 관리, 접근통제, 개인정보의 암호화, 접속기록 보관 및 점검, 개인정보 파기 등. - 정보 보안항목에 대한 구현 및 안전보호 조치 여부. • 보안항목에 대한 안전보호 통제, 관리의 적정성. • 고유식별 정보, 민감 정보 등에 대한 암호화 등 안전성 확보 조치의 적정성. - AI 학습 데이터 익명/가명화 및 비식별화 준수성. • 수집 데이터의 비식별성. • 가명정보 처리 기준 준수. • 익명화 및 비식별화 기법 적정성. • 재식별 방지 및 모니터링 체계 구축.	나.(7)번 참조.

2) 생성형 AI 기반의 고위험 고립위기 청(소)년 사전 예방 시스템



종료단계 감리 주요 내용

AI 모델 성능 평가 결과

[도표 2-1-19] 모델 성능 평가 결과 현황

구분	항목	구축현황
페르소나 대화 모듈	SSA 평균	- 기본 LLM: 7.1 - 파인튜닝 LLM: 7.83 - 기본 LLM 대비 향상%: 12%
	페르소나 적합도	- 기본 LLM: 3.02 - 파인튜닝 LLM: 3.91 - 기본 LLM 대비 향상%: 31%
	주제 적합도	- 기본 LLM: 3.11 - 파인튜닝 LLM: 3.74 - 기본 LLM 대비 향상%: 24%
	검수자간 일치도	- 0.85
페르소나 대화 RAG 시스템 모듈	Precision@5	- 목표 성능: 77.4 - 평가시스템 성능: 79.5
	MRR@50	- 목표 성능: 0.440 - 평가시스템 성능: 0.46
고립 맥락 탐지 모델	정밀도(Precision)	- 0.958
	재현율(Recal)	- 0.968
	F1 Score	$2\times\frac{0.968\times0.968}{0.968+0.968}$ = 0.963
자살 맥락 탐지 모델	정밀도(Precision)	- 0.968
	재현율(Recal)	- 0.947
	F1 Score	$2\times\frac{0.968\times0.947}{0.968+0.947}$ = 0.957

AI 윤리 점검 결과

[도표 2-1-20] AI 윤리 구축현황

구분	항목	구축현황
데이터 윤리	데이터 윤리 품질 검토	- 전문가 검증 프로세스 구축. - 정성적·정량적 평가 병행. - 데이터 편향 및 윤리적 문제 점검. - 승인·보완 기록 유지.
개인정보 보호	서버 접근 통제	- 서버 접근 권한 제한. - 접근 이력(log)을 자동으로 기록·감사.
	개인정보 취급 방침	- 개인정보 및 민감 정보 관련 법규(개인정보 보호법, 청소년 보호법 등) 준수. - 데이터 보관 기간, 활용 범위, 파기 방법 등 투명 고지,목적이 달성된 정보는 파기.
투명성 및 책임성	윤리적 책임성	- 외부 신고 및 상담 창구 마련. - 유해 키워드 실시간 모니터링 시스템 구축: 유해 키워드(차별적·폭력적·음란성 표현, 극단적 선택 관련 단어 등)감지 및 실시간 필터링. - 지속적인 개선 방안 마련: 필터링 키워드 추가 및 수정, 유해 발화 사례 분석 및 사후 검토.
윤리적 사용 제한	부적절한 모델 활용 대응	- 악의적인 활용 방지: 불법·유해 행위(사이버 폭력, 가짜뉴스 유포, 특정 집단 혐오 발화 등) 방지 정책. - API 키 발급 및 접근 관리: API 키를 통한 인증 체계, 이상 사용 패턴 발견시 즉각 차단하거나 제한 조치. - 위반 시 조치: 부적절한 용도로 모델 사용(개인정보 침해, 차별적 발언, 폭력·음란물 유포 등)시 사용 중단, API 키 회수 등 강력 조치 이행.

개인정보 비식별화 점검 결과

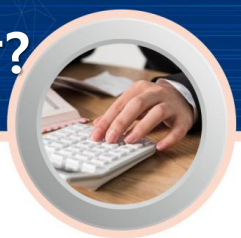
[붙임도표 7-4] AI 학습 데이터의 개인정보 비식별화 적정성 점검 결과

구분	점검 항목	점검 결과	비고
원시 데이터 관리	- 원시 데이터 보호 조치 적정성	- 원시데이터의 경우 AI 서비스 내 학습 완료 후 모두 파기하고 사업 종료 시 주관기관에 파기 확인서 제출 예정임	
초기 학습 데이터 구축	- 초기 데이터 비식별 처리 및 구축 적정성. - 초기 데이터 셋 품질검증 적정성.	- 개인정보가 포함된 원시 데이터를 전처리 및 비식별 처리한 후 초기 학습 데이터 셋(14,000건) 구축을 완료함 ※**육안으로 개인정보를 식별하고 식별된 모든 개인정보에 대해 개인정보의 일부가 아닌 전체를 비식별 처리함** - 초기 데이터 셋은 모든 개인정보의 전체 데이터를 비식별 처리하므로 별도 암호화 저장하고있지 않음 - 데이터 유효성 및 구조 정확성에 대한 검증을 완료함	[붙임도표 7-5] 증적 ①②
확득 데이터 관리	- 대화 이력 데이터 보호 조치 적정성.	- 텍스트 대화 내용을 챗GPT 형식으로 변환 후 DB에 저장함. - DB 저장 시 암호화 적용은 시범운영(3월~8월경) 후 오픈 전에 암호화 솔루션(D' Guard for Cloud 예정) 적용하여 암호화할 예정임.	
공개 데이터 구축	- 공개 데이터 비식별 처리 및 구축 적정성.	- 대화 이력 데이터에서 키워드를 추출(자연어 처리: 조사, 명사 등 키워드 추출)하여 공개 데이터 구축 예정. ※공개 데이터 구축은 자체 제작한 모델을 사용하며 구축 시점은 미정임. ※공개 데이터 셋의 개인정보 비식별 처리 방식은 초기 데이터 구축 방식과 동일하며 별도 암호화 저장 안함.	
재식별 방지 조치	- 재식별 방지 조치 적정성.	- 비식별화된 정보가 유출되는 경우 다른 정보와 결합, 다양한 추론을 통해 개인이 식별될 우려가 있음. - **비식별화된 정보의 유출에 대비하여 재식별 방지를 위한 조치 계획이 수립되어 있지 않음.**	

CHAPTER

6. (생성형) AI 감리 수행시 고려사항

1) 정보시스템 감리원에게 필요한 AI 리터리시란?



정보시스템 감리원에게 AI 리터러시란?

“AI 기술을 이해하고, AI 시스템의 품질을 평가하며, AI 관련 법과 기준에 따라 감리를 수행 할 수 있는 전문성과 책임의식”

감리원에게 요구되는 AI 리터러시 핵심요소

1 AI 기술의 이해

- 머신러닝, 딥러닝, 자연어 처리 등 AI의 기본 개념과 작동 방식의 이해
- AI 시스템의 학습 데이터, 모델 구조, 출력 결과 해석 등 이해 필요

2 AI 기반의 감리역량

- AI가 포함된 시스템의 품질, 신뢰성, 성능 기준을 검토
- 설명가능성, 모델 편향성, 데이터 품질 등 AI 고유 리스크의 판단 능력

3 AI 윤리 및 책임감

- AI의 판단이 차별적이거나 부당한 영향을 주지 않는지 평가
- 개인정보 보호, 알고리즘 투명성, 책임소재 명확화 등 AI 윤리 기준에 따라 감리 수행

4 AI 관련 법제도 이해

- 개인정보보호법, AI 윤리 가이드라인 등 관련 법률•가이드라인 이해
- 감리시, 시스템이 법적 기준에 부합 하는지 확인 할 수 있는 능력

5 AI 프로젝트 관리역량

- AI 프로젝트의 데이터 수집, 전처리, 모델링, 테스트, 운영 전 과정에 대한 이해
- 전통적 시스템과 AI 시스템의 프로젝트 관리 방식 차이에 대한 인식 필요

• 출처: chatgpt.com

2) AI 감리를 위한 감리기준/방법론, AI 전문가, 제도적 뒷받침



생성형 AI는 기존 정보시스템 감리와는 다른 고려사항이 존재합니다.
특히 산업 특성, 기술의 급속한 발전, 전문 인력의 협업, 제도기반이 중요한 요소로 작용합니다.

산업별•서비스별 감리기준 필요

- 생성형 AI는 활용 분야에 따라 요구되는 정확성, 책임성, 위험 수준이 다릅니다.
  - 의료: AI가 진단을 보조할 경우, 잘못된 정보는 생명에 직결될 수 있습니다
  - 금융: 챗봇이나 리스크 모델에서의 오류는 금전적 피해로 이어질 수 있습니다
  - 교육: 생성형 AI가 학생들에게 잘못된 지식을 제공할 수 있습니다

AI 기술 발전 속도에 따른 감리 방법론 필요

- 생성형 AI 기술은 매우 빠르게 진화하고 있어, 최신의 위험요소나 기술 구조를 이해하기 위한 감리 방법론이 필요합니다.
  - GPT-2 -> GPT-4 -> GPT-4o 처럼 모델의 성능과 사용 범위가 급격하게 변합니다.
  - 새로운 리스크 등장: 환각(hallucination) 문제, 데이터 편향, 저작권 침해, 보안 취약점 등 이 있습니다.
  - AI 기술 특성에 맞는 별도 점검 항목과 평가 지표가 필요하며, 주기적으로 업데이트 되어야 합니다

단계별 감리원•AI 전문가 협업

- 감리원은 정보 시스템 구축•운영에 대한 평가 전문성을 가지고 있지만, AI 기술의 내부 구조나 알고리즘의 타당성 판단은 AI 전문가의 영역입니다.
  - 감리원 단독으로는 모델 검증, 데이터 윤리성 평가, 결과 해석의 타당성 검토 등에 한계가 있습니다.
  - 감리원은 제도•법령 관점에서, AI 전문가는 기술적 관점에서 협업함으로 문제 대응이 가능해집니다

제도적 뒷받침 필요

- 현재 감리 기준은 정보시스템 구축 중심이며, AI 프로젝트 특유의 위험요소(데이터 편향, 환각, 설명 불가 등) 에 대한 공식 평가 체계가 미비 합니다.
- AI 판단에 대한 피해나 오류가 발생 했을때, 개발자•운영자•이용자 등 누가 책임질 것인지 모호 합니다.
- AI가 의도치 않게 차별적 판단을 내릴 수 있기 때문에, 공정성•비차별성 확보 기준이 필요 합니다.

• 출처: chatgpt.com

02 | AI 생성 결과물의 신뢰성 확보를 위한 감리 프레임워크

최석우 | 수석
한국정보기술단
한국정보공학기술사회

2025년도 정보시스템감리 공동 심포지엄

AI 생성 결과물의 신뢰성 확보를 위한 감리 프레임워크

최석우 | 수석
한국정보기술단
한국정보공학기술사회

CONTENTS

1 생성형 AI, 혁신과 위험의 시작
2 AI 생애주기별 감리 체크리스트
3 사례분석 및 산업별 적용
4 감리 로드맵 및 결론



CHAPTER

생성형 AI, 혁신과 위험의 시작



생성형 AI란?



생성형 AI(Generative AI)의 정의

- 생성형 AI란, 기존 데이터를 바탕으로 새로운 글, 그림, 소리, 코드 등 '원래 없던' 콘텐츠를 만들어내는 인공지능입니다.
- 이 AI는 방대한 데이터를 학습해 패턴과 규칙을 익히고, 사용자의 요청(프롬프트)에 따라 새로운 결과물을 만들어냅니다.

주요 예시와 활용 분야

1. 글쓰기(텍스트 생성)	2. 이미지 생성	3. 코드 생성	4. 음악/음성 생성	5. 영상 생성
• 예시: ChatGPT, Google Gemini, Copilot • 활용: 기사 작성, 이메일·보고서 자동 작성, 시·소설 창작 • 실제 예: "대학생에 대한 시를 써줘" → AI가 직접 시를 창작	• 예시: DALL-E, Midjourney, Stable Diffusion, Canva • 활용: 광고·디자인 시안, 캐릭터·배경 그림, 제품 콘셉트 • 실제 예: "유니콘을 타는 사서 그림" → AI가 새로운 이미지를 그림	• 예시: ChatGPT, Copilot • 활용: 프로그래밍 코드 자동 완성, 웹사이트/앱 개발 • 실제 예: "무지개 HTML 코드 만들어줘" → AI가 코드 생성	• 예시: Murf.ai, Suno • 활용: 배경음악 제작, 음성 더빙, 가상 DJ, 오디오북 • 실제 예: "여성이 시를 읽는 오디오" → AI가 음성 파일 생성	• 예시: Runway, Sora • 활용: 광고·홍보 영상, 애니메이션, 짧은 동영상 • 실제 예: "시험 공부하는 유니콘 영상" → AI가 동영상 제작

생성형 AI 시장 동향



1. 시장 규모 및 성장률

- 2025년 시장 규모: 378.9억 달러
- 2034년 예상 규모: 1조 50억 달러
- 연평균 성장률(CAGR): 44.2%
- 주요 성장 동인:
  - ✓ 텍스트→이미지·영상 변환 기술 확산
  - ✓ 클라우드 인프라(AWS, Azure) 및 AI 반도체(NVIDIA) 투자 확대

2. 산업별 확산 현황

산업	적용 사례	성과
제조업	현대차 디지털 트윈	결함 94% 사전 감지
금융	한국은행 챗봇	월 50개 편향성 테스트
헬스케어	존스홉킨스 암 진단 AI	진단 정확도 40% 향상
미디어	Netflix AI 추천	시청 시간 35% 증가

3. 지역별 점유율

- 북미: 41% (2025년 기준, Microsoft·Google 주도)
- 아시아 태평양: 최고 성장률(연 27.6%)

4. 미래 전망

- 2030년: 글로벌 GDP의 7% 기여 예상 (McKinsey)

생성형 AI의 위험과 감리 필요성



1. 데이터 중독 (Data Poisoning)

- 정의: 학습 데이터에 악성 데이터를 주입해 AI가 잘못된 판단을 하도록 유도
- 실생활 사례:
  - ✓ MS Tay 챗봇 (2016): 24시간 내 15만 건 악성 프롬프트 입력 → 인종차별 발언 학습 후 서비스 중단
  - ✓ 유럽 물류기업 (2024): 최적 경로 AI 변조 → 연료비 470만 달러 추가 발생
- 해결책 : 블록체인 기반 데이터 출처 추적 (NIA v3.5)

2. 편향 (Bias)

- 정의: 학습 데이터의 편견이 AI 결과에 반영되어 특정 집단을 차별
- 실생활 사례:
  - ✓ UCL 연구 (2024): AI가 여성 성과를 23% 저평가, 백인 남성 고소득 직업 과대평가
  - ✓ 의료 AI (2025): 흑인 환자 진료 권고율 40% 낮춤 → 존스홉킨스 병원 연구
- 해결책 : SHAP/XAI 기반 Fairness Index ≥0.8 검증

3. 환각 (Hallucination)

- 정의: AI가 존재하지 않거나 부정확한 정보를 실제처럼 생성하는 현상
- 실생활 사례:
  - ✓ 법률 AI (2023): 존재하지 않는 판례 6건을 생성해 변호사 제출 → 법원 제재
  - ✓ Google Bard (2024): 제임스웹망원경이 외계행성 첫 촬영 주장 → 사실무근
- 해결책 : RAG(Retrieval-Augmented Generation) 구조 도입

CHAPTER

AI 생애주기별 감리 체크리스트



글로벌 규제와 국내 동향



1. 국제동향

1.1. EU AI법 (2025년 시행)

- 핵심 내용:
  - ✓ **4단계 위험 분류** (금지·고위험·제한·최소)
  - ✓ 고위험 AI (의료·교육·법률 등) → 사전 감리·투명성·설명가능성 의무화
  - ✓ 금지 항목: 사회적 점수제, 실시간 원격 생체인식(일부 예외)
- 시행 일정:
  - ✓ 2025년 2월: 금지 조항 적용
  - ✓ 2025년 8월: GPAI(일반용 AI) 규제 시작

1.2. NIST AI RMF (2023년 발표)

- 목적: **AI 위험 종합 관리** (안전성·공정성·투명성)
- 적용 분야:
  - ✓ **생성형 AI 특화 프로파일** (2024년 7월 추가)
  - ✓ 데이터 품질·모델 강인성·사용자 보호 강조
- 특징: 자발적 준수 → ISO/IEC 23894와 호환

특징	EU AI 법	NIST AI RMF	TTA CAT 인증	NIA 가이드라인
핵심 내용	위험을 네 단계로 분류	포괄적인 위험 관리 프레임워크	AI 시스템, 개발자, 사용자에 대한 인증	데이터 품질 지표 및 관리
시행 일정	단계적 시행, 2025년 2월 시작	자발적 준수, 업데이트된 프로파일	2024년부터 진행 중	2025년 3.5판
평가 기준	고위험 AI에 대한 사전 시장 감사	안전성, 공정성, 투명성을 강조	ISO 23894 및 ISO 42001 기반	다중 모달 데이터 품질 지표

2. 국내 동향

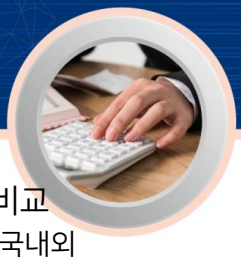
2.1. TTA CAT 인증 (2024년~)

- 인증 대상: AI 시스템·개발사·사용자
- 평가 기준:
  - ✓ ISO 23894 (위험관리) + ISO 42001 (경영시스템)
  - ✓ 기능 시험 강화: 실 운영 환경 검증

2.2. NIA 가이드라인 v3.5 (2025년)

- 1권 : 멀티모달 데이터 품질지표 (F1-Score ≥0.95)
- 3권 : 생성형 AI 특화 관리법 (*프롬프트 버전 관리*)

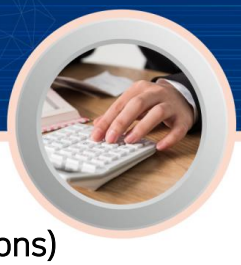
품질관리 프레임워크 개요



- **품질관리 프레임워크 개요** 국제(ISO/IEC 23894, OWASP)와 국내(TTA CAT, NIA) 표준 비교
  - ✓ EU AI법과 NIST AI RMF가 제시하는 위험 관리, 투명성, 데이터 품질 등 핵심 원칙은 아래와 같이 국내외 품질관리 프레임워크(ISO/IEC 23894, TTA CAT 등)에서 구체적으로 실현되고 있음

구분	국제		국내	
	ISO/IEC 23894	OWASP LLM	TTA CAT	NIA 가이드라인 v3.5
초점	AI 생애주기 위험관리	LLM 특화 보안 검증	AI 시스템 신뢰성 인증	멀티모달·생성형 AI 데이터 품질
적용 범위	글로벌 기업·모든 산업	생성형 AI 시스템	국내 공공·금융 간	한국 AI 데이터 구축 프로젝트
주요 요구사항	- 리스크 식별·평가·처리·모니터링 - ISO 31000 기반 통합 관리 - 동적 위험 대응 (예: 재학습 트리거)	- 프롬프트 인젝션 방어 (L2-REQ-03) - 모델 암호화 (L3-REQ-12) - 실시간 이상 탐지	- 15개 요구사항 (거버넌스 40%, 데이터 30%) - 고위험 AI 연 2회 감리 - 레드팀 평가	- 멀티모달 F1-Score ≥0.95 - 합성데이터 블록체인 추적 - 생성형 AI 프롬프트 버전 관리
적용 단계	기획 → 폐기 전 과정	개발 → 배포 → 모니터링	인증 심사 (문서·기능 시험)	데이터 수집 → 검증 → 활용
핵심 도구	리스크 매트릭스·KL 발산	CleverHans·Prometheus	자가검증 체크리스트	GAN 검증 프로토콜·SHAP 분석
인증 대상	AI 시스템·조직	LLM 애플리케이션	AI 제품·개발사·사용자	데이터 구축 기관·생성형 AI
규제 강제성	자발적 준수	기술적 권고	의무 대상: 공공·금융	산업별 권고

데이터 수집·전처리 체크리스트



1. 합성데이터 출처 추적 (🔗 블록체인)

NIA 가이드라인 v3.5 (2025) 기준

- 검증 프로세스(변조불가):
  - ✓ 생성: GAN/VAE 모델 해시값 기록
  - ✓ 전달: 데이터 이동 경로 블록체인 추적
  - ✓ 저장: 분산 원장(DLT)에 영구 보관
- 주요 도구:
  - ✓ 하이퍼레저 패브릭, 이더리움 기반 스마트 계약
- 실제 사례:
  - ✓ 현대차 디지털 트윈: X-ray 검사 데이터 해시값 기록 → 결함 감지 정확도 94% 달성

2. 편향성 분석 (⚖️ SHAP)

SHAP(SHapley Additive exPlanations)

- 목적: 각 데이터 특성이 AI 예측에 미치는 영향 정량화
- 기준: Fairness Index ≥0.8 (NIA 가이드라인)
  - ✓ AI 시스템의 공정성을 평가하기 위한 최소 허용 기준

분석 항목	검증 방법	사례
성별 편향	SHAP 값 기여도 분석	한국은행 챗봇, 성별에 따른 대출 승인율 차이 12% 감소
지역 편향	지역별 데이터 분포 비교	국회도서관 SLM, 지역 방언 테스트 120종 적용

3. 핵심 원칙 요약

- 합성데이터: 블록체인으로 출처 투명성 확보
- 편향성: SHAP으로 공정성 검증 → "왜 이런 결과가 나왔나?" 설명 가능
- 도구: Hyperledger + SHAP + Prometheus/Grafana 연동

모델 개발·학습 체크리스트



1. 적대적 훈련 (Adversarial Training)

- 목적: AI가 악성 공격(해킹·조작)에 강인해지도록 학습
- 실행 방법:
  - ✓ FGSM 기법으로 공격 샘플 5% 생성 후 혼합 학습
- 목표: Robust Accuracy ≥95% (예: 현대차 결함 감지 94%)

2. GAR(Generative AI Retrieval) 전환

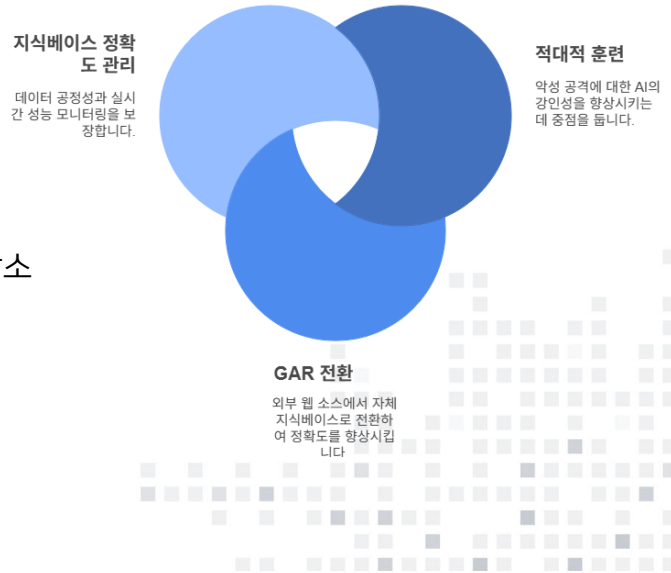
- RAG → GAR 구조 개선

구분	RAG	GAR
데이터 소스	외부 웹	자체 지식베이스
장점	넓은 정보 범위	높은 신뢰도·정확도

- 적용 사례:
  - ✓ 삼성 가우스: 자체 벡터 DB 연동 → 환각 현상 70% 감소

3. 지식베이스 정확도 관리

- 검증 항목:
  - ✓ SHAP 기반 Fairness Index ≥0.8
  - ✓ 월간 오류 데이터 재학습 (변경 사항 24시간 내 반영)
- 도구:
  - ✓ iWeaver: 데이터 버전 관리 (Git 기반)
  - ✓ Prometheus: 실시간 성능 모니터링



운영·모니터링 체크리스트



1. 실시간 모니터링 (Prometheus + Grafana)

모니터링 항목

지표	설명	임계값
정확도	예측/생성 정확도	≥98%
지연 시간	사용자 요청 처리 시간	〈0.5초
CPU/GPU 사용률	리소스 효율성	〈80%
데이터 분포	입력 데이터 변화 (KL 발산)	〈0.3

주요 기능

- Prometheus: 실시간 데이터 수집 (메트릭 저장/쿼리)
- Grafana: 대시보드 시각화 + 슬랙/이메일 경고 통합
- 사례: Apple AI 뉴스 요약 도구 오류 시 5분 내 서비스 중단 방지

2. 모델 드리프트 대응

드리프트 탐지 방법

- KL 발산(Kullback-Leibler): 입력 데이터 분포 변화 측정 (≥0.3 시 경고)
- ADWIN 알고리즘: 데이터 스트리밍 환경에서 점진적 변화 감지
- 오류율 추적: 일일 오류율 ≥5% 시 재학습 트리거

재학습 프로토콜

- 트리거 조건: KL 발산 ≥0.3 또는 오류율 5% 초과
- 실행 주기: 매주 금요일 00:00 자동화 스케줄링
- **도구:**
  - ✓ Scikit-Learn: partial_fit으로 증분 학습
  - ✓ TensorFlow Model Salvage: 기존 가중치 기반 재학습

3. 핵심 원칙

- 자동화: 경고 → 재학습 파이프라인 구축 (Apache Airflow)
- 투명성: SHAP/XAI로 드리프트 원인 설명 (예: "연령별 데이터 편향 15% 증가")
- 협업: 개발팀 + 데이터팀 주간 리뷰 회의 진행

CHAPTER

사례분석 및 산업별 적용



데이터 오염 실제 사례



1. MS Tay 챗봇 (2016)

문제:

- 24시간 내 15만 건 악성 프롬프트 입력 → 인종차별·성차별 발언 학습
- 원인: 오픈 소셜 학습 시스템 취약성 + 실시간 필터링 부재
- 결과: 서비스 16시간 만에 중단, Microsoft 신뢰도 추락

교훈:

- 사전 필터링 시스템 도입 (금지어 패턴 2,500개)
- 인간 검수팀 실시간 모니터링 의무화



2. 국내 육아앱 챗봇 (2023)

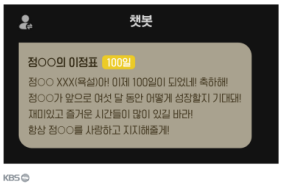
문제:

- 커뮤니티 데이터 12% 비속어 포함 → "XXX아 축하해!" 욕설 메시지 발송
- 원인: "친근한 톤" 구현을 위한 무분별 데이터 학습
- 결과: 사용자 6만 6천 원 환불 요구 + 앱 평점 4.8→1.2 급락

교훈:

- 학습 데이터 품질 검증 프로토콜 강화 (예: SHAP 기반 편향성 분석)
- 고객 대응 매뉴얼 개선 (삭제 요청 시 즉시 조치)





산업별 AI 품질관리 적용 사례



1. 제조업: 현대차 디지털 트윈
핵심 기술

- IoT + NVIDIA Omniverse 연동:
  - ✓ 가상 공장과 실제 센서 데이터 실시간 동기화
  - ✓ 초당 10만 데이터 포인트 분석 → 결함 94% 사전 감지
- 모니터링: Prometheus/Grafana 기반 KL 발산 ≥0.3 시 재학습
- 성과
  - ✓ CO_2 배출량 37% 감소 (HMGICS, 2025)
  - ✓ 에너지 소비 최적화 시간 78% 단축

2. 금융·공공 분야

분야	전략	성과
한국은행 챗봇	SHAP 기반 월 50개 편향성 테스트	응답 정확도 90%
국회도서관 SLM	BERT-SLM 방언 120종 테스트	고령층 접근성 89% ↑

기술 세부

- 금융: FINPROMPT 표준 + EU AI법 준수
- 공공: TF-IDF 가중치 적용 방언 인식 (예: "고마우따" → "고맙습니다")

CHAPTER

감리 로드맵 및 결론



미래 기술 및 감리 로드맵

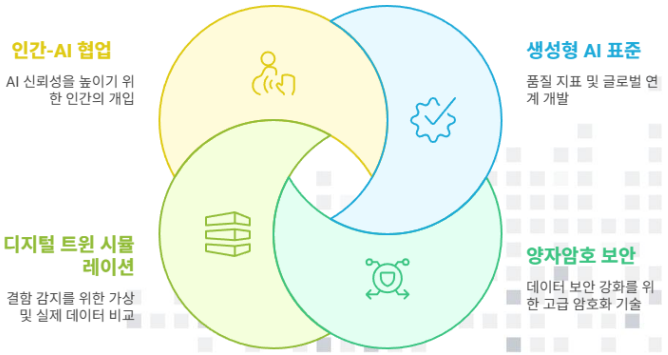


1. 2026년 목표: ISO/IEC 25040 생성형 AI 전용 표준 제정
- 생성형 AI 특화 품질지표: 데이터 무결성, 모델 강인성, 설명가능성
- 글로벌 연계: EU AI법, NIST RMF와 연계한 감리 체계 구축
- 적용 분야: 의료, 금융, 자율주행 등 고위험 AI 시스템

2. 2030년 비전: 양자암호(QRC) 기반 보안 검증
- 삼성 QRC 칩 도입: 기존 RSA 대비 127배 빠른 복호화
- 모델 가중치 및 학습 데이터 암호화 검증 의무화
- 보안 강화: 양자내성암호(PQC) 적용, 물리적 변조 감지

3. 디지털 트윈 기반 시뮬레이션 감리
- 가상 공장과 실제 센서 데이터 비교: 결함 94% 사전 감지
- KL 발산 임계치 초과 시 자동 재학습 트리거
- 산업 적용: 현대차 HMGICS, 공장 설계 최적화

4. 인간-AI 협업 모델(HITL)
- 기술적 검증과 윤리적 판단의 융합
- HITL(Human-in-the-Loop) 프레임워크:
  - ✓ AI 출력 신뢰도 낮을 시 인간 검수원 개입
  - ✓ 예: 의료 진단 결과 최종 승인



결론



1. 데이터 품질의 중요성
- 깨끗하고 신뢰할 수 있는 데이터가 AI 성능의 출발점
- 데이터 중독·편향 방지, 출처 추적(블록체인) 필수
- 예시: MS Tay 챗봇, 국내 육아앱 사례

2. 실시간 점검의 필요성
- Prometheus + Grafana 대시보드로 성능·오류 실시간 모니터링
- 모델 드리프트 발생 시 신속 재학습으로 품질 유지
- 예시: KL 발산 ≥0.3 시 주간 자동 재학습

3. 글로벌 표준 준수
- ISO/IEC 25040, EU AI법 등 국제 기준에 맞춘 감리 체계
- 윤리적 AI 인증제도 확대, 사회적 신뢰 확보
- 예시: TTA CAT, NIA 가이드라인 v3.5





용어설명 (1)



용어	설명
생성형 AI	글이나 그림을 스스로 만들어내는 똑똑한 인공지능 (예: 챗GPT, 이미지 생성 AI 등)
CAGR	매년 평균으로 얼마나 성장했는지를 나타내는 숫자 (연평균 성장률)
Data Poisoning	AI한테 일부러 틀린 데이터를 줘서, 이상하게 배우도록 속이는 공격 방법
SHAP	AI가 왜 그런 예측을 했는지, 어떤 특징이 영향을 줬는지 알려주는 도구
XAI	AI가 한 행동을 사람이 이해할 수 있게 설명해주는 기술 (Explainable AI)
RAG	검색해서 필요한 정보를 찾아와서 대답하는 AI 구조 (예: 챗GPT + 검색 기능)
Fairness Index	AI가 사람들에게 공평한 판단을 하고 있는지를 숫자로 평가하는 기준
EU AI 법	유럽에서 만든 AI 법. 위험한 AI는 감시하고 규칙을 지키게 함
NIST AI RMF	미국에서 만든 AI 리스크 관리 지침 (AI 안전하게 쓰는 가이드라인)
TTA CAT	우리나라에서 AI 기능을 인증해주는 시험 제도
ISO/IEC 23894	AI를 쓸 때 위험을 관리하는 국제 표준
ISO/IEC 42001	AI를 회사나 조직에서 잘 운영하도록 도와주는 경영 국제 기준

용어설명 (2)



용어	설명
Hyperledger Fabric	기업에서 쓰는 블록체인 프로그램 (오픈소스 기반)
GAN / VAE	그림, 소리, 데이터 등을 새로 만들어내는 AI 모델
Adversarial Training	AI를 일부러 공격해보며 더 튼튼하게 훈련시키는 방법
GAR	검색해서 만든 정보를 더 잘 정리해서 답해주는 구조
Prometheus	서버나 시스템 상태를 실시간으로 감시해주는 프로그램
Grafana	시스템 상태를 예쁜 차트로 보여주는 프로그램
KL 발산	AI가 배운 게 이상해졌는지 수치로 확인하는 방법
ADWIN	데이터가 달라졌는지 자동으로 감지하는 똑똑한 알고리즘
iWeaver	AI가 어떤 데이터를 언제 학습했는지 관리해주는 도구
Apache Airflow	여러 작업을 자동으로 순서대로 처리해주는 프로그램
TF-IDF	어떤 단어가 문서에서 얼마나 중요한지를 계산하는 방법
QRC	완전 랜덤한 숫자를 양자기술로 만들어 보안에 쓰는 암호 방식
PQC	양자컴퓨터로도 풀기 어려운 암호 기술
HITL	AI가 스스로 다 하지 않고, 사람이 중간에 확인하고 개입하는 방식

감사합니다



03 | AI 시대의 정보시스템 감사

이동기 | 부회장
한국정보시스템감사통제협회

2025년도 정보시스템감리 공동 심포지엄

AI 시대의 정보시스템 감사
-신뢰성과 통제를 위한 프레임워크

이동기 | 부회장
한국정보시스템감사통제협회

CONTENTS

1 왜 지금 AI 감사인가?
2 ISACA AAIA 감사 역량
3 AI 감사 프레임워크



CHAPTER

왜 지금 AI 감사인가?

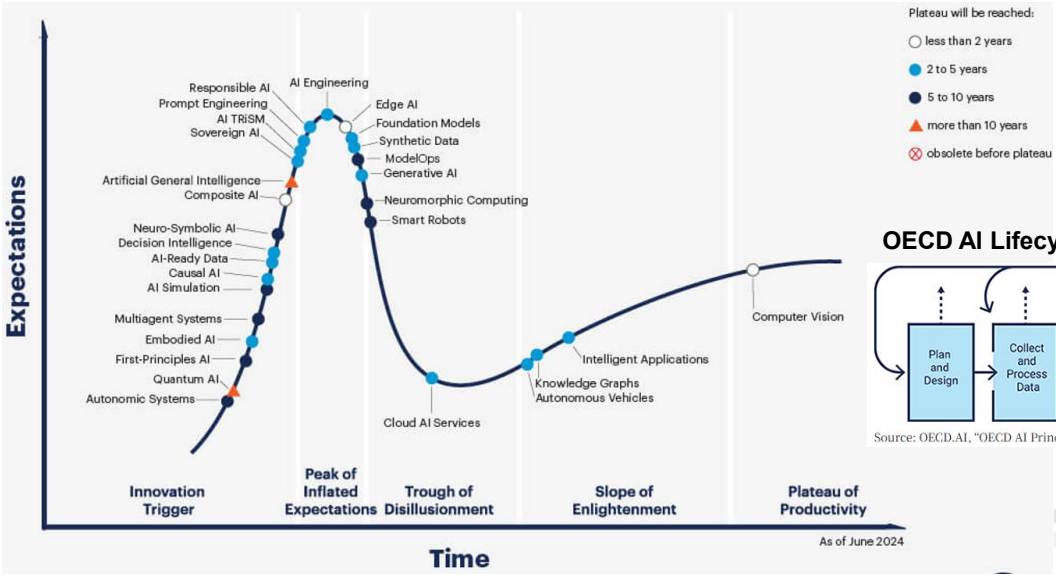


왜 지금 AI 감사인가?

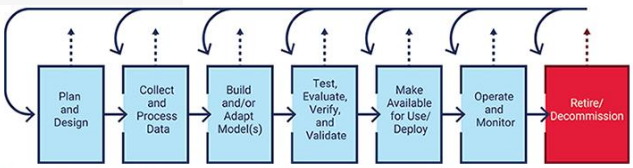


AI는 '새로운 기회' 를 제공하는 동시에, 리스크 관리가 필요한 '감사 대상'

- 생성형 AI의 확산으로 인해 신뢰성, 책임성, 리스크 통제가 조직 운영에서 중요한 감사 이슈로 부상
- 특히, AI 생애주기(Life Cycle)는 단순한 기술 도입을 넘어, 설계-개발-운영-종료 전 과정에서 통제가 필요하며, 비즈니스 목표와의 연계 여부 또한 감사/감리의 핵심 평가 기준



OECD AI Lifecycle



Source: OECD.AI, "OECD AI Principles overview," https://oecd.ai/en/ai-principles

Explore Beyond GenAI on the 2024 Hype Cycle for Artificial Intelligence, Afraz Jaffri | November 11, 2024

CHAPTER

ISACA AAIA 감사 역량

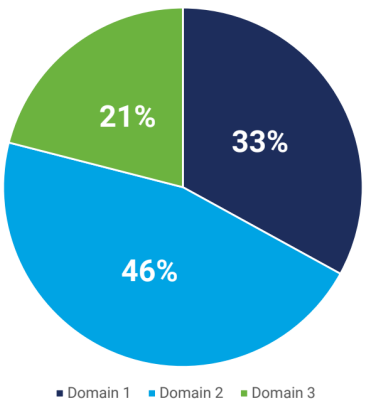


ISACA AAIA 감사 역량



국제 감사 전문가들에게 요구되는 AI 감사 역량

- AI 감사의 핵심 도메인은 거버넌스, 운영, 감사기법으로 구성



Domain	Subdomain
1. AI Governance and Risk (거버넌스 및 리스크)	AI Models, Considerations, and Requirements AI Governance and Program Management AI Risk Management Privacy and Data Governance Programs Leading Practices, Ethics, Regulations, and Standards for AI
2. AI Operations (운영)	Data Management specific to AI AI Solution Development Methodologies and Lifecycle Change Management specific to AI Supervision of AI Solutions (e.g., outputs, impacts, and decisions) Testing Techniques for AI Solutions Threats and Vulnerabilities specific to AI Incident Response Management specific to AI
3. AI Auditing Tools and Techniques (감사기법)	Audit Planning and Design Audit Testing and Sampling Methodologies Audit Evidence Collection Techniques Audit Data Quality and Data Analytics AI Audit Outputs and Reports

CHAPTER

AI 감사 프레임워크

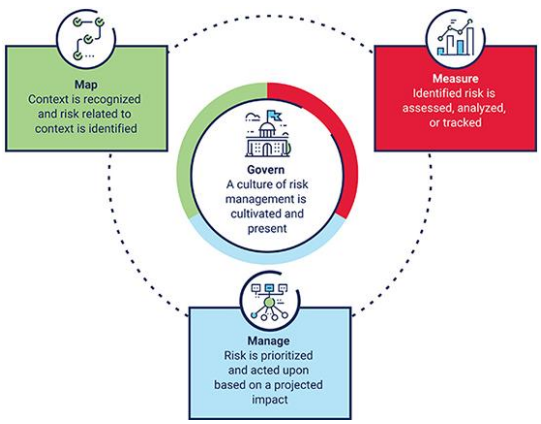


AI 감사 프레임워크 – 거버넌스 및 리스크 (Domain 1)



AI 감사의 출발점은 'GRC' (거버넌스, 리스크 관리, 통제와 책임)

- 조직 내 AI 책임 분배, 윤리 정책, 규제 준수 여부 등은 감사 대상의 핵심이며, 이를 통해 신뢰 기반의 거버넌스 구축이 가능
- AI 리스크 관리의 목적: AI 솔루션이 사람, 조직, 전체 생태계에 미칠 수 있는 부정적인 영향을 최소화
- AI 개발자와 사용자는 개발 초기 단계부터 전 생애주기에 걸쳐 AI 리스크를 선제적으로 식별, 문서화,관리



NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023 Jan

- AI 리스크 식별은 조직의 전사적 리스크 관리 프로그램에 통합
  - 정기적인 리스크 평가, 시스템 변경 및 신규 솔루션 도입 시 리스크 식별 및 평가 등
  - AI 리스크는 위협 주체별, 리스크 형태, 기술적·운영상 특성을 함께 고려
  - 리스크 식별 및 관리를 담당하는 자는 인간이 AI 솔루션을 '본질적으로 신뢰할 수 있는 것'으로 잘못 간주할 수 있다는 리스크에 주의 (professional skepticism)

[AI 리스크 관리의 주요 기능]

- AI 거버넌스 (Govern) — 거버넌스와 리스크 관리 프로그램을 수립
  AI 리스크 식별 (Map) — 리스크 및 그 잠재적 영향을 식별하고 이해
- AI 리스크 평가 (Measure) — 리스크를 정량화하고 분석하며 대응
- AI 리스크 모니터링 (Manage) — 리스크 관리 프로그램을 지속적으로 감독하고 업데이트

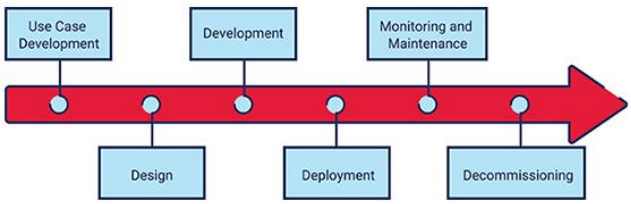
AI 감사 프레임워크 – 운영 (Domain 2)



AI의 가치'는 '운영 통제'에 의해 결정

- AI 운영 전반(데이터 품질, 편향, 변경관리, 사고 대응, 외부 모델 활용 등)에 대한 통제가 감사의 주요 대상
- 전통적인 소프트웨어 개발은 논리에 기반을 두고 있으며, 개발자가 코드로 작성한 명령어에 따라 애플리케이션이 작동하도록 설계
- AI 솔루션 개발은 데이터와 알고리즘에 기반, 데이터 학습하고 패턴을 인식, 추가적인 프로그래밍 없이도 시간이 지남에 따라 스스로 변할 수 있으므로(model drift) 고유한 속성, 리스크 해결 위한 차별화된 개발 생애주기 요구
- AI 프로젝트 실패는 일반적으로 사람, 프로세스, 데이터 측면의 문제로 귀결되며, 다음과 같은 공통적인 실패 요인이 식별됨

Figure 2.4—AI Solution Development Life Cycle



[예시 - 공통 실패 요인]

- AI를 활용해 해결할 문제를 잘못 이해하거나 잘못 전달
- 데이터 품질 부족
- 비즈니스 문제 자체를 해결하지 못함
- AI 모델을 관리·배포할 인프라 부족
- AI가 해결하기엔 지나치게 어려운 문제에 AI를 적용

AI 감사 프레임워크 - 감사기법 및 도구 (Domain 3)



'AI 를 대상으로 감사' & 'AI 를 활용한 감사'

- 감사계획, 샘플링, 감사 증적 수집, 데이터 분석, 결과 보고서 작성 등 기존의 감사 절차와 유사하지만, AI 특유의 복잡성과 기술적 특성으로 인해, 추가적인 고려사항에 유의하여, 실무적 이해도 중심의 감사 수행 역량 요구
- AI 감사에서는 감사 대상 뿐 아니라, 감사 기법 자체도 고도화되어, AI 시스템을 감사하기 위해서는, 기존 시스템 감사보다 더 정교한 계획 수립, 표본 추출, 데이터 분석, 증적 수집이 필요
(알고리즘의 결정 방식 검증, 설명 가능성을 확보하기 위한 데이터 로깅 등)
- 'AI를 감사'하는 동시에 'AI를 활용한 감사'가 되는 이중 구조로 발전
(감사인이 직접 AI 기반 툴 활용해 이상 탐지나 자동화된 감사 절차를 실행하는 사례 증가 중)
- AI 특유의 복잡성과 불확실성으로 인해, 전통적인 "감사(audit)" 역할만으로는 조직이 직면한 위험과 통제의 실효성을 충분히 확보하기 어렵기 때문에, "자문(advisory)" 보고서 형태가 감사 영역에 병행 또는 보완적으로 도입

ISACA AAIA ISACA Advanced in AI Audit.

AI 감사 프레임워크 - 감사기법 및 도구 (Domain 3)



'AI를 대상으로 감사' & 'AI를 활용한 감사'

Figure 3.14—AI Audit Report

Report Area	Details
Executive summary	Key findings and action items summarization
	Audit conclusion and opinion
	Overall audit rating
Scope and objectives	Scope overview: • Accountabilities in scope • Systems in scope • Models in scope • Data in scope • Processes in scope • Controls in scope
	Methodology and objectives: • Data management ▪ Quality and integrity ▪ Data privacy • Model performance testing ▪ Performance measures ▪ F1 score ▪ Bias and fairness outcomes ▪ Transparency and explainability ▪ Security testing • Technology and process testing ▪ System development life cycle ▪ Change management ▪ Configuration management ▪ Release management • Ethical testing outcomes • Compliance and regulatory adherence ▪ Internal for policies and procedures ▪ External for obligations
Detailed findings	Linked to risk assessment and controls assessment mappings
	Recommended action items and timelines for remediation • Operational practicalities and financial considerations
	Management owners and responses
Appendices	Areas to be captured for further information or to enumerate technical or statistical details

- **자문 보고서(Advisory Reports)**는 내부감사 부서가 컨설팅 형태로 수행하는 업무를 통해 작성하고, 일반 구성원이 아닌 AI 전문가(감사인)이 최고감사책임자(CAE)를 통해 이사회 또는 감사위원회에 직접 보고하는 구조
- 조직의 프로세스, 사람, 기술을 잘 알고 있기 때문에 외부 컨설턴트보다 훨씬 더 깊이 있는 전문성(subject matter expertise)을 제공.
- AI와 관련된 자문 업무를 감사인이 수행하게 되면, 예상하지 못한 부작용을 줄이고 전반적인 AI 시스템 개발 및 도입의 완성도를 제고
  (많은 조직에서 AI 는 아직 새로운 영역이며, 조직 전체의 다양한 이해관계자들이 관여)
- **범위** - 자문 보고서의 범위는 참여의 성공과 산출물에 대한 기대치 조정에 있어 매우 중요
- **리스크 고려사항** - 리스크 관리 및 통제 개선 자문 (독립성 고려)
  : 시정조치(remediation) 절차 중 필요한 가이드 및 조언을 포함할 수 있으나,
  : 조직의 통제를 직접 설계, 프로세스/통제에 대한 소유권 X

ISACA AAIA ISACA Advanced in AI Audit

감사합니다

