챗GPT 등
생성형 AI 활용
보안 가이드라인

국가정보원
NATIONAL INTELLIGENCE SERVICE

NSR 국가보안기술연구소

챗GPT 등
생성형 AI 활용
보안 가이드라인

목차

1. 개요

1.1. 생성형 인공지능 기술 소개

- 생성형 인공지능 기술이란 인공지능 기술의 한 종류로서 이미지, 비디오, 오디오, 텍스트 등을 포함한 대량의 데이터를 학습하여 사람과 유사한 방식으로 문맥과 의미를 이해하고 새로운 데이터를 자동으로 생성해 주는 기술을 의미한다.

- 기존 AI 기술이 회귀(regression), 분류(classification), 군집화(clustering) 등 판별적(discriminative) AI 기술이었다면, 생성형 AI 기술은 이용자가 요구한 질문이나 과제를 해결하기 위해 주어진 데이터를 기반으로 패턴과 규칙을 학습하고 이를 통해 새로운 컨텐츠를 생성하는 기술이다.

- 생성형 인공지능 기술 중 하나인 대규모 언어모델(Large Language Models; LLMs)은 일반적으로 수백억 개 이상의 파라미터를 포함하는 인공지능 모델을 의미하며 복잡한 언어 패턴과 의미를 학습하고 다양한 추론 작업에 대해 우수한 성능을 보유하고 있다.

- 최근 주요 생성형 대규모 언어모델은 〈표 1〉과 같으며 대다수의 모델은 API 등을 통해 제한적으로만 접근 가능하다. 대규모 언어모델 기반의 대표적인 AI 서비스로는 Stable Diffusion (Stability AI; 2022), Midjourney (Midjourney; 2022), ChatGPT (OpenAI; 2022), Bard (Google, 2023), Firefly (Adobe, 2023) 등이 있다.

표 1 . 최근 주요 생성형 대규모 언어모델

공개일	모델명	기관	파라미터 (개)	학습 토큰 (개)	공개 여부
2020.05	GPT-3	OpenAI	1,750억	3,000억	비공개
2021.10	GShard	Google	6,000억	1조	비공개
2022.01	LaMDA	Google	1,370억	2.81조	비공개
2022.05	OPT	Meta AI	1,750억	1,800억	공개
2022.02	AlphaCode	DeepMind	410억	비공개	비공개
2022.03	InstructGPT	OpenAI	1,750억	비공개	비공개
2022.03	Chinchilla	DeepMind	700억	1.4조	비공개

2022.03	InstructGPT	OpenAI	1,750억	비공개	비공개
2022.03	Chinchilla	DeepMind	700억	1.4조	비공개
2022.04	PaLM	Google	5,400억	7,680억	비공개
2022.09	Sparrow	DeepMind	700억	비공개	비공개
2023.02	LLaMA	Meta	650억	1.4조	공개
2023.03	GPT-4	OpenAI	비공개	비공개	비공개
2023.03	PanGu-Σ	Huawei	1.085조	3,290억	비공개
2021.09	HyperCLOVA	NAVER	820억	5,616억	비공개
2021.11	KoGPT	Kakaobrain	60억*	2,000억	공개
2021.12	EXAONE	LG	3,000억	6,000억	비공개
2023.06 (예정)	MI:DEUM	KT	300억	비공개	-
2023.07 (예정)	HyperCLOVA X	NAVER	2,040억	5,600억	-

* KoGPT는 파라미터 수가 100억 개 미만이지만, 동향 분석을 위해 추가

1.2. 생성형 인공지능 기술의 기본 개념 – GPT를 사례로

• GPT(Generative Pre-trained Transformer)는 대규모 언어모델로서 도서, 웹 문서 등에서 수집한 방대한 텍스트 데이터베이스를 기반으로 학습하여 언어의 통계적 패턴을 모방하고, 이를 토대로 설득력 있는 문장을 생성하는 기술이다.

• GPT는 2018년 OpenAI에서 처음 제안하였으며, 2019년 GPT-2, 2020년 GPT-3가 각각 발표되면서 점차 학습에 사용되는 데이터의 크기 및 모델의 파라미터(매개변수) 수가 증가하는 추세를 보이고 있다(〈그림 1〉, 〈표 2〉).

  – 파라미터를 확장시 보다 많은 정보를 학습하고 복잡한 문제를 해결할 수 있게 되는데 특히 GPT-3는 이전의 AI 모델이 가지고 있던 파라미터를 크게 확장하면서 알고리즘의 성능을 혁신적으로 증가시켰다.

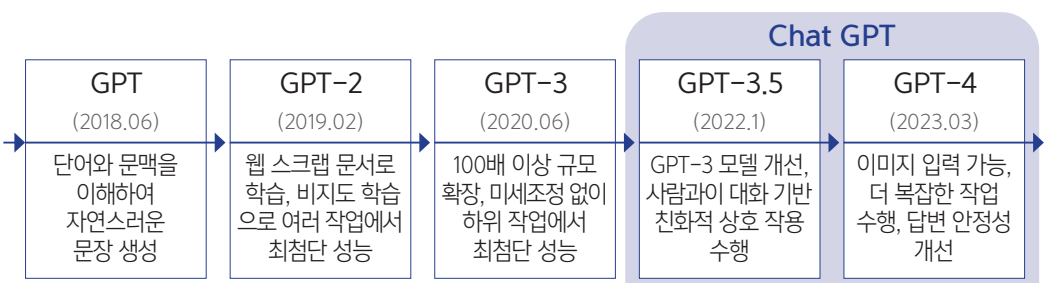

그림 1. GPT 모델 군(family)의 발전 동향

표 2. 각 GPT 버전에 대한 세부적인 내용

항목	GPT	GPT-2	GPT-3	GPT-3.5	GPT-4
학습 데이터 크기 (GB)	5GB	40GB	570GB	비공개	비공개
학습 데이터 수집 기간	~2015	~2017.12	~2019.10	~2021.09	~2021.09
파라미터 수 (개)	1.17억	15억	1,750억	1,750억	비공개
최대 입력 길이 (토큰*)	512	1,024	2,048	4,096	32,768
모델 공개 여부	공개	공개	제한 (API)	제한 (API)	제한 (API)

* 1,000개 토큰은 약 750단어에 해당

• GPT-3.5 및 GPT-4는 OpenAI의 AI 챗봇 서비스인 챗GPT를 통해 각각 2022년 11월과 2023년 3월에 공개되었다. GPT-3.5는 2022년에 공개된 InstructGPT와 유사한 방식으로 개발되었으며, 이 모델은 GPT-3를 강화학습(Reinforcement Learning from Human Feedback; RLHF) 방식을 통해 미세 조정한 모델이다.

  - 이로 인해, 이전 버전의 AI 모델들보다 더욱 자연스러운 대화가 가능해졌으며, 사람이 원하는 답변을 더 잘 생성할 수 있게 되었다.

  - GPT-4는 GPT-3.5보다 한 단계 발전한 모델로서, 고수준 추론 작업에 탁월한 성능을 보이며, 더 긴 응답을 생성할 수 있고, 이미지와 텍스트를 동시에 입력으로 받아 텍스트를 반환할 수 있는 멀티모달(multi-modal) 모델이다.

| 사례 | 챗GPT 구조와 동작 방식

• GPT 모델 군은 모두 트랜스포머 디코더(decoder) 블록을 여러 층으로 쌓은 형태를 가지고 있다. 이 모델들은 사전 학습(pre-trained) 과정을 통해 이전 단어(token)를 보고 다음 단어를 예측하는 방식으로 학습된다.

• GPT 버전에 따른 구조 차이는 〈그림 2〉와 같다.

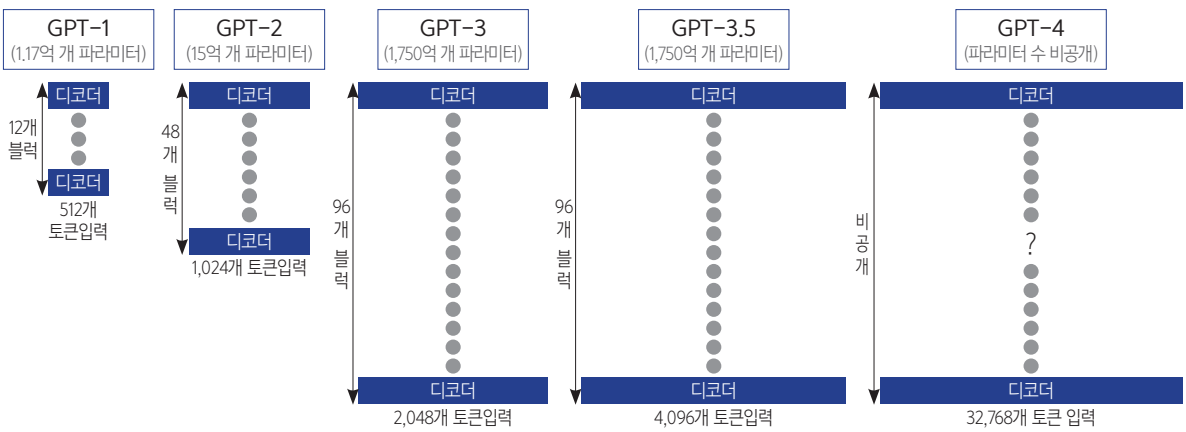

그림 2. GPT 버전에 따른 모델 구조 차이

• 또한, 〈그림 3〉에서 알 수 있듯이 GPT-3 이전의 모델들은 입력에 대한 다음 단어를 예측하지만, GPT-3.5 이후 버전의 모델들은 사용자의 질의 입력에 답변하는 챗봇 시스템으로 동작한다는 점에서 주요 차이점이 존재한다.

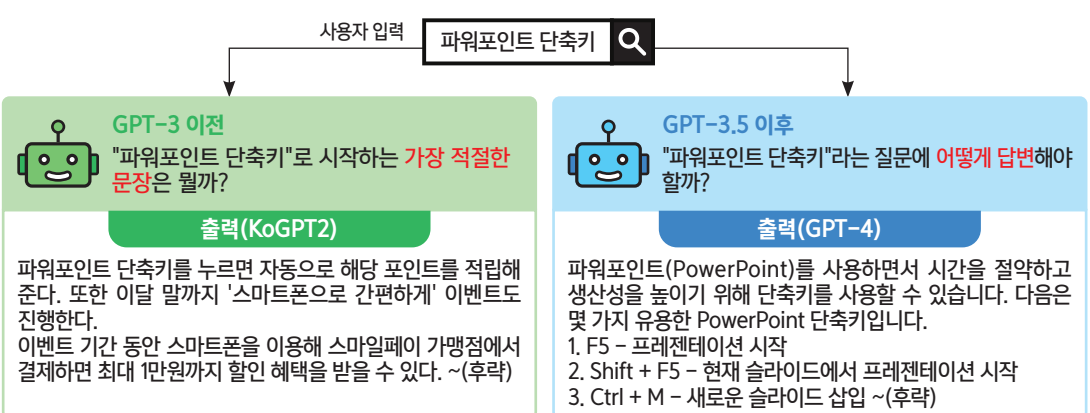

그림 3. GPT-3 이전 모델과 GPT-3.5 이후 모델의 동작 방식 차이

• 챗GPT 서비스는 〈그림 4〉와 같은 흐름으로 동작한다.

- 사용자가 입력한 데이터는 OpenAI 서버에 존재하는 비공개 모델인 GPT-3.5 또는 GPT-4에 입력되고, 해당 모델은 입력을 분석하고 적절한 응답을 생성한 후 이를 사용자에게 반환한다.

- 이러한 과정에서 사용자 입력과 서버 응답은 필요에 따라 챗GPT 확장 프로그램(extension) 또는 챗GPT 플러그인(plug-in)을 거쳐 사용자에게 출력될 수 있다.

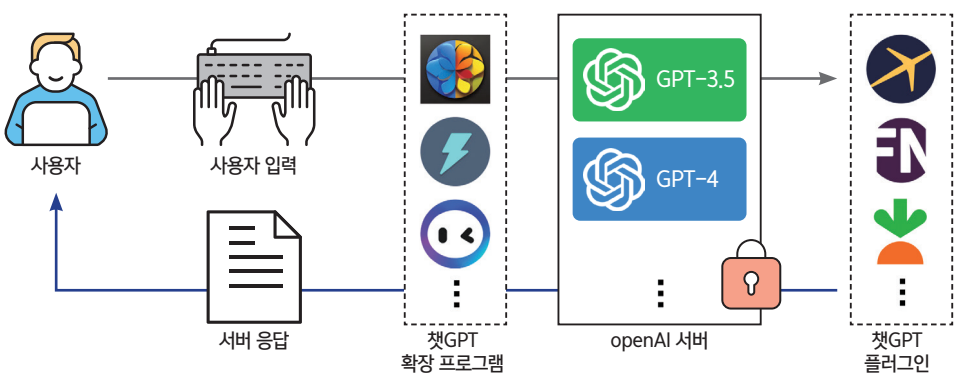

그림 4. 플러그인 및 확장 프로그램을 고려한 데이터 흐름도 관점의 챗GPT 서비스 동작 방식

• 챗GPT 플러그인은 챗GPT 서비스 이용 과정에서 최신 정보에 액세스하거나 계산을 실행하고, 타사 서비스와의 상호 작용을 가능하게 해 주는 역할을 수행한다.

• 챗GPT 플러그인은 브라우징(browsing), 코드 인터프리터(code interpreter), 검색(retrieval), 그리고 서드 파티(third-party) 플러그인 등 4가지 항목으로 구분된다. 이러한 플러그인들은 각각 다른 기능과 역할을 수행하여 챗GPT 서비스의 다양한 기능과 확장성을 제공한다.

- 브라우징 기능은 챗GPT가 웹에서 최신 정보를 검색하고 접근할 수 있도록 지원한다. 챗GPT는 검색 정보를 보여주며, 관련된 참조 링크도 함께 제공하여 사용자에게 비교적 최근의 데이터에 대한 정보를 제공한다.

- 코드 인터프리터 기능은 챗GPT가 파이썬(Python) 코드를 바로 실행할 수 있도록 지원한다. 이를 통해 챗GPT는 그래프를 그리거나, CSV 또는 엑셀 데이터 파일을

업로드하여 데이터 분석을 수행하며, 이미지를 포함하여 실시간으로 파일 편집이 가능하다.

- 챗GPT 플러그인 중 검색 항목은 챗GPT에 내장된 검색기능을 강화하고 확장하는 역할을 수행한다. 이 항목은 챗GPT가 웹에서 실시간으로 정보를 검색하고, 해당 정보를 사용자에게 제공할 수 있도록 한다. 검색 결과에는 요약된 내용, 관련 링크, 이미지, 비디오 등이 포함될 수 있다.

- 챗GPT는 외부 개발자가 특정 기능을 확장할 수 있도록 하기 위해 제한적으로 서드 파티 플러그인 개발을 허용하고 있다. 대표적인 서드 파티 플러그인은 〈표 3〉과 같다.

표 3. 챗GPT에서 제공하는 서드 파티 플러그인 일부

기관	설명
익스피디아(Expedia)	호텔 · 항공권 등 여행에 관한 온라인 예약 처리
피스컬노트(FiscalNote)	법률, 정치, 규제 데이터 및 정보에 관한 데이터 세트 제공
인스타카트(Instacart)	지역 식료품점에서 상품 주문 및 배달
카약(KAYAK)	항공편, 숙박, 렌터카 검색 및 여행지 추천
오픈테이블(OpenTable)	레스토랑 검색 및 예약
쇼피파이(Shopify)	온라인 쇼핑몰과 브랜드의 상품 검색 및 구매
슬랙(Slack)	클라우드 컴퓨팅 기반 팀 협업 툴
울프람(Wolfram)	간단한 수학 연산 수행 및 그래픽 결과 시뮬레이팅
자피어(Zapier)	웹 애플리케이션과 함께 자동화 워크플로우 제공

• 챗GPT 확장프로그램은 외부 프로그램에서 챗GPT의 기능을 사용할 수 있도록 제공하는 서비스로, 챗GPT에서 외부의 기능을 사용하도록 하는 플러그인과는 대조된다.

• 구글 크롬(Chrome) 웹 브라우저를 기준으로, 사용자는 크롬 웹 스토어에서 원하는 챗GPT 확장 프로그램을 설치할 수 있으며 이때 챗GPT의 확장 프로그램은 챗GPT를 서비스하는 OpenAI에서 직접 관리하지 않으므로, 확장 프로그램 악용사례에 유의해야 한다.

1.3. 해외 주요국 동향

- **(EU)** 유럽 연합은 2023년 5월 11일 인공지능에 대한 세계 최초의 규제 프레임워크인 '인공지능 법(AI Act)'제정의 첫 단계로서 해당 법 초안을 유럽의회 상임위원회에서 통과시켰다.
  - EU 인공지능 법에 따르면, 챗GPT 등 대규모 언어모델 및 생성 AI와 같은 이른바 '기초 모델(foundation models)' 제공업체 개발자는 모델을 공개하기 전에 안전 점검, 데이터 거버넌스 조치 및 위험 완화 조치를 적용해야 한다.
  - 시스템을 훈련하는데 사용되는 학습 데이터 셋을 공개하고 생성 AI가 만든 콘텐츠는 인간이 생성한 것이 아님을 밝혀야 한다는 조항을 추가하였다.
- **(미국-EU)** 미국과 유럽연합은 2023년 5월 31일 챗GPT와 같은 생성형 인공지능의 부작용을 줄이기 위한 '자발적 행동강령'을 마련하기로 하였다.
- **(미국)** 美 백악관 과학 기술 정책실(Office of Science and Technology Policy)은 2023년 8월 개최되는 데프콘(DEFCON)에서 챗GPT를 비롯한 생성형 인공지능 시스템을 공개 평가하여 잠재적인 취약점(potential harms)을 테스트한다고 밝혔다.
  - 여기서는, 대규모 언어모델을 포함한 다양한 생성 AI에 내재한 혼란(confabulations), 탈옥, 편견과 같은 위험을 발견하고, 기업 개발자가 발견된 문제를 해결하도록 장려하는 것을 목표로 한다.
- **(캐나다)** 캐나다 개인정보 보호 규제당국(Canada privacy regulators)은 2023년 5월 26일 챗GPT의 모기업인 OpenAI의 데이터 수집 및 사용에 관한 공동 조사를 시작하였다.
  - 연방 개인정보 보호 규제기관은 퀘벡, 브리티시 컬럼비아, 앨버타의 규제 기관과 함께 OpenAI가 챗GPT를 통해 사용자(residents)의 개인정보 수집, 사용 및 공개에 대한 동의를 얻었는지 여부를 조사할 것이라고 밝혔다.
- **(이탈리아)** 이탈리아 개인정보 감독기구(Italian Data-Protection Authority; Italian DPA)는 3월 31일, 챗GPT의 개인정보 수집, 처리 및 사용자 연령 확인 부재로 인해 개인정보 보호 규정(GDPR) 위반 사유로 판단하여 챗GPT의 사용을 금지하는 조치를 취했다.

- 그러나 이후 4월 28일, OpenAI가 DPA의 요구사항을 이행함에 따라 이탈리아 정부는 챗GPT의 접속 차단을 해제하였다.

- OpenAI는 가입 시 이탈리아에서 사용자의 나이를 확인할 수 있는 도구를 제공하고, 유럽 연합 사용자가 모델 학습을 위해 개인 데이터를 사용하는 것에 반대할 권리를 행사할 수 있는 새로운 양식을 제공하는 등 이탈리아 DPA에서 제기한 문제를 해결하거나 명확히 하였다.

• **(일본)** 일본 내각부는 2023년 4월 22일 OpenAI의 챗GPT와 유사한 생성형 인공지능 챗봇의 확산에 대응하기 위해 경제산업성, 총무성, 문부과학성, 디지털청 등 관계 부처가 참여하는 'AI 전략팀' 신설 계획을 발표하였다. 이 전략팀은 2023년 4월 24일 회의에서 AI의 업무 활용과 관련된 과제를 정리하고 부처 간 의사소통을 강화하는 방안을 논의한 바 있다.

2. 생성형 인공지능 기술의 대표적인 보안 위협

2.1. 개요

• 챗GPT와 같은 대규모 언어모델 등 생성형 AI 기술(이하 'AI 모델')의 보안 위협으로는 (1) 잘못된 정보, (2) AI 모델 악용, (3) 유사 AI 모델 서비스 빙자, (4) 데이터 유출, (5) 플러그인 취약점, (6) 확장 프로그램 취약점, (7) API 취약점 등이 존재한다(〈표 4〉 참조).

표 4. 대규모 언어모델 등 생성형 AI 기술의 대표적인 보안 위협

대표 보안 위협	주요 원인	가능한 보안 위협
잘못된 정보	• 편향 • 최신 데이터 학습 부족 • 환각 현상	• 사회적 혼란 조장 • 고위험 의사 결정 • 잘못된 의사 결정 유도
AI 모델 악용	• 적대적 시스템 메시지	• 피싱 이메일 및 인물 도용 • 사이버 보안 위협 코드 작성 • 대화형 서비스를 악용한 사이버 범죄 커뮤니티 활성화 • 사회 공학적 영향 • 가짜 뉴스 생성
유사 AI 모델 서비스 빙자	• 유사 악성 서비스 접근 유도	• 스쿼팅 URL 및 확장 프로그램 • 가짜 애플리케이션
데이터 유출	• 데이터 합성 과정의 문제 • 과도한 훈련 데이터 암기 문제 • 대화 과정에서 개인정보 및 민감 정보 작성	• 훈련 데이터 유출 • 데이터 불법 처리 우려 • 기밀 유출 • 대화 기록 유출 • 데이터베이스 해킹 및 회원 추론 공격
플러그인 취약점	• AI 모델의 적용 범위 확장 • 안정성 확인 미흡 • 해커 공격 범위 확장 • 취약점이 있는 서비스와 연결	• 새로운 도메인에서의 모델 오작동 • '에이전트'화 된 AI 모델의 악용 • 멀티모달 악용
확장 프로그램 취약점	• 확장 프로그램 내부의 악성 서비스 설치 • 서비스 제공 업체의 보안 조치 미흡	• 개인정보 수집 • 시스템 공격 • 호스팅 서버 및 스토리지 시스템 위협
API 취약점	• 미흡한 API 키 관리 • 데이터와 명령 사이의 불분명한 경계	• API 키 탈취 • 악의적인 프롬프트 주입

2.2. 잘못된 정보

• 위협의 주요 원인(〈그림 5〉 참조)

1) 편향

학습 데이터의 불균형 또는 데이터 내의 직 · 간접적인 편향은 AI 모델의 편향으로 이어질 수 있으며, 이는 모델이 만들어내는 결과물에 영향을 미쳐 사용자에게 사실과 다른 정보를 제공할 수 있다. 예를 들어, AI 모델이 편향된 텍스트 데이터를 학습한다면, 해당 모델이 특정 그룹이나 주제에 대해 편견이 있는 결과를 생성할 수 있다.

2) 최신 데이터 학습 부족

특정 생성형 AI 모델은 최종 학습 데이터 시점까지의 정보만 가지고 있기 때문에 학습 이후에 발생한 사건이나 정보에 대해서는 알지 못하거나, 부정확한 정보를 제공할 수 있다. 특히 AI 모델은 현재 진행 중인 사건이나 빠르게 변화하는 정보에 대해서는 정확한 답변을 생성하는데 제한이 있을 수 있다.

3) 환각 현상

환각(Hallucination) 현상은 AI 모델의 결과물이 정확한 것처럼 생성되었으나 실제로는 거짓인 경우를 말한다. 즉, AI 모델은 새로운 콘텐츠를 생성할 수 있는 능력으로 인해 잘못된 정보나 존재하지 않는 정보를 생성할 수 있다는 특징을 가지고 있다. 이러한 환각 현상은 AI 모델의 신뢰성을 저하시키는 원인이 될 수 있다.

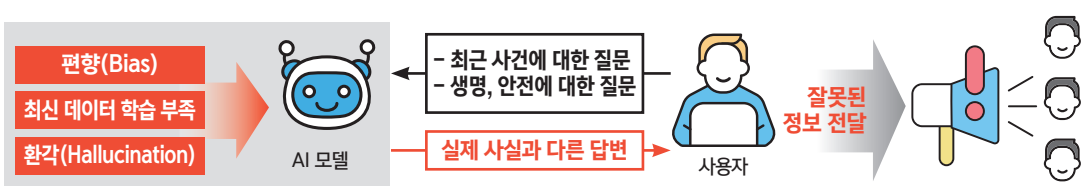

그림 5. AI 모델의 잘못된 정보 제공

• 대표 보안 위협

1) 사회적 혼란 조장

AI 모델의 악용 혹은 환각 현상 등에 따라 잘못된 결과물은 사회적 혼란을 조장하거나 해커에 의해 악용될 수 있다. 특정 집단을 대상으로 하는 혐오 결과물 생성 및 민감한 사회적 이슈에 대해 논란을 조장하는 결과물 생성 등이 예가 될 수 있다.

2) 고위험 의사 결정

많은 생성형 AI 서비스들은 고위험 정치적 의사 결정이나, 법률 또는 건강 관련 조언을 제공하는 경우 인공지능 서비스 활용을 제한하는 이용 정책을 수립하고 안내하고 있다. 또한, 「지능정보화 기본법 시행령」 제16조 2항에서는 “국민의 생명 또는 신체안전 등에 밀접한 지능정보기술”을 ‘지능정보기술을 이용하는 사람의 생명 · 신체를 보호하는데 현저한 지장을 줄 우려가 있는 지능정보기술’로 정의하고 있으므로, 대규모 언어모델 등 생성형 AI 사용에 충분한 주의가 필요하다.

3) 잘못된 의사 결정 유도

생성형 AI 서비스 사용시 검증되지 않은 결과물의 공유는 조직 · 개인의 잘못된 의사 결정을 유도할 수 있다.

2.3. AI 모델 악용

• 위협의 주요 원인

1) 적대적 시스템 메시지

AI 모델은 사용자의 요청에 따라 다양한 유형의 출력을 생성할 수 있다. 이러한 유연성은 해커가 모델에 적대적 시스템 메시지를 보내 유해한 답변을 생성하는 데 사용할 수 있다는 위험성을 내포하고 있다. 사용자에 요청에 대한 적절한 제약 조건 없이 AI 모델이 사용될 경우, 심각한 결과를 초래할 수 있다.

그림 6. AI 모델의 적대적 시스템 메시지 생성

• 대표 보안 위협

1) 피싱 이메일 및 말투 도용

AI 모델은 사용자가 제공하는 텍스트 스타일을 흉내 내는 데 매우 효과적이며, 이를 통해 피싱 이메일을 생성하거나, 특정 개인이나 단체의 말투를 도용하는 데 악용될 수 있다. 이렇게 생성된 정교한 피싱 이메일은 기존의 보안 시스템을 속이고, 사람들이 이를 진짜로 인식하게 만들 수 있다.

2) 사이버 보안 위협 코드 작성

AI 모델은 소프트웨어 코드를 생성할 수 있다. 이는 AI 모델이 악성 코드 작성에 이용될 수 있음을 의미한다. 해커는 이를 통해 보안 시스템을 회피하는 데 필요한 복잡한 악성 코드를 작성할 수 있으며, 이로 인해 정보 보안에 심각한 위협이 될 수 있다.

3) AI 모델을 악용한 사이버 범죄 커뮤니티 활성화

일부 사이버 범죄 커뮤니티는 AI 모델을 이용하여 악성 소프트웨어를 재현하거나, 범죄 행위에 대한 지침을 생성하고 있다. 이로 인해 이러한 커뮤니티의 활동이 활성화되고, 사이버 범죄의 위험성이 더욱 증가하게 된다.

4) 사회 공학적 영향

AI 모델은 사실적이고 믿기 쉬운 콘텐츠를 생성한다. 이를 통해 사용자는 다른 사람들을 조작하거나 사기를 칠 수 있는 잠재적인 도구를 얻을 수 있다. 이는 사회 공학적 공격을 더욱 쉽게 만들고, 사람들이 피해자가 될 위험성을 증가시킨다.

5) 가짜 뉴스 생성

생성형 AI 모델은 가짜 뉴스를 생성하는 데 사용될 수 있다. 특히, 챗GPT 등 대규모 언어모델은 인간이 생성하는 수준의 현실적이고 설득력 있는 결과물을 생성함으로써 허위 정보를 퍼뜨리는 데 악용이 가능하다.

2.4. 유사 AI 모델 서비스 빙자

• 위협의 주요 원인(〈그림 7〉 참조)

1) 유사 악성 서비스 접근 유도

해커들은 정상적인 AI 모델 서비스인 것처럼 모사된 악성 AI 서비스 URL 제공을 통해 악성코드를 내포한 악성 서비스를 퍼뜨릴 수 있다. 사용자는 반드시 공식 웹사이트를 통해서 AI 모델 서비스에 접근하여야 하며, 비정상적 경로를 통해 악의적 목적을 가진 유사 AI 서비스에 접근할 경우, 악성 확장 프로그램을 다운로드하는 등 피해를 볼 수 있다.

그림 7. 악성 AI 모델을 통한 사기 행위

• 대표 보안 위협

1) 스쿼팅 URL 및 확장 프로그램

해커는 신뢰할 수 있는 AI 모델의 이름을 도용하여 사용자의 실수로 스쿼팅 URL에 접속하거나, 악성 확장 프로그램을 설치하도록 유도할 수 있다. 이후 해커는 스쿼팅 URL 및 악성 확장 프로그램을 통해 사용자의 개인정보를 도용하거나 악성 코드를 전파할 수 있다.

* 스쿼팅 : 사이버 스쿼팅 또는 도메인 스쿼팅을 포함한 유사 이름 위협을 의미함 (예: 철자가 조금 다른 URL을 불법 취득하여 여기에 실수로 접속한 사용자를 위협하는 행위)

2) 가짜 애플리케이션

최근 챗GPT 등 생성형 AI 모델에 대한 관심이 증가함에 따라, 특정 서비스 이름을 도용한 가짜 애플리케이션도 증가하고 있다. 이러한 가짜 애플리케이션은 사용자의 개인정보를 도용하거나 사용자 기기에 악성코드를 설치하는 등의 악의적인 목적으로 활용될 수 있다. 이로 인해 대화내용 및 카드번호와 같은 민감한 정보가 유출되거나 악성코드에 감염될 수 있으므로 사용자는 이러한 위험에 주의를 기울여야 한다.

2.5. 데이터 유출

• 위협의 주요 원인(〈그림 8〉 참조)

1) 데이터 합성 과정의 문제

AI 모델 학습에 사용되는 매우 방대한 양의 훈련 데이터에는 기관 내 다양한 주요 정보가 포함될 수 있다. AI 모델은 여러 단계의 추론 과정을 거치면서 입력 데이터에 따라 가장 신뢰도가 높은 문자열을 합성하여 답변하는데, 합성된 문자열은 AI 모델이 자체 생성할 수도 있지만 예상치 못하게 기존 훈련 데이터의 부분 혹은 전체를 그대로 답변에 포함할 가능성이 있다.

2) 과도한 훈련 데이터 암기 문제

AI 모델이 훈련 데이터를 과도하게 암기하면 추론 과정에서 훈련 데이터에 대한 신뢰도가 더욱 높아지므로 내부 정보를 유출하는 경향이 커지게 된다. 훈련 데이터 내 민감정보와 개인정보의 비식별 처리를 했더라도 AI 모델의 출력문과 외부 정보와의 결합을 통해 비식별화된 정보를 추측할 수 있다.

3) 대화 과정에서 개인정보 및 민감 정보 작성

외부 AI 모델 서비스를 이용할 경우, 사용자는 내부 업무정보나 개인정보를 입력할 수 있으며 이는 다양한 보안 위험을 초래할 수 있다. 사용자가 입력한 정보는 AI 모델의 훈련 데이터로 사용될 수도 있으며, 다른 정보와 결합되어 타인의 요청에 따라 추론 작업에 활용되거나 해당 정보가 모델에 학습되어 타인과의 대화 과정에서 해당 중요 정보가 유출될 수 있다.

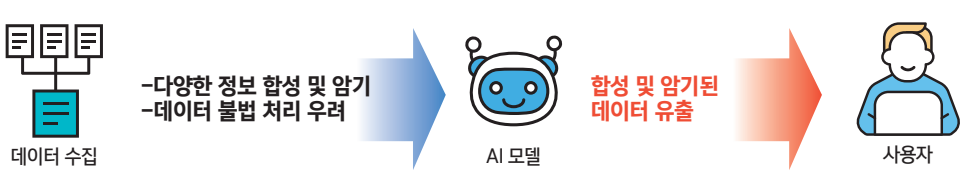

그림 8. 데이터 수집 및 AI 모델 자체에 의한 데이터 유출

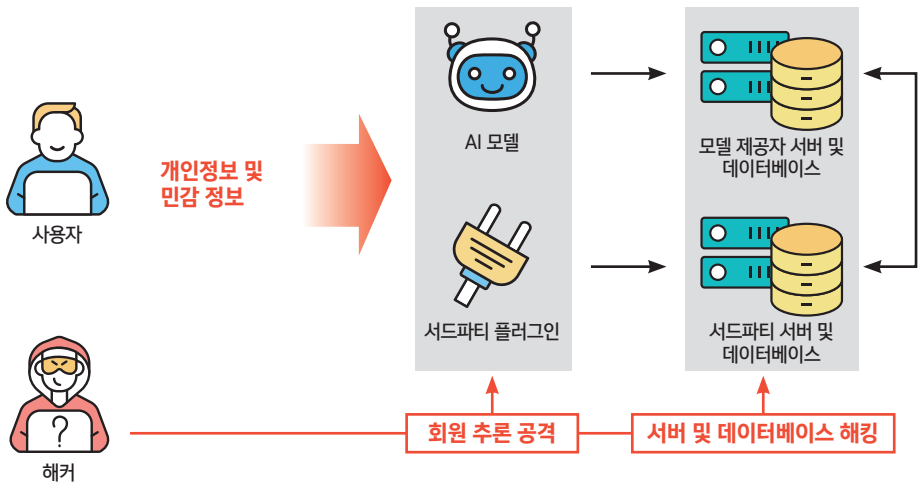

그림 9. 사용자의 민감 정보 입력 및 해커에 의한 데이터 유출

• 대표 보안 위협

1) 훈련 데이터 유출

AI 모델은 수집된 훈련 데이터를 일부 암기할 수 있으며, 이는 정보의 유출 위험을 가지고 있다. 중요한 URL, 개인정보, 연락처, API 키와 같은 민감한 정보들이 모델 내에서 데이터를 처리하고 응답을 생성하는 과정에서 외부에 노출될 수 있으며, 이로 인해 더 심각한 보안 위협 상황을 초래할 수 있다.

2) 데이터 불법 처리 우려

최근 국가들은 대규모 언어모델 기반 생성형 AI 기술과 관련, 데이터의 불법 처리에 대해 우려를 표명하고 있다. 이탈리아 정부를 비롯하여 일부 국가들은 AI 모델 개발 및 운영 회사들에게 데이터 처리에 대한 법적 준수와 개인정보 보호 규정(GDPR 등)을 엄격히 준수할 것을 요구하고 있으며, 이에 따라 일시적인 서비스 중단이나 조사를 요구하는 경우도 있다.

3) 기밀 유출

생성형 AI 서비스를 사용하는 과정에서 개인 또는 기관의 기밀 정보를 AI 모델에 제공하면, 이러한 정보가 다른 사용자에게 유출될 위험이 있다. 이는 사용기관의 중요한 내부 정보가 외부 공개되거나 개인정보가 노출될 수 있는 위험을 초래할 수 있다. 따라서

사용자는 AI 모델을 사용할 때 기관의 기밀 정보를 입력하지 않아야 하며, 기관 정보 보안 책임자는 이러한 정보 유출 방지를 위해 적절한 보안 대책 및 수단을 마련해야 한다.

4) 대화 기록 유출

일부 AI 서비스에서 타인의 대화 기록이 다른 사용자에게 보이는 버그 현상이 발생한 사례가 있었다. 이 사례에서는 대화 기록과 함께 금융 관련 정보, 이름, 이메일 결제 정보, 카드 정보, 카드 4자리 숫자 정보 등이 일반 사용자에게 노출되었다. 이처럼 AI 서비스 시스템 오류로 인해 개인 · 민감정보가 타인에게 노출될 수 있는 위협이 있다.

5) 데이터베이스 해킹 및 회원 추론 공격

대규모 언어모델 기반 AI 서비스를 이용하는 사용자의 질문은 데이터로 남고, 모델 혹은 서비스를 제공하는 자(회사)는 이를 확인할 수 있다. 또한 최악의 경우 해커가 해당 모델 혹은 서비스 제공자의 데이터베이스를 해킹하거나 모델에 대한 회원 추론(membership inference) 공격을 수행하여 기밀성(confidentiality)을 침해할 수 있다.

2.6. 플러그인 취약점

• 위협의 주요 원인(〈그림 10〉 참조)

1) AI 모델의 적용 범위 확장

AI 모델은 플러그인을 통해 다양한 데이터와 플랫폼과 결합되어 활용 범위가 확장되고 있다. 이를 통해 단순한 질문-답변 시스템을 넘어서 모델의 기능을 확장하거나 특정 작업을 수행하기 위한 추가적인 기능을 제공할 수 있다. 또한, 모델 입력 데이터의 전처리, 결과의 후처리, 외부 서비스 통합 등 다양한 기능을 수행할 수도 있으나 이러한 확장성은 동시에 새로운 취약점을 초래하여 보안위협이 증가될 수 있다.

2) 안정성 확인 미흡

AI 모델의 새로운 기능 도입 시에는 사전 해당 기능의 안정성을 검증하는 과정이 필요하다. 그러나 AI 모델의 빠른 기능 확장과 발전 속도로 인해 사전 안정성 검증이 충분히 이루어지지 않을 수 있다. 이는 AI 모델이 예상치 못한 방식으로 작동하거나 새로운 보안 위협에 노출될 위험이 있음을 의미한다.

3) 해커 공격 범위 확장

생성형 AI 모델의 플러그인 기능은 해커에 의해 새로운 공격 도구로 악용될 우려가 있다. 모델이 생성하는 다양한 컨텐츠와 제공 정보를 이용하여, 사기 · 도용 · 피싱 · 개인정보 유출 · 악성코드 작성 등 공격 행위를 고도화할 수 있으며, 플러그인과 취약 서비스에 존재하는 취약점을 공격하여 AI 모델과 연동된 다른 시스템의 보안을 위협할 수 있다.

4) 취약점이 있는 서비스와 연결

AI 모델이 플러그인을 통해 다른 서비스와 연결될 때, 해당 서비스에 존재하는 보안 취약점이 사용자에게 전파될 수 있다. 이는 AI 모델이 연결된 서비스의 취약점을 통해 해커가 AI서비스 사용자 시스템 내부에 접근하여 정보를 유출하거나, 시스템 오작동을 유발하여 악의적인 결과물 생성을 유도하여 사용자를 속이는 등의 위협이 존재할 수 있다.

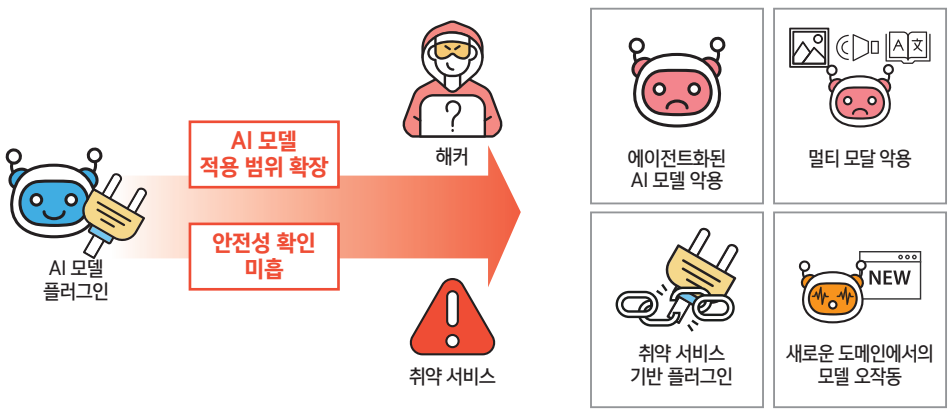

그림 10. 플러그인 취약점

• 대표 보안 위협

1) 새로운 도메인에서의 모델 오작동

생성형 AI 모델이 새로운 도메인에 적용될 때(예, 뉴스 기사를 학습했던 모델을 의학 분야에 적용하는 경우 등), 모델이 예상치 못한 실수를 하거나 잘못된 동작을 수행할 수 있다. 이로 인해 사용자에게 부정확한 정보를 제공하거나, 시스템의 안정성을 해치거나, 악용되어 부정적인 결과를 초래할 수 있다.

2) '에이전트'화된 AI 모델의 악용

사용자가 챗GPT 등 생성형 AI 서비스를 특정 업무를 지원하는 신뢰성 높은 시스템으로 간주하고 활용할 경우, 해커는 이 점을 악용하여 사용자의 AI 모델이 부정확하거나 위험한 정보를 생성하거나, 악성코드를 전파하는 등에 악용할 수 있다.

3) 멀티모달 악용

AI 모델은 플러그인을 통해 이미지와 음성 등 다양한 형태의 데이터를 처리할 수 있다. 이는 모델이 텍스트 정보뿐만 아니라 시각적이고 음성적인 콘텐츠를 이해하고 생성할 수 있는 능력을 갖고 있다는 것을 의미한다. 하지만 이로 인해 모델은 단순한 텍스트 정보뿐만 아니라 다양한 형태의 데이터를 악용하여 유해한 행동을 수행할 위험이 있다. 예를 들어, 모델이 악의적인 이미지를 생성하거나 음성 인증 시스템을 속일 수 있는 음성을 생성하는 등의 악용 가능성이 있다.

2.7. 확장 프로그램 취약점

• 위협의 주요 원인(〈그림 11〉 참조)

1) 확장 프로그램 내부의 악성 서비스 설치

AI 모델 확장 프로그램은 입력 데이터를 전처리하거나, 결과를 후처리하여 특정 요구사항에 맞게 가공하는 기능을 제공하는 등 사용자에게 높은 활용성과 편의성을 제공하려는 목적으로 개발되었다. 하지만 이러한 확장 프로그램은 해커에게 유용한 공격 수단으로도 사용될 수 있다. 해커는 확장 프로그램에 백도어(backdoor)를 설치하거나, 코드를 변조하여 원래 기능 외에도 개인정보 수집, 시스템 취약점 공격 등의 악용을 수행할 수 있다.

2) 서비스 제공업체의 보안 조치 미흡

생성형 AI 서비스 제공업체가 적절한 사이버 보안 대책을 적용하지 않거나 업데이트를 제대로 수행하지 않으면 해커나 악의적 사용자가 저장된 사용자 데이터에 무단으로 액세스함으로써 개인정보 유출, 데이터 변조 등의 위험이 발생할 수 있다. 또한, 악성 코드가 유입되어 전체 시스템에 악영향을 미칠 수 있으며 저장된 데이터 손실이 발생하여 주요 정보나 업무에 심각한 영향을 줄 수 있다.

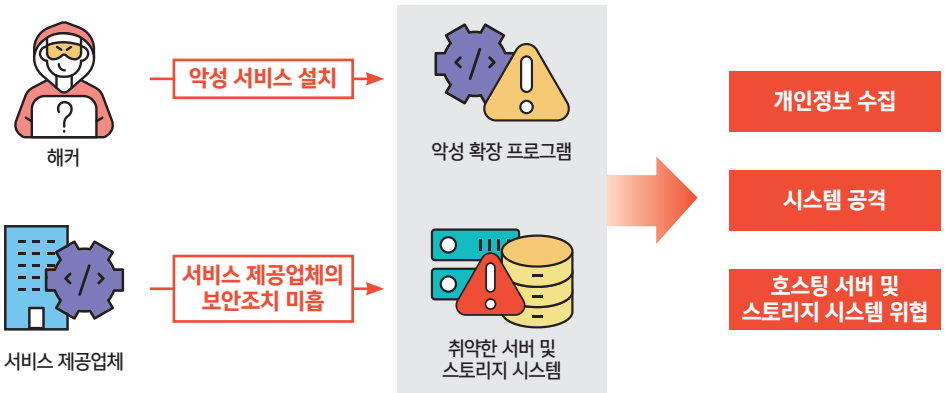

그림 11. 확장 프로그램 취약점

• 대표 보안 위협

1) 개인정보 수집

사용자가 AI 모델 확장 프로그램을 사용하는 동안 악성 플러그인이 설치되거나 사용자의 권한을 남용하여 부적절한 작업을 수행함으로써 발생할 수 있다. 이를 통해 사용자의 모든 행동이 무단으로 모니터링될 수 있으며, 사용자의 개인정보가 침해될 수 있다. 수집된 정보는 악의적인 목적으로 사용될 수 있으며, 예를 들어 사용자의 신원 도용, 금융 손실 등의 사이버 범죄에 악용될 수 있다.

2) 시스템 공격

취약한 확장 프로그램을 악용하는 해커는 시스템을 공격하는데 확장 프로그램을 이용할 수 있다. 예를 들어, 해커는 악의적인 확장 프로그램을 설치한 사용자를 대상으로 랜섬웨어 공격을 수행할 수 있으며, 사용자 시스템을 좀비 PC로 만들어 DDoS 공격을 발생시키는 데 악용할 수 있다.

3) 호스팅 서버 및 스토리지 시스템 위협

AI 모델 확장 프로그램은 안전한 서버와 스토리지 시스템에서 호스팅 되어야 한다. 호스팅 및 스토리지 제공업체가 적절한 사이버 보안 조치를 적용하지 않거나, 지속해서 이를 업데이트하지 않을 경우 해커의 호스팅 서버 침입으로 인한 내부 주요 데이터 유출, 서버 마비, 서비스 중단 등이 발생할 수 있으며 호스팅 서버의 자원을 악용하여 다른 침해 행위에 악용될 수 있다.

2.8. API 취약점

• 위협의 주요 원인(〈그림 12〉 참조)

1) 미흡한 API 키 관리

API 키는 소프트웨어 개발자가 서드파티 애플리케이션과 서비스 간에 통신을 수행하기 위해 사용되는 인증 수단이다. 미흡한 API 키 관리는 취약점을 노출시킬 수 있으며, 데이터 및 시스템 보안을 위협할 수 있다.

2) 데이터와 명령 사이의 불분명한 경계

AI 모델은 일반적인 프로그래밍 언어와는 달리 데이터와 명령을 명확하게 구분하지 않는 특성을 가지고 있다. 이로 인해 AI 모델은 입력된 데이터를 명령으로 오인하거나 명령을 잘못 이해하여 예상하지 못한 응답을 생성할 수 있다. 예를 들어 해커는 공격 명령이나 악의적 의도를 숨긴 데이터를 입력으로 제공하여 AI 모델이 부적절한 응답을 생성하도록 할 수 있다.

그림 12. API 취약점

대표 보안 위협

1) API 키 탈취

AI 모델 API는 인증을 위해 API 키를 사용하며, 이 키를 적절하게 관리하지 않으면 다양한 보안 위협이 발생할 수 있다. 우선, API 키가 노출되면 악의적인 사용자가 해당 키를 이용하여 민감한 정보에 접근하거나 비인가된 작업을 수행할 수 있다. 또한, API 키가 보호되지 못한다면 인증 및 권한 부여 프로세스의 취약점으로 악용되어 인가되지 않은 액세스 또는 권한을 탈취한 공격자의 시스템 침입 등이 이루어질 수 있다.

2) 악의적인 프롬프트 주입

데이터와 명령 사이에 명확한 경계가 존재하지 않는 AI 모델의 특성으로 인해 발생하며, 해커가 악의적인 목적으로 AI 모델의 API를 호출할 때 제어할 수 없는 프롬프트를 주입하여 민감한 정보를 탈취할 수 있다. 이를 통해 공격자는 API 키를 탈취하거나 민감한 정보에 접근할 수 있다. 이 공격은 다양한 형태로 이루어질 수 있는데 예를 들어 해커는 프롬프트에 악의적인 코드나 명령을 삽입함으로써 AI 모델이 예상치 않은 작업을 수행하도록 유도할 수도 있다.

3. 안전한 생성형 인공지능 기술 사용 가이드라인

3.1. 개요

- 본 장에서는 챗GPT와 같은 대규모 언어모델 등 생성형 AI 기술(이하 'AI 모델')의 안전한 사용을 위한 기본 지침을 제공하여 각급기관의 정보화 · 정보보안 담당자들이 보다 효율적이고 안전하게 서비스를 이용하도록 하는데 그 목적이 있다.

- 본 가이드라인을 참고하여 관련 서비스 잠재적인 위험을 최소화하고, 올바른 활용을 통해 업무의 효율성을 높일 수 있을 것으로 기대한다.

- 본 가이드라인은 관련 기술을 효율적이고 안전하게 사용하는 방법을 안내하기 위해 FAQ 형식으로 구성하였으며, (1) 서비스 사용 주의사항, (2) 서비스와의 대화 시 주의사항, (3) AI 모델 플러그인 사용 주의사항, (4) AI 모델 확장 프로그램 사용 주의사항, (5) AI 모델 생성기반 공격 대처 방안을 포함한다. 전체적인 구성은 〈표 5〉와 같다.

표 5. AI 모델 및 관련 서비스 사용 가이드라인 구성

가이드라인 주제	포함 내용
서비스 사용 주의사항	• 서비스 접근 • 계정 관리법
서비스와의 대화 시 주의사항	• 답변 검증 (정확성 및 유해성) • 개인정보 및 민감 정보 처리 • AI 모델이 생성한 데이터 관리 • 책임감 있는 사용 • 업무에서 올바르게 AI 모델 활용하기 (이용목적 및 협업)
AI 모델 플러그인 사용 주의사항	• 올바른 서비스 플러그인 사용 및 관리 • 개인정보 및 민감 정보 처리
AI 모델 확장 프로그램 사용 주의사항	• 올바른 서비스 확장 프로그램 사용 및 관리 • 개인정보 및 민감 정보 처리
AI 모델 생성기반 공격 대처 방안	• AI 모델 생성기반 공격 정의 • AI 모델 생성기반 공격 대처

3.2. 서비스 사용 주의사항

• 서비스 접근

Q 서비스에 어떻게 접근해야 하는가?

A 서비스 접근은 AI 모델 사용의 첫 단계로 중요한 보안 요소이다. 안전한 사용을 위해 다음 사항을 준수해야 한다.

공식 사이트 사용	» 서비스 사용 시 공식 사이트를 통한 접속 » 피싱 공격 및 허위 사이트 방지를 위한 SSL 인증서 유무 확인
안전한 네트워크 환경	» 암호화된 Wi-Fi 및 VPN 사용 등을 통한 안전한 네트워크 환경 확보 » 공공 Wi-Fi 등 암호화되지 않은 네트워크 사용 시 보안에 유의
웹 브라우저 보안	» 최신 웹 브라우저 사용 및 정기 업데이트 수행 » 브라우저 보안 설정 검토 및 광고, 악성코드 차단 프로그램 사용
내부망 사용 제한	» 기관 내부(업무)망에서의 서비스 사용시 별도 내부 전용 서비스로 한정하여 사용 » 기관 인터넷망에서의 서비스 사용시 기관 민감데이터 미활용 등 유의 사용
적절한 버전 사용	» 최신 버전 서비스 이용을 통한 보안 취약점 해소

• 계정 관리법

Q 계정은 어떻게 관리해야 하는가?

A 서비스를 사용하기 위해서는 계정 보안을 강화하는 것이 중요하다. 계정 보안을 위해 다음과 같은 보호 조치를 취해야 한다.

강력한 비밀번호 설정 및 관리	» 강력한 비밀번호 설정(충분한 길이 및 대소문자, 숫자, 특수문자 혼합) » 정기적인 비밀번호 변경 및 재사용 금지
이메일 인증	» 이메일 인증을 통한 본인 인증 완료
이중 인증 (2FA) 설정	» 이용 로그인 계정의 이중 인증(2FA) 설정을 통한 계정 침입 방지

3.3. 서비스와의 대화 시 주의사항

• 답변 검증 (정확성 및 유해성)

Q1 서비스 답변의 정확성은 어떻게 검증해야 하는가?

A1 인공지능 기반 예측형 서비스는 항상 정확한 정보를 제공한다고 보장할 수 없다. 따라서, 중요한 결정이나 작업을 수행하기 전에 다른 출처를 참조하거나 검색 등 다른 방법을 활용하여 정보의 정확성을 확인 및 점검하는 것이 필요하다.

다중 출처 비교	» 인터넷 검색, 전문가 의견 자문 등 다양한 출처 참조를 통한 AI 모델 답변 정확성 검증
공식 문서 및 정책 참조	» 공식 문서 및 정책 참조를 통한 AI 모델 답변 신뢰성 검증
최신 데이터 확인	» 서비스 결과 데이터에 대한 최신 여부 확인을 통한 AI 모델 답변 유효성 검증
답변 현실성 검토	» 사용자의 상식 및 기본 지식 검토를 통한 AI 모델 답변 현실성 검증

Q2 서비스 답변의 유해성은 어떻게 피할 수 있는가?

A2 사용자는 서비스가 생성한 결과물을 항상 주의 깊게 검토하고, 부적절한 내용이나 유해한 요소가 포함되어 있는지 확인해야 한다. 또한, 관련 규정을 준수하고 있는지에 대해서도 주의를 기울여야 한다.

결과물 검토	» AI 모델 답변 검토를 통한 유해성 차단
사용자 정의 필터 설정	» 사용자 정의 필터 설정을 통한 사용자 기준에 맞는 답변 생성
서비스 제공자의 필터링 기능 활용	» 서비스 제공자의 필터링 기능 활용을 통한 유해 답변 필터링 및 안전한 사용 환경 유지
사용자 신고 기능 활용	» 사용자 신고 기능 활용을 통한 유해한 서비스 학습 데이터 및 알고리즘 개선 참여
교육 및 인식 향상	» 기관 차원의 사용자 교육 및 인식 향상을 통한 신중한 서비스 사용 당부 및 유해 답변 전파 차단
법률 및 규정 숙지	» 법률 및 규정 숙지를 통해 서비스 사용 시 법적 문제 발생 방지
사용 목적 검토	» 서비스 사용 목적 적절성 검토를 통한 올바른 사용
지적 재산권 존중	» 타인의 저작물 인용 및 사용 시 출처 표기 및 저작자 동의 선행 » 부적절한 방식으로 타인의 저작물 사용 및 수정 금지

• 개인정보 및 각급기관 내 민감 정보 처리

Q 서비스와 대화 시 개인정보 및 소속기관 내 민감 정보는 어떻게 처리해야 하는가?

A 서비스와 대화 시 주민등록번호, 신용카드 정보, 비밀번호 등의 개인정보뿐만 아니라 소속기관의 내부 민감 정보를 입력하지 않도록 주의해야 한다. 기술 취약점으로 인해 개인정보 침해, 금융적 손실뿐만 아니라 각급기관의 내부 주요 정보가 유출될 위험이 있기 때문이다.

데이터 제어 설정	» 데이터 제어 설정을 통한 채팅 기록 및 모델 학습 비활성화
개인정보 입력 금지	» 개인정보 침해 및 악용 방지를 위한 민감한 개인정보(건강, 종교, 정치적 성향 등) 입력 금지 » 금융적 손실 및 신용도 피해 방지를 위한 금융 관련 정보(신용카드 번호, 은행 계좌 정보, 비밀번호 등) 입력 금지
기관 내 민감 정보 입력 금지	» 기관 내부자료 등 민감정보 입력 금지로 기관 정보 유출 차단 » 가명화 및 익명화를 통한 실제 개인정보와 기관과의 관계 유추 가능성 차단
보안 질문 회피	» AI 모델과 대화 중 보안 관련 질문 발생 시 답변 자제 » 필요시 고객 지원 센터 문의를 통한 올바른 절차 준수
인증 정보 입력 금지	» 계정 침해 위협 방지를 위한 인증 정보(사용자 이름, 비밀번호, 인증 코드 등) 입력 금지

• 서비스가 생성한 데이터 관리

Q 서비스가 생성한 데이터는 어떻게 관리해야 하는가?

A 서비스 제공자의 데이터 취급 정책을 준수해야 한다. 데이터 보관 기간이 지난 후에는 서비스 제공자가 제공하는 방법을 통해 데이터를 삭제해야 한다.

데이터 보관, 삭제 정책 이해 및 준수	» 올바른 데이터 관리 및 보호를 위한 서비스 제공자의 데이터 보관 및 삭제 정책 숙지 » 불필요한 데이터 누출 위협 감소를 위한 데이터 보관 기간 만료 후 서비스 제공자가 제공하는 방법에 따라 데이터 삭제
데이터 관리 및 보안	» 데이터 보관시 암호화와 접근 제어 등의 보안 메커니즘 사용 여부 확인 » 데이터 백업과 복원 방안 마련 여부 확인 » 데이터 보관 및 관리시 관련 법적 규정 및 개인정보보호법 등의 규정 준수 여부 확인 » 데이터 사용 및 이동에 대한 로그 기록 및 감사 추적 기능 제공 여부 확인 » 데이터 유출 등 침해사고 발생 시 대응 방안 수립 여부 및 대응 매뉴얼 존재 여부 확인

• 책임감 있는 서비스 사용

Q 서비스를 책임감 있게 사용하는 방법은 무엇인가?

A 서비스는 많은 유용한 기능을 제공하지만, 사용자는 책임감 있는 방식으로 서비스를 사용해야 한다. 사기, 디도스 공격(DDoS), 스팸, 협박 등 악의적 행위를 목적으로 이 서비스를 사용하지 말아야 한다.

윤리적 사용	» 서비스 사용 시 윤리적 가치 고려 및 상호 존중 » 서비스 사용시 윤리적 가치 고려를 통해 긍정적 사용자 경험 및 유익 결과 지향
악의적 행위 금지	» 악의적 행위(사기, DDoS. 스팸, 협박 등) 저지를 통한 법적 문제 방지 및 서비스 이용 경험 보호
서비스 약관 준수	» 서비스 제공자의 이용 약관 숙지 및 준수

• 업무에서의 올바른 서비스 활용(이용목적 및 협업)

Q 서비스의 이용목적을 어떻게 구별할 수 있는가?

A 서비스를 사용할 때 개인적인 용도 및 업무 용도를 구분해야 한다. 업무용으로 사용 시 소속 기관의 보안 정책을 준수해야 하며, 개인정보 및 기관 내부 정보활용을 금지하여야 한다.

서비스 이용 목적 명확화	» AI 모델 사용시 개인적 이용과 업무용 이용 구분 필요
계정 분리	» 사적 이용과 업무용 이용 별도의 계정 사용 권장

3.4. 서비스 플러그인 사용 주의사항

• 올바른 서비스 플러그인 사용 및 관리

Q1 안전한 서비스 플러그인을 사용하려면 어떻게 해야 하는가?

A1 서비스 플러그인을 사용할 때는 반드시 정식 배포 이후에 사용해야 한다. 또한, 플러그인 제공 업체의 신뢰성을 확인해야 하고, 악성 플러그인 또는 취약 프로그램과 연결된 플러그인 사용을 주의해야 한다.

정식 배포된 플러그인 사용	» 신뢰 가능한 소스에서 플러그인 다운로드 및 설치 » 정식 배포 플러그인을 통한 품질 보증 및 안전성 확보
플러그인 제공 업체 신뢰성 확인	» 플러그인 제공업체 신뢰성 확인 필요(보안 조치 및 지속 업데이트 여부, 고객 지원 종류 등) » 신뢰할 수 없는 업체 플러그인 사용 시 데이터 노출 가능성 존재
악성 플러그인 주의	» 정상 플러그인 여부 확인 후 사용 » 외부 데이터에 접근 가능한 플러그인의 악의적 활용 주의
취약 프로그램과 연결된 플러그인 주의	» 취약 서비스 기반 플러그인 사용 금지 » 연결 서비스 취약점 및 업데이트 버전 수정 내역 확인

Q2 서비스 플러그인을 어떻게 관리해야 하는가?

A2 정기적으로 플러그인을 업데이트하고 사용하지 않거나 필요하지 않은 플러그인은 제거해야 한다. 또한, 플러그인을 설치하기 전, 해당 플러그인이 사용 중인 웹 브라우저 및 시스템과 호환되는지를 확인하고, 정상적인 플러그인이더라도 어떤 권한을 가지고 있는지, 어떤 데이터에 접근하는지를 수시 확인해야 한다.

업데이트 알림 활성화 및 주기적 확인	» 업데이트 알림 활성화 및 업데이트 정보 주기적 확인을 통한 최신 버전 유지
사용 빈도/성능 저하 점검 및 삭제	» 사용 빈도 낮은 플러그인 제거 » 성능 저하 플러그인 제거
플러그인 권한 확인	» 플러그인 권한 검토 및 부적절한 권한 거부
접근 데이터 확인	» 플러그인 접근 데이터 주기적 확인

• 개인정보 및 민감 정보 처리

Q 서비스 플러그인을 사용할 때 개인정보나 민감 정보는 어떻게 처리해야 하는가?

A 플러그인을 통해 서비스가 외부 데이터에 접근할 수 있으므로, 플러그인이 접근하는 데이터를 보호하여야 한다. 또한, 플러그인이 데이터를 어떻게 처리하는지 이해하고, 개인정보 보호를 위한 적절한 조치를 취해야 한다.

개인정보 및 민감 정보 보호 조치	» 플러그인 접근 데이터에 대한 가명화 및 익명화 » 플러그인 접근 데이터에 대한 인증 및 권한 관리 설정
플러그인의 데이터 처리 이해	» 플러그인 제작사의 데이터 처리 방식 이해

3.5. 서비스 확장 프로그램 사용 주의사항

• 올바른 서비스 확장 프로그램 사용 및 관리

Q1 안전한 서비스 확장 프로그램을 사용하려면 어떻게 해야 하는가?

A1 서비스 확장 프로그램을 사용할 때는 반드시 제작자를 확인하고, 공식 출처에서 다운로드 받아야 한다. 이때 요구되는 권한을 주의 깊게 검토해 불필요한 권한을 요구하는 확장 프로그램은 개인정보 유출이나 데이터 손상의 위험이 있으므로 사용을 피해야 한다. 또한, 다른 사용자들의 확장 프로그램 사용 후기와 평가를 참고하여 플러그인의 안전성과 효과를 확인할 필요가 있다.

신뢰할 수 있는 출처에서 다운로드	» 신뢰할 수 있는 출처(공식 웹 사이트, 검증된 다운로드 사이트 등)를 사용한 확장 프로그램 다운로드 » 불법 사이트 및 알 수 없는 출처의 확장 프로그램 설치 금지
권한 목록 확인 및 불필요한 권한 거부	» 확장 프로그램 요구 권한 목록 확인 및 불필요한 권한 거부
사용자 리뷰 및 전문가 평가 확인	» 사용자 리뷰 및 전문가 평가 확인을 통한 확장 프로그램 신뢰성 검증
보안 소프트웨어와 함께 사용	» 보안 소프트웨어를 통한 악성 확장 프로그램에 대한 경고 기능 설정

Q2 서비스 확장 프로그램을 어떻게 관리해야 하는가?

A2 정기적인 업데이트와 제거가 필요하다. 확장 프로그램 업데이트를 통해 각종 보안 취약점이 업데이트될 수 있고, 불필요한 확장 프로그램은 시스템 성능 저하 및 보안 취약점을 초래할 수 있다.

업데이트 알림 활성화 및 주기적 확인	» 업데이트 알림 활성화 및 주기적 확인을 통한 최신 버전 사용
사용 빈도/성능 저하 점검 및 삭제	» 사용 빈도 낮은 확장 프로그램 제거 » 시스템 성능 저하 확장 프로그램 제거

• 개인정보 및 민감 정보 처리

Q 서비스 확장 프로그램을 사용할 때 개인정보나 민감 정보는 어떻게 처리해야 하는가?

A 확장 프로그램은 사용자의 데이터에 접근할 수 있으므로, 데이터 보호 측면에서 확장 프로그램을 사용할 때 개인정보나 민감 정보를 공유하지 않아야 한다. 또한, 확장 프로그램이 데이터를 어떻게 처리하는지 이해하고, 개인정보 보호를 위한 적절한 조치를 취해야 한다.

데이터 처리 방식 확인	» 확장 프로그램 제작자의 개인정보 처리 방침을 통한 데이터 저장 및 전송 방식, 보안 정책 확인
데이터 공유 주의	» 확장 프로그램의 데이터 접근 권한 제한 » 개인정보 및 민감 정보 수집 확장 프로그램 사용 금지
개인정보 보호 조치	» 확장 프로그램 제작자의 보안 기능 확인 » 필요시 데이터 가명화, 익명화, 인증 및 권한 관리 설정
데이터 처리 방침 준수	» 개인정보 및 민감 정보 관련 법률 및 규정 준수 » 적절한 데이터 보호기술 사용 » 필요시 데이터 가명화, 익명화 적용

3.6. AI 모델 생성기반 공격 대처 방안

• AI 모델 생성기반 공격 정의

Q AI 모델 생성기반 공격이란 무엇인가?

A AI 모델은 생성기반 공격에 사용될 수 있다. 생성기반 공격은 인공지능 모델이 생성한 텍스트를 이용하여 악의적인 목적을 달성하는 공격이다. 이러한 공격은 편견, 혼란 및 불신 조장, 피싱 및 사기를 통한 개인정보 및 기밀정보 탈취, 사람들의 신념을 조작하거나 특정 의견을 전파하는 사회적 공학 공격 등 다양한 목적으로 수행되어 정부, 기업 및 개인에게 심각한 영향을 미칠 수 있다.

• AI 모델 생성기반 공격 대처

Q AI 모델 생성기반 공격에 어떻게 대처해야 하는가?

A 사용자들은 AI 모델 생성기반 공격에 대해 비판적으로 접근하고, 편견이나 허위 정보를 퍼뜨리는 목적으로 사용되는 콘텐츠를 구별할 수 있어야 한다. 또한, 다양한 도구를 사용하고 보안 전문가와 인공지능 개발자들이 함께 시스템을 구축하여 이러한 공격에 대응해야 한다.

인공지능 생성 콘텐츠 특징 파악	» AI 모델 생성 콘텐츠 특성(급격한 주제 전환, 논리적 모순 등) 이해를 통한 사람 작성 콘텐츠와 구별
인공지능 생성 콘텐츠 감지 도구 사용	» 인공지능 생성 콘텐츠 감지 도구(GPTZero, DetectGPT 등)를 사용한 AI 모델 생성 여부 구별
공격 대응체계 구축	» 생성기반 공격 대응체계 구축을 통한 실시간 공격 감지 및 대응 » 인공지능 기반 보안 솔루션 도입 및 수동 감시

인공지능 생성 콘텐츠 위험 인지	» 사실 여부 확인 및 객관적 분석을 통한 비판적 콘텐츠 수용 » 감정적 자극 유도 및 소셜 미디어 서비스 콘텐츠에 대한 비판적 자세 견지
보안 정책 및 교육 강화	» 생성기반 공격 관련 정책 수립 » 기관 차원의 생성기반 공격 교육을 통한 사용자 인식 향상

4. 생성형 인공지능 기반 정보화사업 구축 방안 및 보안 대책

4.1. 개요

- 본 장에서는 각급기관에서 OpenAI 챗GPT 등 생성형 대규모 언어모델(이하 'AI 모델')을 기반으로 안전한 서비스를 구축 · 활용하기 위한 가이드라인을 제공하고자 한다. 이를 통해 관련 API 사용과 서비스 구축 및 운영 과정에서 발생하는 보안 위협을 최소화하도록 한다.

- AI 모델은 의료, 복지, 금융, 교육 등 다양한 분야에서 활용될 수 있으며, 대국민 서비스 또는 각급기관의 효율적 업무를 위한 내부 업무 시스템 등의 형태로 구축될 수 있다. 본 가이드라인은 이러한 AI 모델 기술 활용간 국내 · 외 민간 업체에서 개발된 AI 모델을 활용하거나, 자체 AI 모델을 구축하는 경우에도 적용될 수 있다(〈그림 13〉 참조).

- 본 가이드라인은 (1) 구축 유형에 따른 AI 모델 도입 시 고려사항, (2) AI 모델 API 활용 및 민간 AI 모델 도입 방안, (3) 자체 데이터 세트 및 AI 모델 구축 단계별 보안 위협 대응 방안을 다루고 있다. 이외 사항은 「국가 정보보안 기본지침」 등 관련 규정을 준수하여야 한다.

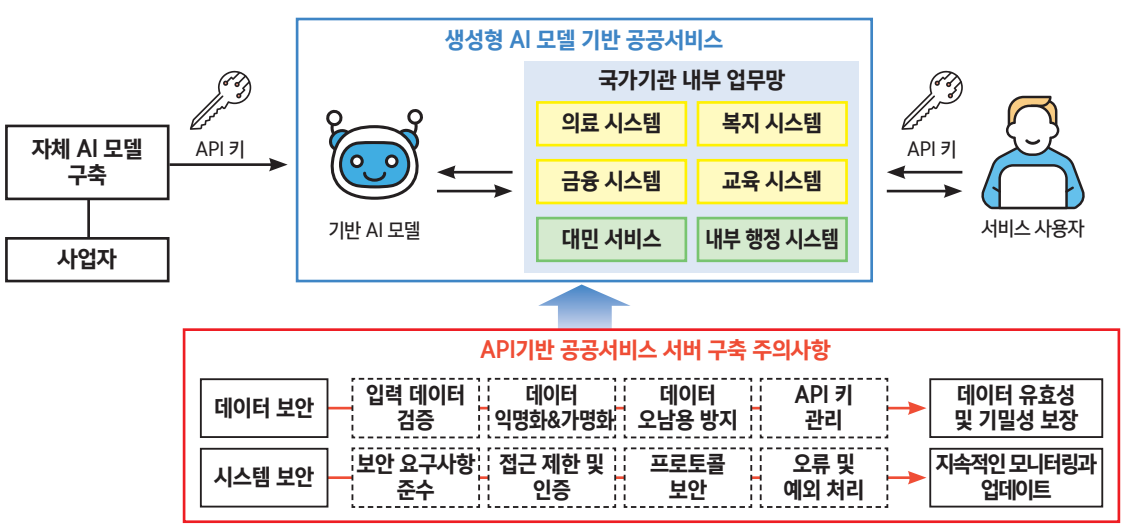

그림 13. 생성형 인공지능 API 기반 공공서비스 구축 시 주의사항

4.2. 구축 유형에 따른 생성 AI 기술 도입 시 고려사항

(가) 내부 행정 업무 활용 시스템 도입 (내부망 구축 시)

【기본 고려사항】

• AI 모델을 기관 내부망에 구축하여 각급기관의 행정 업무 등에 활용할 경우 각급 기관의 비공개 데이터를 비롯한 민감 정보가 AI 모델을 통해 처리되기 때문에 신중한 구축 방안 수립이 필요하다.

• AI 모델을 적용 · 활용하기 위한 내부 업무 시스템은 인터넷 등 외부망과 분리된 상태로 운영되어야 한다.

【기관 자체 구축·정부 자원 활용 시】

• 기관에서 생성형 AI 모델 기반 인공지능 서비스를 개발하기 위해서 정부 기관 소유의 정보시스템을 통해 추진할 수 있으며, 기관 자체의 AI 모델을 학습 · 강화 · 가공 등 지속 개발 · 서비스할 수 있다.

  - 이 경우 숏 절차에서 AI 모델 학습에 이용되는 기관 데이터 및 사용자 질의문 · 생성물 등 개발 · 운용 시 발생하는 데이터는 기관 외부로 직접적으로 이동할 수 없다. 다만, 보안성이 담보된 안전한 망연계시스템 활용 등 망분리 관련 보안대책을 준수하는 경우 해당 기관 외부로 데이터가 전달될 수 있다.

【클라우드컴퓨팅 기술·환경 활용 시】

• 내부 행정업무를 클라우드컴퓨팅 기술을 활용해 운용하는 기관의 경우, 「국가 클라우드 컴퓨팅 보안 가이드라인」의 보안등급 구분 및 보안대책을 준수하여 개발 · 운영할 수 있다.

  - 이때, 인터넷을 기반으로 서비스되는 관련 기술은 사용될 수 없으며, 기관 내부 행정망 이외의 네트워크에 연결되어 있거나 물리적 영역 분리가 이루어지지 않은 경우 이용할 수 없다.

【국가정보원 보안성 검토 절차 준수】

• 국가정보보안기본지침」 제15조 제1항 제19호에 의거, 첨단 정보통신기술을 활용한 정보화사업을 추진하기 위해서는 국가정보원의 사전 보안성 검토 절차를 준수하여야 한다.

(나) 대민서비스 등 외부 업무 활용 시스템 도입(외부·전용망 구축 시)

【기본 고려사항】

• AI 모델 기반 AI 서비스를 기관망 외부 · 전용망에 구축하여 대민서비스 및 홈페이지 등 외부 공개 업무에 활용할 경우 AI 모델 학습에 필요한 기관 데이터 유출이 야기될 수 있으므로 충분한 주의가 필요하다.

【기관 자체 구축·정부 자원 활용 시】

• 기관에서 생성형 AI 모델 기반 인공지능 서비스를 개발하기 위해서 정부 기관 소유의 정보시스템을 통해 추진할 수 있으며, 기관 자체의 AI 모델을 학습 · 강화 · 가공 등 지속 개발 · 서비스할 수 있다.

  – 이때 정부 기관 소유 정보시스템상에서 상용 AI 모델을 도입하여 관련 서비스를 개발하고자 하는 경우, 학습에 이용되는 데이터 및 사용자 질의문 등 개발 · 운용 시 발생하는 모든 데이터에 대한 기관 자체 보안등급 분류를 통해 상용 AI 모델을 도입할 수 있는지 여부를 판단하여야 하며, 개발 이후에도 주기적 데이터 등급 지정 · 점검을 통해 철저한 관리가 필요하다.

【클라우드컴퓨팅 기술·환경 활용 시】

• 클라우드컴퓨팅 기술을 활용해 관련 서비스를 개발하려는 기관의 경우, 「국가 클라우드 컴퓨팅 보안 가이드라인」의 보안등급 구분 및 보안대책을 준수하여 개발 · 운영할 수 있다.

• 클라우드컴퓨팅 서비스 내에서 제공되는 관련 상용 서비스를 활용하여 개발할 수 있으며, AI 모델 학습에 이용되는 기관 데이터 및 사용자 질의문 등 개발 · 운용 시 발생하는 모든 데이터는 기관 소유로서 민감한 내부 정보는 AI 모델 학습에 활용되지 않도록 해야 하며 이를 위한 데이터 관리에 주의를 기울여야 한다.

【국가정보원 보안성 검토 절차 준수】

• 「국가정보보안기본지침」 제15조 제1항 제19호에 의거, 첨단 정보통신기술을 활용한 정보화사업을 추진하기 위해서는 국가정보원의 사전 보안성 검토 절차를 준수하여야 한다.

4.3. AI 모델 API 활용 및 민간 AI 모델 도입 방안

• AI 모델 API의 데이터 보안 고려사항

– AI 모델 API 활용 시 데이터 보안을 위해 (1) 입력 데이터 검증, (2) 개인정보 처리, (3) 데이터 오남용 방지, (4) API 키 관리 등의 사항을 고려해야 한다.

입력 데이터 검증	» API 호출 시 입력되는 데이터에 대한 유효성 검사를 수행하여 SQL 주입(SQL Injection), XSS(Cross-site Scripting) 공격 등 위해 행위 방지 » API 호출 시 입력되는 데이터에 대한 기관 내부 정보 혹은 개인정보 포함 여부를 확인
개인정보 처리	» 「개인정보 보호법」 등 개인정보의 처리에 관한 규정을 준수하여 데이터 처리 » 민감한 정보를 처리하기 전에 사용자의 동의 수집 » 데이터 가명화 및 익명화를 통해 데이터 누출 시 개인을 직접적으로 식별 또는 연관 불가토록 조치
데이터 오남용 방지	» API에 전달하는 사용자 입력 데이터 최소화 » API를 통해 반환되는 결과를 항상 검토하고, 안전성 및 정확성을 확인 » 불필요한 데이터 저장 및 공유 금지 » 필요한 최소한의 기간만 사용자 입력과 API 응답 결과를 저장하고, 보관 기간 이후에는 기관 및 API 양측 서버간 데이터 완전 삭제
API 키 관리	» API 키를 암호화된 파일에 보관하여 보안 강화 » API 키를 사용하는 프로젝트, 애플리케이션, 사용자에 대한 접근 권한 최소화 및 정기 점검 » 정기 및 수시 API 키 갱신으로 보안성 유지 » API 키 사용 내용을 지속 모니터링하여 허가되지 않은 사용이나 악의적 활동을 신속 발견 · 대응

• AI 모델 기술 API의 서버 보안 고려사항

– AI 모델 API의 서버 보안 고려사항으로는 (1) 보안 요구사항 준수, (2) 접근 제한 및 인증, (3) 예외 및 오류 처리, (4) 사용량 관리, (5) 기타 고려사항이 있다.

보안 요구사항 준수	» 「개인정보보호법」, 「전자정부법」, 「사이버안보 업무규정」 등 국내 보안 관련 법률과 규정을 정확히 이해하고 이에 따라 시스템을 설계하고 운영 » API를 사용하기 전 반드시 이용약관을 확인하고 이를 준수
접근 제한 및 인증	» 사용자, 역할, 권한에 따라 시스템의 특정 부분에 대한 접근 제어 » 기본 인증, 이중 인증(2FA), 생체 인증 등 다양한 방법을 통해 사용자의 신원 보장 및 무단 접근 방지
예외 및 오류 처리	» 올바른 예외 처리를 통해 시스템이 예외 상황에서도 안정적으로 작동하도록 시스템 신뢰성 확보 » 시스템의 안정성을 유지하고, 예상치 못한 문제로부터 시스템을 보호하기 위해 발생한 오류를 적절하게 처리
사용량 관리	» API 사용량을 추적하여 필요에 따라 사용량 제한 또는 사용 패턴 최적화 » 사용량을 제한하여 특정 사용자가 시스템을 과도하게 사용하여 다른 사용자의 서비스 이용을 방해하는 행위 방지 » 사용 패턴 최적화를 통해 서비스 비용 효율 최대화

기타 고려사항	» HTTPS, SSL/TLS 등의 보안 프로토콜을 사용하여 데이터를 안전하게 전송 » API 키 관리, 토큰 기반 인증, OAuth 등을 통해 API를 보호 » 서버 자체의 보안을 위해 OS, 네트워크, 데이터베이스 등의 시스템 보안 유지 » 시스템의 이상 행동을 감지하고, 보안 문제를 조기에 발견 · 대응하기 위해 로그 수집, 분석, 및 모니터링 » 코드 주입, XSS, CSRF 등의 공격을 방지하기 위해 안전한 코딩 기법 사용 » 최소한의 권한 원칙 적용 및 권한 변경 철저 관리 » 직원의 보안 교육을 지속적으로 실시하여 조직 내부에서 발생할 수 있는 보안 위협 대비 » 시스템 장애나 보안 사고의 신속하고 효과적인 대응을 위해 사고 대응 프로세스 정의 및 데이터 백업과 복구 전략 수립

• AI 모델 API 및 상용 AI 모델 도입 전략

[도입을 위한 기술 검토]

- AI 모델 API 및 상용 AI 모델 도입 이전에 해당 모델이 조직의 요구사항과 일치하는지 확인해야 한다. 이를 위해 (1) 기술적 특성, (2) 성능, (3) 보안, (4) 비용 등 다양한 측면에서 모델을 검토해야 한다.

기술적 특성	» AI 모델의 출력 품질과 유용성 향상을 위해 훈련 데이터 관리 » AI 모델의 출력 형태를 확인하고, 조직의 요구사항과 알맞은 출력 형태의 모델 선택 » API, 미세 조정 또는 사전 학습 등 AI 모델의 사용 형태를 사전 검토하여 알맞은 방법 도입 » 조직의 요구사항에 알맞은 미세 조정 가능 여부 확인 » AI 모델 사용 형태 및 크기 등 모델 구동에 필요한 리소스 사전 파악
성능	» AI 모델 출력결과 정확성 확인 » AI 모델 결과 응답 속도 및 지연 시간 측정 » AI 모델 출력결과 일관성 확인
보안	» 상용 AI 모델 제공자의 개인정보처리방침 및 사용자 약관을 참고하여 데이터의 수집 및 처리 방법 확인 » AI 모델 제공업체에 대한 국내 IT 보안에 대한 법적 요구사항 준수 확인 » AI 모델 제공업체의 보안 인프라 및 긴급 위기 대비태세 등 관련 정책 여부 확인
비용	» API 사용 및 민간 AI 모델 도입에 합리적 비용 고려

[테스트 전략]

- AI 모델 도입 전에는 반드시 해당 모델 결과에 대한 정확성, 적절성 등에 관한 테스트 거쳐야 한다. 이를 위해 모델의 성능을 측정하는 (1) 테스트 케이스를 설계하고, (2) 모델의 출력을 사람이 검토할 수 있는 방안을 마련해야 한다.

테스트 케이스 설계	» AI 모델이 기능적으로 필요한 모든 요구사항을 충족하는지 확인 » AI 모델이 극단적이거나 예상치 못한 입력에 대해 적절하게 반응하는지 확인 » AI 모델이 부적절하거나 위험한 요청을 적절하게 거부하거나 처리하는지 확인
AI 모델 출력 검토	» AI 모델이 정확한 · 일관된 · 적절한 정보를 제공하는지 확인 » AI 모델에 대한 미세 조정이 가능한 경우, 사용자의 피드백을 통해 AI 모델의 성능을 평가하고 개선시킬 수 있는지 확인

[시스템 통합 전략]

- AI 모델을 기존 시스템과 연계하기 위한 전략이 필요하며, 이를 위해 시스템의 아키텍처를 이해하고, 모델을 통합 · 연계하기 위한 기술적 요구사항을 파악해야 한다.

시스템 아키텍처 이해	» 시스템에 대한 문서화된 정보 검토 및 시스템 관계자 인터뷰 수행 » 데이터 흐름 및 프로세스 시각화를 통해 시스템 아키텍처 이해
기술적 요구사항 파악	» 기술적 요구사항을 정의 및 시스템과 모델 간 인터페이스 설계 » 통합 테스트 계획 수립으로 기술적 요구사항 파악

4.4. 자체 데이터 세트 및 AI 모델 구축 단계별 보안 위협 대응 방안

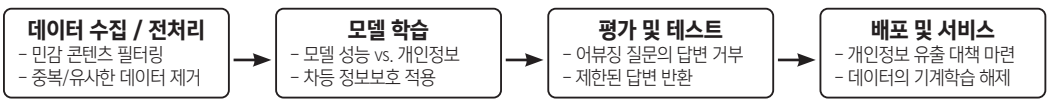

그림 14. AI 모델의 구축 단계에 따른 보안 위협의 대응 방안

• 데이터 수집/전처리 단계

[자동화된 민감 데이터 전처리]

- 개발자는 자동화된 콘텐츠 필터링 도구를 적용하여 원시 데이터에 각급기관의 기밀정보 및 개인정보 등 민감 데이터가 포함되지 않도록 완전히 제거(sanitize)해야 한다. 이때 근래의 대규모 언어모델을 학습하기 위해 사용되는 데이터 세트는 그 크기가 방대할뿐더러 상당한 비정형 데이터가 포함되어 있으며, 수동으로 필터링하는 것은 상당한 시간이 소요될 수 있어 자동화된 데이터 전처리 도구를 마련 · 활용한다.

[중복/유사 학습 데이터 제거]

- 개발자는 학습 데이터를 구축할 때 중복된 텍스트를 제거하는 것이 필요하다. 이는 대규모 언어모델이 훈련 데이터에 단어, 문자 또는 문서 수준에서 중복된 내용이 있을 경우, 모델이 이를 암기하고 예측에 활용하려는 경향이 있기 때문이다. 중복 데이터를 사용하게 되면, 정보 노출, 데이터 추론 등을 통해 훈련 데이터의 기밀성이 침해될 뿐 아니라, 민감한 개인정보가 노출될 수 있다.

• 모델 학습 단계

[AI 모델 성능 고려한 기밀 정보 유출 차단 전략 수립]

- 개발자는 AI 모델의 성능과 기밀 정보의 유출 가능성에 대한 상충 관계를 적절히 조절할 수 있는 학습 전략을 수립해야 한다. 공격자는 AI 모델이 사용한 학습 데이터와 평가 데이터 간의 손실 차이를 분석하여 특정 데이터가 훈련에 사용되었는지를 추론할 수 있다. 따라서, 개발자는 이러한 구별 가능성(distinguishable)을 최소화하여 기관 내 주요 정보 혹은 개인정보 유출 가능성을 줄이는 한편, AI 모델의 성능이 지나치게 저하되지 않게 주의해야 한다.

[정보 유출 방지를 위한 방안 마련]

- 개발자는 AI 모델의 학습 데이터에 의도하지 않게 포함될 수 있는 기관 내 주요 정보 또는 개인정보를 보호하기 위해, 학습 과정에서 임의의 노이즈(noise)를 삽입하여 정보 노출의 위험도를 낮춰주기 위한 정보보호 알고리즘(differential privacy)을 적용할 수 있다.

• 평가 및 테스트 단계

[어뷰징 질문에 대한 답변 거부]

개발자는 AI 모델이 의도하지 않은 질문에 대한 답변을 거부할 수 있도록 신중한 미세 조정 방식을 고려해야 한다. 예를 들어, 내부 정보의 유출 유도 또는 혐오 발언, 선정성 등과 같이 사회적으로 논란이나 거부감을 일으킬 수 있는 답변은 거부하거나, 윤리적인 규범에 따라 조정되어야 한다. 또한 빈 프롬프트(입력이 공백 문자열로만 이루어진 경우)를 받았을 때는 메시지가 끊긴 것으로 간주하고 답변을 다시 확인해달라는 응답을 반환하여 훈련 데이터 추출 공격을 예방할 수 있다.

[AI 모델의 출력 결과 제한]

- 개발자는 입력에 대한 AI 모델의 모든 출력을 사용자에게 그대로 전달하는 것을 제한해야 한다. 만약 입력에 대한 AI 모델의 모든 출력 결과를 사용자에게 전달한다면, 공격자는 이러한 쿼리를 반복하여 수행함으로써 비공개 모델의 복제본을 생성할 수 있다. 이를 방지하기 위해 사용자에게 반환하는 각 단어의 신뢰도에 노이즈를 추가하여 교란하거나, 사용자가 가장 선호할 것으로 예상되는 단일 또는 일정 개수의 답안을 반환하는 형태로 응답을 제한해야 한다.

• 배포 및 서비스 단계

[정보 유출 대책 마련]

- 개발자는 AI 모델이 생성한 데이터에 개인정보 또는 비공개 업무자료 등 민감한 정보로 의심되는 표현이 등장할 경우를 대비하여 적절한 대응체계를 사전에 준비해야 한다. 각급기관은 국가정보원의 「국가 정보보안 기본지침」 및 개인정보보호위원회의 「개인정보 유출 대응 매뉴얼」 등을 참고하여 자료유출 등 사고 발생 대비 대책을 마련

하여 시행하는 것을 권고한다.

[데이터의 기계학습 해제]

- 대규모 언어모델이 민감한 데이터를 암기 및 유출하는 현상을 확인하고, 학습 데이터 세트에서도 이러한 민감 정보를 발견했을 때, 개발자는 기계학습 해제(machine unlearning) 등의 방법을 이용한 모델 재학습을 고려해야 한다.

5. 부록

5.1. 약어

2FA	Two-Factor Authentication
AI	Artificial Intelligence
API	Application Programming Interface
CSV	Comma Seperated Value
DDoS	Distributed Denial of Service
DNS	Domain Name System
EDPB	European Data Protection Board
EU	European Union
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HTTPS	Hypertext Transfer Protocol Secure
IP	Internet Protocol
IT	Information Technology
LLMs	Large Language Models
ML	Machine Learning
MS	Microsoft
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
PII	Personally Identifiable Information
RBAC	Role-Based Access Control
RLHF	Reinforcement Learning from Human Feedback
SQL	Structured Query Language
SSL	Secure Sockets Layer
URL	Uniform Resource Locator

VPN	Virtual Private Network
Wi-Fi	Wireless Fidelity
XSS	Cross Site Scripting

챗GPT 등 생성형 AI 활용

보안수칙

01 **민감한 정보**(非공개 정보, 개인정보 등) **입력 금지**

* 설정에서 「대화 이력&학습」 기능 非활성화

02 생성물에 대한 **정확성·윤리성·적합성** 등 再검증

03 가짜뉴스 유포·불법물 제작·해킹 등 **범죄에 악용 금지**

04 생성물 활용 시 지적 재산권·저작권 등 **법률 침해·위반 여부 확인**

05 악의적으로 거짓 정보를 입력·학습 유도하는 등 **非윤리적 활용 금지**

06 연계·확장프로그램 사용 시 보안 취약여부 등 **안전성 확인**

07 로그인 계정에 대한 **보안설정 강화 및 보안관리 철저**

* '다중 인증(Multi-Factor Authentication)' 설정 등

국가정보원

챗GPT 등
생성형 AI 활용
보안 가이드라인