AI 보안 위협 대응 매뉴얼

AI Security Threat Mitigation Manual

TrustworthyAI Lab

https://trustworthyai.co.kr/

김재성, 송예훈, 윤채원 석사과정 (지도교수: 김호기)

중앙대학교 산업보안학과

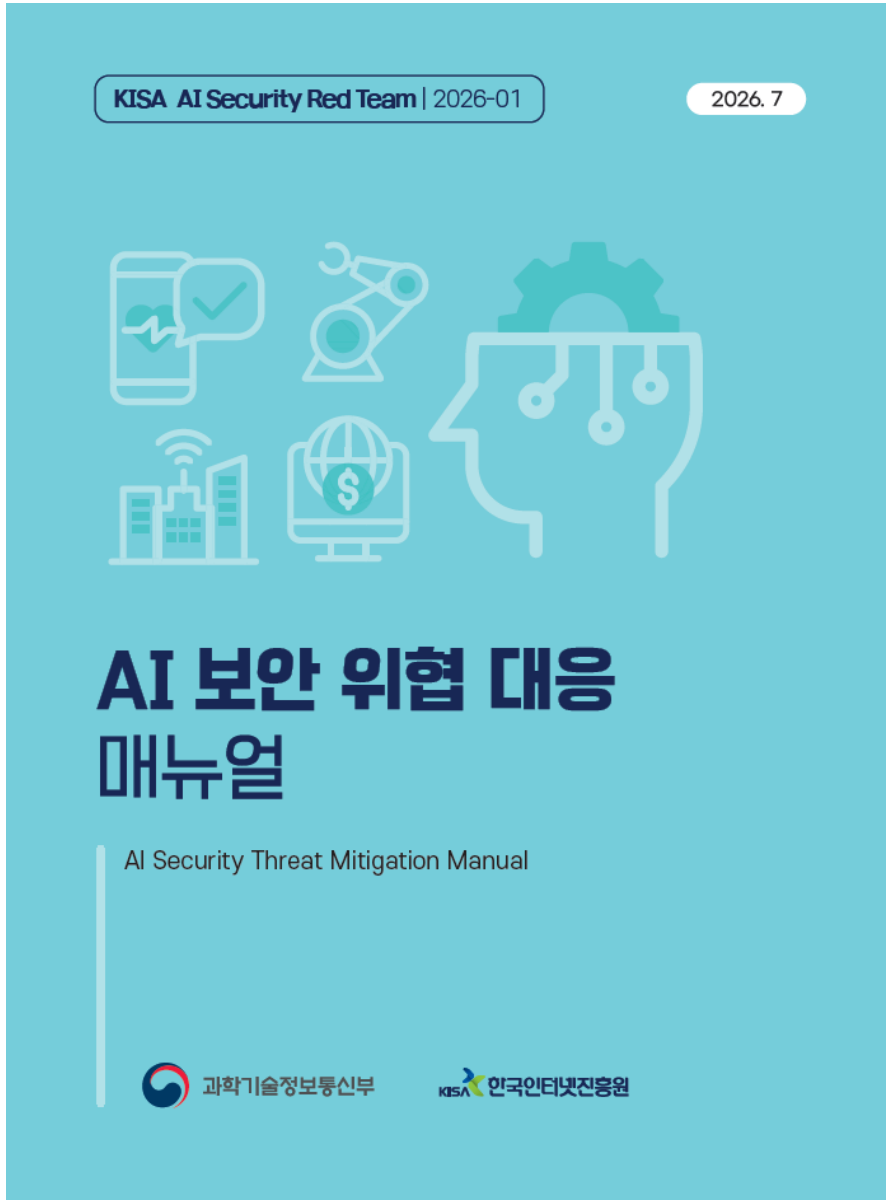
Content

1 개요

개요

발간 목적 및 대상

- **AI 보안 위협 대응 매뉴얼**
  - **발행일:** 2026년 7월
  - **발행처:** 과학기술정보통신부, 한국인터넷진흥원 AI탐지대응팀
  - **참여기관:** SK쉴더스 EQST Lab, 중앙대 TrustworthyAI Lab
  - **발간 목적**
    - 거대 언어모델(Large Language Model, 이하 LLM)의 발전과 관련 법의 제정
    - 이러한 국내외 환경 속에서 현재까지 알려진 LLM 보안 위협을 체계적으로 분류하고, 각 위협에 대한 분석 방법과 완화 방안을 제시하기 위함
  - **온라인 체크리스트**
    - https://trustworthyai.co.kr/ai-security-checklist/



개요

발간 목적 및 대상

- **적용 범위**
  - **분석 대상:** LLM과 이를 포함하는 LLM 시스템으로 한정함

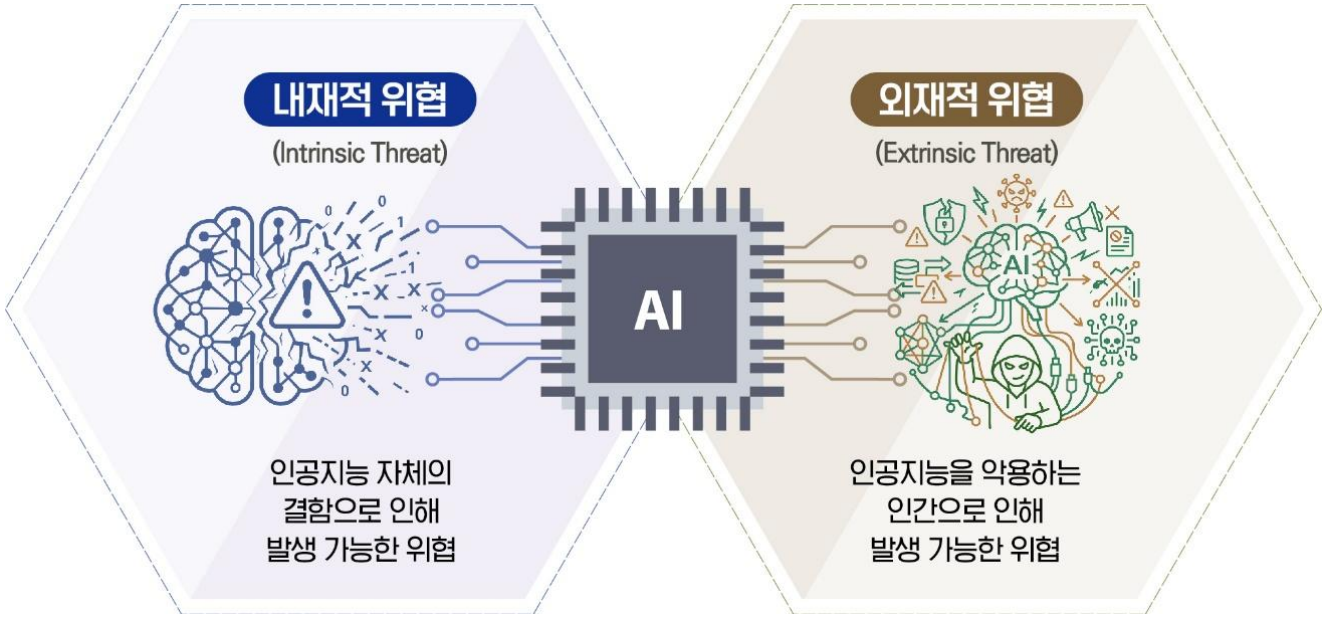
AI 세분류에 관한 용어집

용어	영문, 약어	설명
인공지능	Artificial Intelligence, AI	사람의 인지·판단·추론 과정을 컴퓨터로 구현하려는 기술 분야 전체.
머신러닝	Machine Learning	사람이 규칙을 설계하지 않고, 데이터로부터 모델이 스스로 규칙을 학습하도록 만든 기법.
딥러닝	Deep Learning	인공신경망을 활용해 데이터의 복잡한 패턴을 자동으로 학습하는 머신러닝의 한 갈래.
생성형 AI	Generative AI, GenAI	텍스트·이미지·코드·음성 등 새로운 콘텐츠를 만들어 내는 딥러닝 모델 집합.
LLM	Large Language Model	방대한 텍스트로 학습되어 다음 단어를 확률적으로 예측·생성하는 인공신경망 모델.
LLM 애플리케이션	LLM Application	LLM을 핵심 두뇌로 두고 도구·에이전트·외부 데이터와 결합해 사용자 요청을 처리하는 단일 서비스.
LLM 시스템	LLM System	LLM 애플리케이션 뿐만 아니라 모델 공급망·운영 인프라·외부 데이터·사용자·운영자 경계를 포함하는 더 넓은 운영 환경.

개요

AI 안전, 보안 그리고 위협

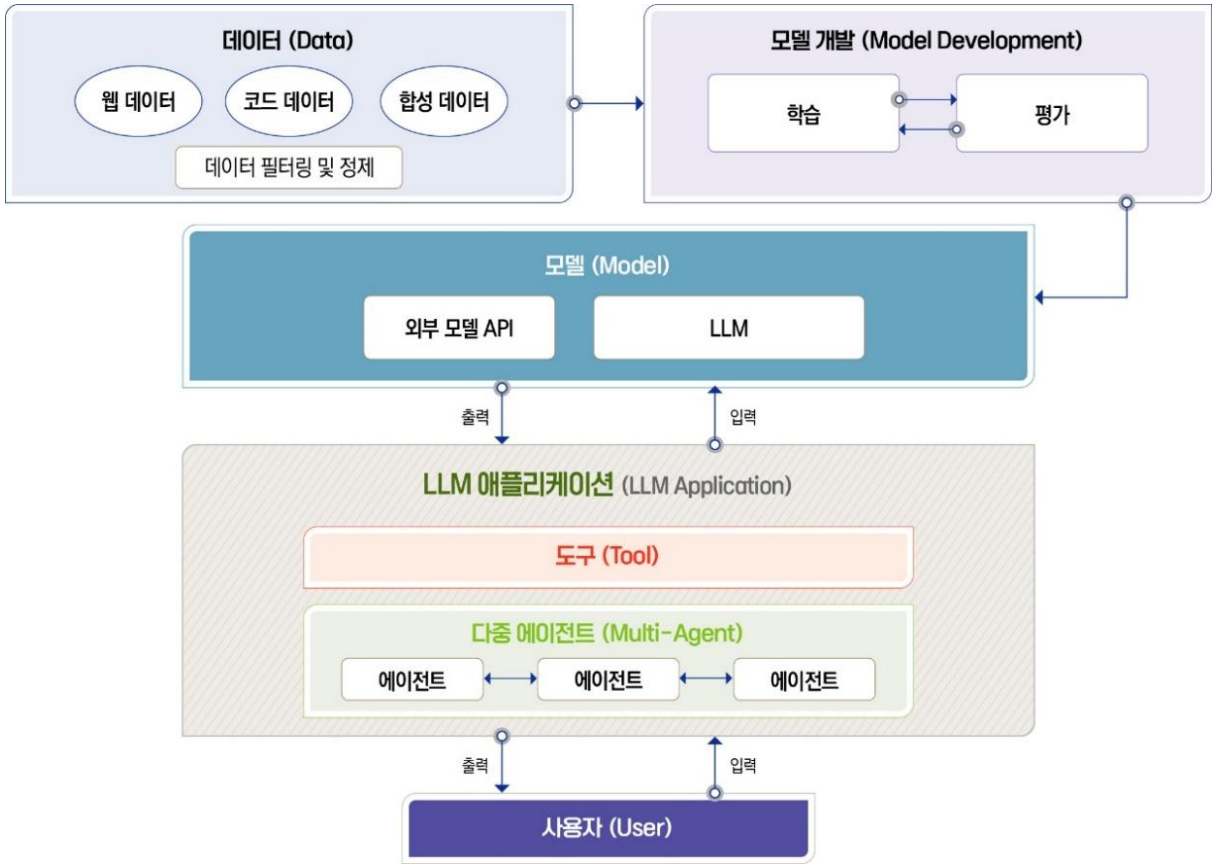
- **위협의 구분**
  - **내재적 위협:** AI 자체의 한계나 고유한 특성에서 비롯되는 위협
  - **외재적 위협:** AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협



개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - **내재적 위협**은 LLM 생명 주기 전 과정에서 발생
    - 데이터
    - 모델 개발
    - 모델
    - LLM 애플리케이션
    - 사용자



개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 내재적 위협: **AI 자체의 한계나 고유한 특성에서 비롯되는 위협**
    - 학습 데이터 편향(Training Data Bias)
    - 환각(Hallucination)
    - 정렬 실패(Alignment Failure)
    - 확률적 비결정성(Non-Determinism)

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 내재적 위협: AI 자체의 한계나 고유한 특성에서 비롯되는 위협
    - **학습 데이터 편향(Training Data Bias)**
      - 학습 데이터의 분포 편향이 모델 응답에 그대로 반영되거나, 학습 과정에서 특정 문자열이 모델에 저장되어 추론 시점에 재현되는(기억화) 위협
    - 환각(Hallucination)
    - 정렬 실패(Alignment Failure)
    - 확률적 비결정성(Non-Determinism)

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 내재적 위협: AI 자체의 한계나 고유한 특성에서 비롯되는 위협
    - 학습 데이터 편향(Training Data Bias)
    - **환각(Hallucination)**
      - 모델이 사실과 무관한 내용을 사실인 것처럼 그럴듯하게 생성하는 위협으로, LLM 모델에서 공통적으로 나타나는 현상
    - 정렬 실패(Alignment Failure)
    - 확률적 비결정성(Non-Determinism)

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 내재적 위협: AI 자체의 한계나 고유한 특성에서 비롯되는 위협
    - 학습 데이터 편향(Training Data Bias)
    - 환각(Hallucination)
    - **정렬 실패(Alignment Failure)**
      - 모델 출력이 운영자의 정책·시스템 프롬프트와 어긋나는 위협
      - 정렬이란 LLM의 출력이 개발자·운영자가 의도한 정책·시스템 프롬프트와 일치하도록 모델을 추가 학습 또는 조정하는 절차를 가리킴
    - 확률적 비결정성(Non-Determinism)

개요

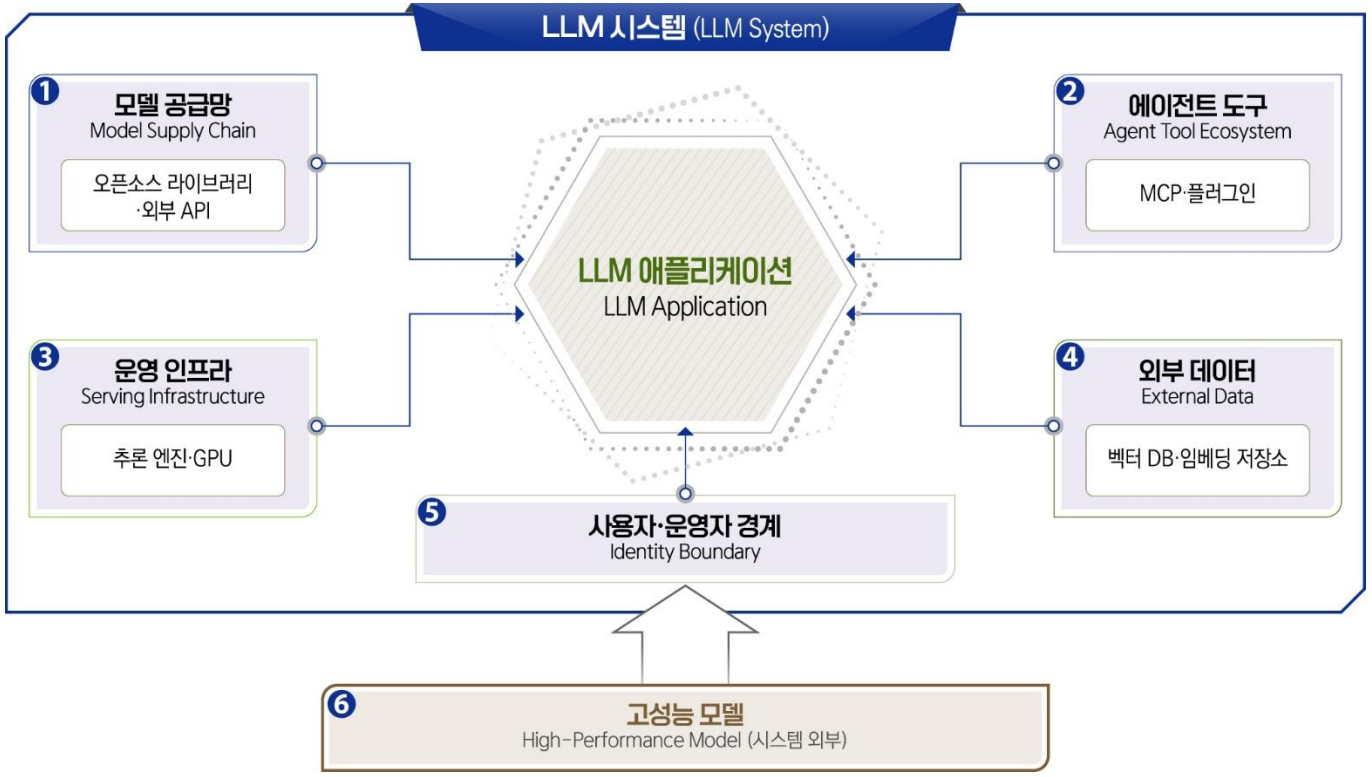
AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 내재적 위협: AI 자체의 한계나 고유한 특성에서 비롯되는 위협
    - 학습 데이터 편향(Training Data Bias)
    - 환각(Hallucination)
    - 정렬 실패(Alignment Failure)
    - **확률적 비결정성(Non-Determinism)**
      - 동일 입력에 대해 모델 출력이 매번 달라지는 현상으로, LLM의 확률적 출력 특성과 GPU 연산의 비결정적 특성 등에 기반함
      - 보안 평가의 재현이 어려워지고 다른 내재적 위협 유형과 복합적으로 발생할 수 있음

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - **외재적 위협**은 애플리케이션 내부가 아닌 LLM 시스템에서 주로 발생
    - 데이터·모델 공급망
    - 에이전트 도구
    - 운영 인프라
    - 외부 데이터
    - 사용자·운영자 경계
    - 고성능 모델



개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 외재적 위협: **AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협**
    - 탈옥(Jailbreak)
    - 모델 추출(Model Extraction)
    - 공급망·인프라 대상
    - 고성능 모델 위협

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 외재적 위협: AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협
    - **탈옥(Jailbreak)**
      - 모델이 정책상 거부해야 할 요청(해킹 툴 제작, 개인정보 추출, 혐오 표현 등)을 무력화하는 기법
    - 모델 추출(Model Extraction)
    - 공급망·인프라 대상
    - 고성능 모델 위협

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 외재적 위협: AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협
    - 탈옥(Jailbreak)
    - **모델 추출(Model Extraction)**
      - 추론 API에 대한 반복 질의로 원본 모델을 복제하여 유출하는 위협
    - 공급망·인프라 대상
    - 고성능 모델 위협

개요

AI 안전, 보안 그리고 위협

- **위협의 구분**
  - 외재적 위협: AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협
    - 탈옥(Jailbreak)
    - 모델 추출(Model Extraction)
    - **공급망·인프라 대상**
      - 모델, 라이브러리, 의존성, 학습 데이터, 벡터 DB 등 AI 시스템의 공급망과 외부 데이터 경로를 오염시키거나, MCP와 같은 외부 도구 연결을 통해 악성코드를 주입하는 위협
    - 고성능 모델 위협

개요

AI 안전, 보안 그리고 위협

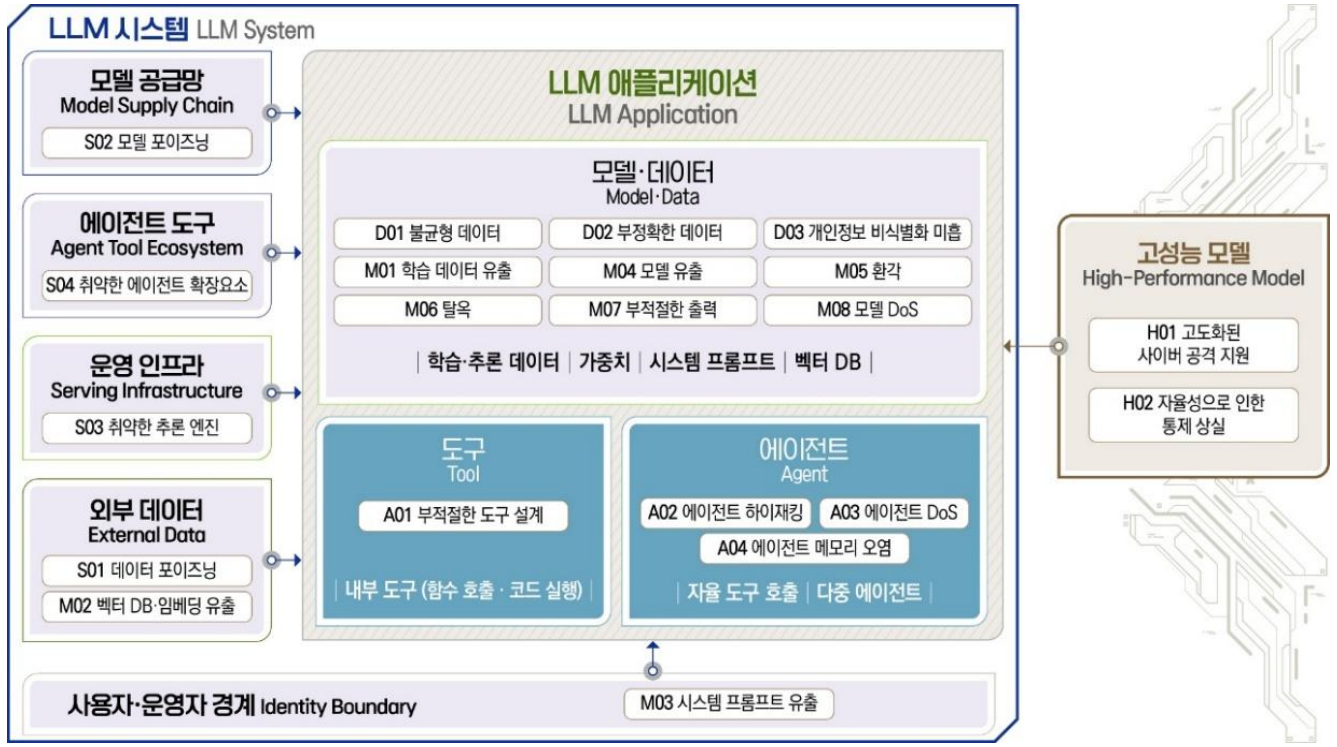
- **위협의 구분**
  - 외재적 위협: AI를 도구로 활용하는 외부 행위자에 의해 야기되는 위협
    - 탈옥(Jailbreak)
    - 모델 추출(Model Extraction)
    - 공급망·인프라 대상
    - **고성능 모델 위협**
      - 고성능 모델 자체가 외부 공격자의 도구로 전용되는 공격
      - 고성능 모델은 사이버 공격 지원과 자율적 의사결정 등에서 위협행위자의 가용 능력 상한을 끌어올릴 수 있는 첨단·대규모 모델 전반을 가리킴

2 AI 보안 위협 분류 및 진단

AI 보안 위험 분류 및 진단

AI 보안의 중요성 확대

- **공격 표면의 확장**
  - LLM 기반 서비스는 다양한 구성요소가 결합된 복합 애플리케이션 형태로 발전
  - 보안 위협 또한 전통적인 소프트웨어 취약점에 국한되지 않고 새로운 유형으로 확대
  - **명확한 대응책 제시를 위해 위협을 재분류**
    1. 데이터 및 모델 위협
    2. 에이전트 및 공급망 위협
    3. 고성능 모델 위협



AI 보안 위험 분류 및 진단

세부 위협 분류

AI 보안 위험 분류 및 진단

세부 위협 분류

AI 보안 위험 분류 및 진단

데이터 위협

- **데이터 위협**
  - LLM은 대규모 데이터를 기반으로 학습됨
  - 학습 데이터의 품질, 대표성, 개인정보 처리 수준은 서비스의 안전성에 직접적인 영향을 미침
  - 데이터 위협은 주로 **학습 및 활용 데이터의 품질 관리 미흡**에서 발생함

AI 보안 위험 분류 및 진단

데이터 위협

- **[D01] 불균형 데이터**
  - 정의
    - 특정 속성이나 클래스에 편중된 데이터로 인해 LLM이 부정확한 결과를 학습하고 출력하는 위협
  - 발생 원인
    - 학습 데이터 수집 과정에서 특정 클래스에 해당하는 데이터가 과도하게 포함되거나 일부 데이터가 충분히 확보되지 않는 경우 발생 가능
  - 영향
    - 모델이 다수 데이터에 치우친 패턴을 학습하고 소수 집단이나 희소 사례에 대해 부정확한 판단을 할 가능성이 있음

AI 보안 위험 분류 및 진단

데이터 위협

- **[D02] 부정확한 데이터**
  - 정의
    - 사실적 오류, 논리적 모순, 잘못된 레이블 등이 포함된 데이터를 사용하여 LLM이 부정확한 정보를 학습하고 출력하는 위협
  - 발생 원인
    - 오래된 문서의 미갱신, 검증되지 않은 외부 자료의 활용, 데이터 등록 및 관리 절차 미흡
  - 영향
    - AI 모델이 사실과 다른 정보나 현재 기준에 맞지 않는 응답을 생성할 수 있음

AI 보안 위험 분류 및 진단

데이터 위협

- **[D03] 개인정보 비식별화 미흡**
  - 정의
    - 데이터 내 식별 정보(이름, 주민등록번호 등)가 제대로 제거되지 않아 개인정보가 노출되는 위협
  - 발생 원인
    - 학습 단계에서 데이터 수집 및 정제 과정에서 개인정보 탐지 및 제거 절차 미흡
    - 추론 단계에서 처리 목적에 필요한 범위를 초과한 정보가 권한 검증 없이 참조될 경우
  - 영향
    - 학습 데이터에 포함된 정보를 재현하거나, 추론 과정에서 권한 없는 사용자에게 개인정보 노출될 위험

AI 보안 위험 분류 및 진단

데이터 위협

- **자가 진단 체크리스트**

항목	기준	Y	N
데이터 및 모델 위협			
D01 불균형 데이터	학습 데이터셋이 특정 출처, 플랫폼, 국가, 언어, 산업 분야, 사용자군 등에 편중이 없는지 검토하는가		
	불균형이 확인된 데이터에 대해 재수집, 제외 등 완화 조치를 수행하는가		
	불균형 데이터 식별 및 완화 조치 결과를 기록하고 관리하는 승인 절차가 있는가		
D02 부정확한 데이터	학습 데이터 내 사실적 오류, 논리적 모순, 불완전한 정보, 잘못된 레이블 등이 포함되어 있는지 검토하는가		
	학습 데이터셋의 사실성, 정확성, 일관성을 검증하기 위한 기준이 있는가		
	부정확한 데이터가 확인된 경우 수정, 제외, 재수집 등 정정 조치를 수행하는가		
	데이터 정정 이력, 검수 결과, 승인 내역 등을 체계적으로 기록하고 관리하는가		
D03 개인정보 비식별화 미흡	학습 데이터에 개인정보 비식별화(가명·익명 처리 등) 조치를 적용하는가		
	원본 데이터와 비식별 데이터를 분리하고, 접근 권한을 철저히 구분하여 통제하는가		
	모델 및 AI 시스템을 대상으로 개인정보 영향평가를 수행하는가		

AI 보안 위험 분류 및 진단

세부 위협 분류

AI 보안 위험 분류 및 진단

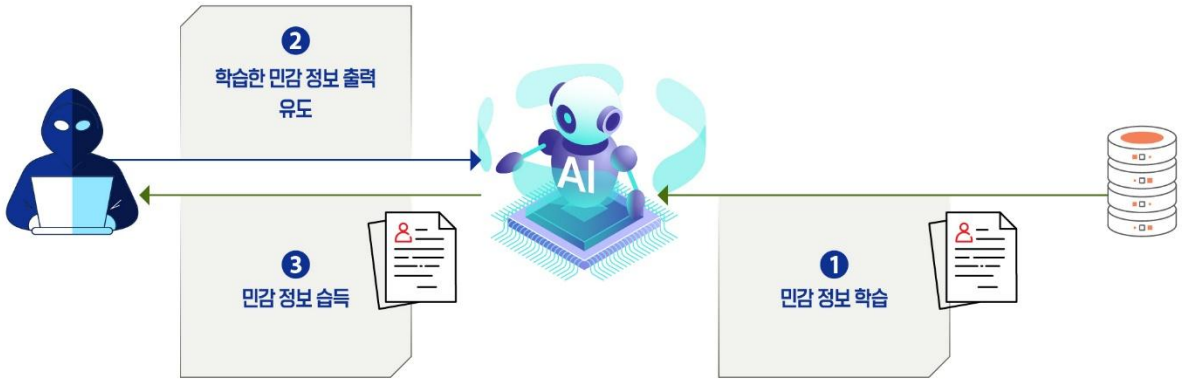
모델 위협

- **모델 위협**
  - 대규모 데이터를 기반으로 학습된 모델은 사용자 입력에 따라 응답을 생성하고 업무 자동화에 활용됨
  - 모델의 응답 특성, 한계, 민감정보 노출 가능성은 서비스의 안전성, 신뢰성, 프라이버시에 큰 영향을 미침
  - 모델 위협은 **AI 모델이 가진 구조적 한계와 생성 특성**에서 발생함

AI 보안 위험 분류 및 진단

모델 위협

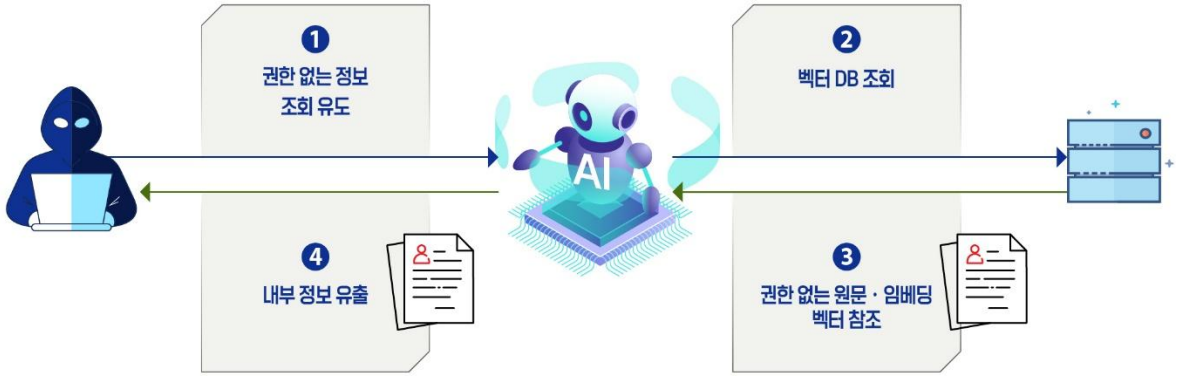
- **[M01] 학습 데이터 유출**
  - 정의
    - 악의적인 질의를 통해 LLM이 학습했던 원본 데이터가 복원되거나 유출되는 위협
  - 발생 원인
    - 모델이 학습 데이터의 일부를 과도하게 암기한 경우 발생 가능
  - 영향
    - 외부에 학습 데이터에 포함된 개인정보, 기밀정보, 내부 문서 등이 유출될 수 있음



AI 보안 위험 분류 및 진단

모델 위협

- **[M02] 벡터 DB·임베딩 유출**
  - 정의
    - RAG 등 검색에 사용되는 벡터 DB에서 임베딩 벡터나 원문 데이터가 외부로 유출되는 위협
  - 발생 원인
    - 벡터 DB 내 임베딩·문서 청크·메타데이터에 대한 접근 권한 통제 미흡
  - 영향
    - 내부 문서, 업무 자료, 개인정보, 기밀 메타데이터 등이 외부에 노출될 수 있으며, 임베딩의 경우 유사도 분석, 반복 질의, 역추론 기법 등에 사용될 수 있음



AI 보안 위험 분류 및 진단

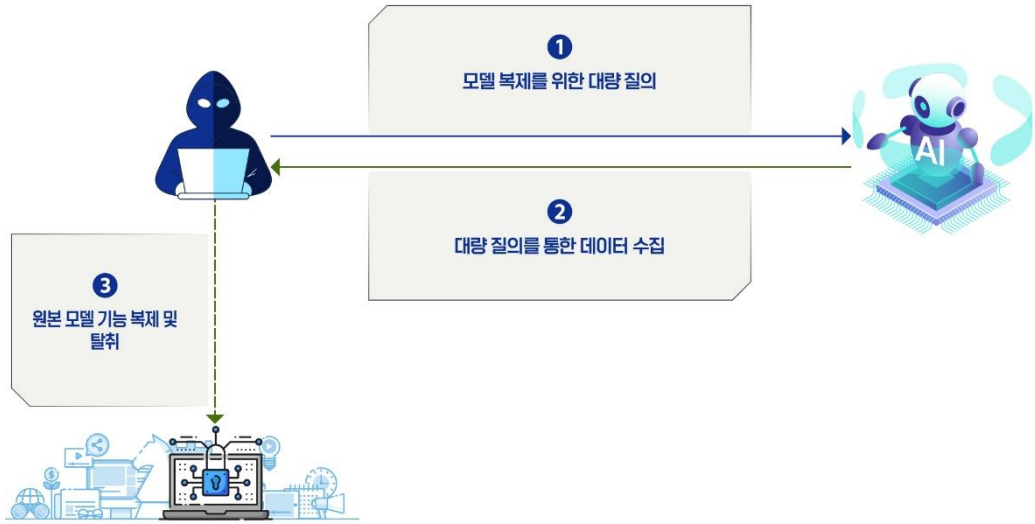
모델 위협

- **[M03] 시스템 프롬프트 유출**
  - 정의
    - LLM이 동작하는데 필요한 시스템 프롬프트가 사용자에게 유출되는 위협
  - 발생 원인
    - 시스템 프롬프트를 보호를 위한 지침이 명확하지 않거나, 모델이 시스템 프롬프트 공개를 요구하는 사용자 요청을 적절히 거부하지 못하는 경우 발생 가능
  - 영향
    - 공격자가 모델의 제한 조건과 동작 방식을 파악하여 탈옥, 정책 우회 등 후속 공격에 활용 가능

AI 보안 위험 분류 및 진단

모델 위협

- **[M04] 모델 유출**
  - 정의
    - 모델 파일, 가중치, 설정 정보 등이 외부로 유출 및 복제되는 위협
  - 발생 원인
    - 모델 저장소에 대한 접근 통제 미흡, 또는 추론 API가 확률값이나 로짓 등 과도한 정보를 제공하거나 반복 질의에 대한 제한이 부족한 경우 발생 가능
  - 영향
    - 모델 가중치, 구조, 응답 패턴 등 핵심 자산이 노출되고, 이로 인해 서비스 이용 제한 우회나 후속 공격에 활용 가능



AI 보안 위험 분류 및 진단

모델 위협

- **[M05] 환각**
  - 정의
    - LLM이 실제 사실이나 주어진 맥락에 일치하지 않는 정보를 그럴듯하게 생성하는 위협
  - 발생 원인
    - LLM은 실제 사실 여부를 검증하기보다 학습 데이터의 패턴과 입력 문맥에 기반하기 때문에 발생 가능
  - 영향
    - 모델이 사슬과 다른 정보, 존재하지 않는 출처, 잘못된 설명이나 판단 근거 등을 생성할 수 있음

AI 보안 위험 분류 및 진단

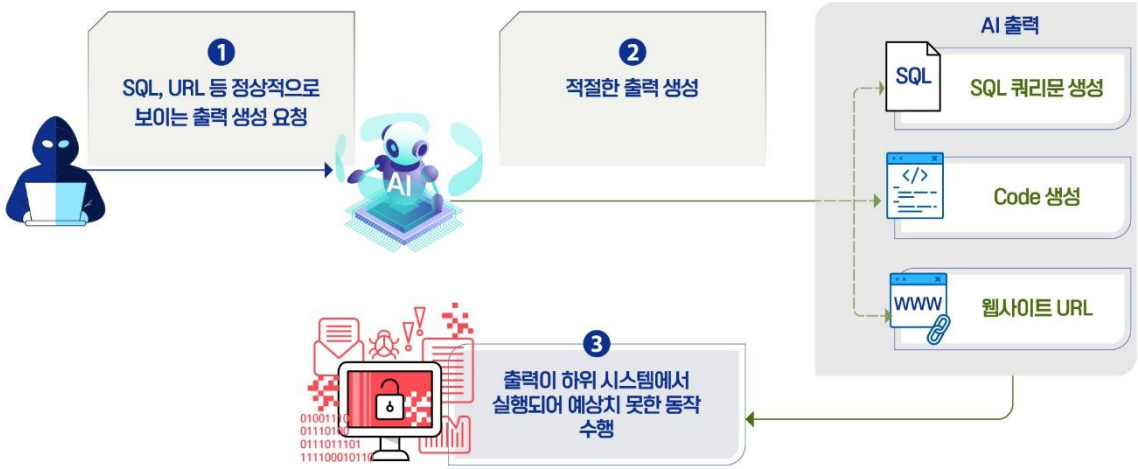
모델 위협

- **[M06] 탈옥**
  - 정의
    - LLM이 실제 사실이나 주어진 맥락에 일치하지 않는 정보를 그럴듯하게 생성하여 사용자 판단 오류나 잘못된 의사 결정을 유발할 수 있는 위협
  - 발생 원인
    - LLM은 실제 사실 여부를 검증하기보다 학습 데이터의 패턴과 입력 문맥을 기반으로 가능성이 높은 응답으로 생성하기 때문에 발생 가능
  - 영향
    - AI 모델이 사실과 다른 정보, 존재하지 않는 출처, 잘못된 설명이나 판단 근거 생성 가능

AI 보안 위험 분류 및 진단

모델 위협

- **[M07] 부적절한 출력 처리**
  - 정의
    - LLM의 출력이 다른 시스템(UI, DB 등)에 그대로 실행 또는 반영되어 시스템 오작동이나 취악점을 유발하는 위협
  - 발생 원인
    - 모델 출력이 신뢰할 수 있는 값으로 간주되거나, 검증이 충분히 수행되지 않을 때 발생 가능
  - 영향
    - 모델이 유해하거나 제한되어야 할 정보를 생성하거나 시스템 프롬프트 우회, 민감정보 유출 등 다른 위협으로 이어질 수 있음



AI 보안 위험 분류 및 진단

모델 위협

- **[M08] 모델 DoS**
  - 정의
    - 과도하거나 복잡한 입력을 주입해 시스템 자원을 고갈시키고 LLM의 서비스 지연 및 중단을 유발하는 위협
  - 발생 원인
    - LLM은 입출력 토큰 수, 추론 복잡도에 따라 연산 자원과 처리 시간이 크게 증가할 수 있음
  - 영향
    - 정상 사용자의 요청 처리가 지연되거나 실패하는 등 가용성이 저하될 수 있음

AI 보안 위험 분류 및 진단

모델 위협

- **자가 진단 체크리스트**

항목	기준	Y	N
데이터 및 모델 위협			
M01 학습 데이터 유출	모델이 고의적, 반복적 질의에도 학습 데이터 복원(에 가까운) 응답을 하지 않는지 확인하는가		
	데이터 추출을 노린 반복적이거나 유사한 악성 입력을 차단하는가		
M02 벡터 DB·임베딩 유출	벡터 DB 및 임베딩 저장소에 대한 접근 권한을 통제하는가		
	벡터 DB 또는 RAG 검색 결과에 포함된 민감정보가 모델 출력에 노출되지 않도록 관리하는가		
M03 시스템 프롬프트 유출	시스템 프롬프트 공개 요청에 대해 응답을 제한하는가		
	모델 출력에 시스템 프롬프트가 노출되지 않도록 통제하는가		
M04 모델 유출	모델 가중치 및 파일을 내부 네트워크에서만 접근 가능하도록 보호하는가		
	추론 API에 대한 비정상적인 호출 및 빈도를 제한하는 설정을 관리하는가		
	필요시 전체 확률값(logits·logprobs) 대신 상위 k개의 토큰만 제한적으로 제공하는가		
	확률값 제공 시 정밀도를 낮추거나 노이즈를 추가하여 원본 모델 추정을 방지하는가		

항목	기준	Y	N
데이터 및 모델 위협			
M05 환각	모델이 거짓 정보(환각)나 허위 근거를 생성할 가능성을 정기적으로 평가하는가		
	중요한 답변에 대해 출처 확인 및 근거를 제시하도록 검증 절차를 수행하는가		
	사용자에게 AI의 답변에 오류나 환각이 포함될 수 있음을 명확히 고지하는가		
M06 탈옥	보안 정책 우회나 역할 변경을 유도하는 악의적 요청을 탐지·차단하는가		
	생성된 응답에 정책 위반 또는 유해 콘텐츠가 포함될 경우 사용자에게 전달되기 전에 차단하는가		
M07 부적절한 출력 처리	모델 출력을 시스템에 반영하기 전 안전 검증(필터링, 이스케이핑 등)을 수행하는가		
	모델 출력이 시스템 명령, 설정값, 코드, 쿼리, 클라이언트 화면 등에 검증 없이 직접 삽입되지 않도록 통제하는가		
M08 모델 DoS	모델 서비스 가용성을 저하시킬 수 있는 과도한 요청 유형을 식별하고 평가하는가		
	입력 검증, 요청 빈도 제한, 사용량 제한 등 과부하 방지 메커니즘을 적용하는가		
	비정상 트래픽 및 대량 요청에 대한 탐지·차단·제한 절차가 있는가		

AI 보안 위험 분류 및 진단

세부 위협 분류

AI 보안 위험 분류 및 진단

에이전트 위협

- **에이전트 위협**
  - LLM 에이전트란?
    - 흔히 사용하는 대화형 LLM과 달리 이메일, 데이터베이스, 외부 API 등 다양한 외부 도구에 직접 접근하여 실제 작업을 수행할 수 있는 자율적 시스템
    - **부적절한 도구 설계, 악의적인 조작을 통한 에이전트 하이재킹** 등 다양한 보안 위협에 노출될 수 있음

AI 보안 위험 분류 및 진단

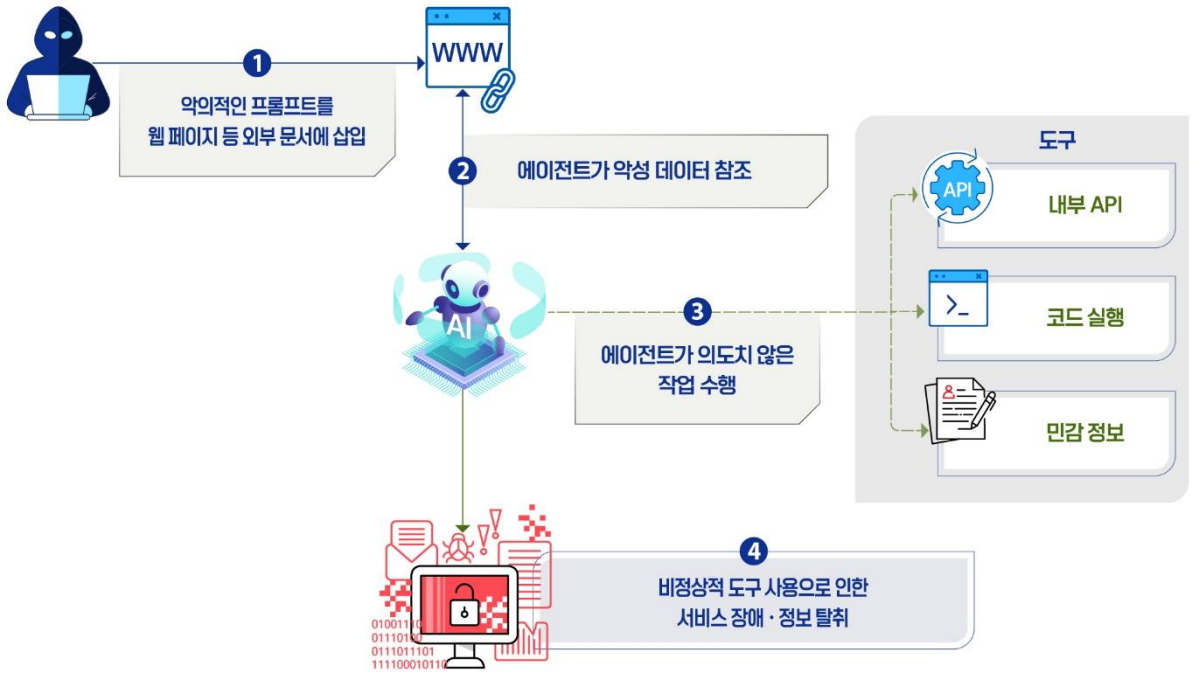
에이전트 위협

- **[A01] 부적절한 도구 설계**
  - 정의
    - LLM 에이전트 도구의 권한 제어 및 검증 미흡으로 악성 입력이 시스템 오동작이나 정보 유출을 일으키는 위협
  - 발생 원인
    - 도구에 필요한 최소 권한보다 과도한 실행 권한이 부여되거나, 수행 가능한 작업 범위가 명확히 제한되지 않은 경우 발생 가능
  - 영향
    - 임의 명령 실행, 악성코드 실행 등을 통해 시스템이 비정상적으로 동작하거나 장악될 수 있음

AI 보안 위험 분류 및 진단

에이전트 위협

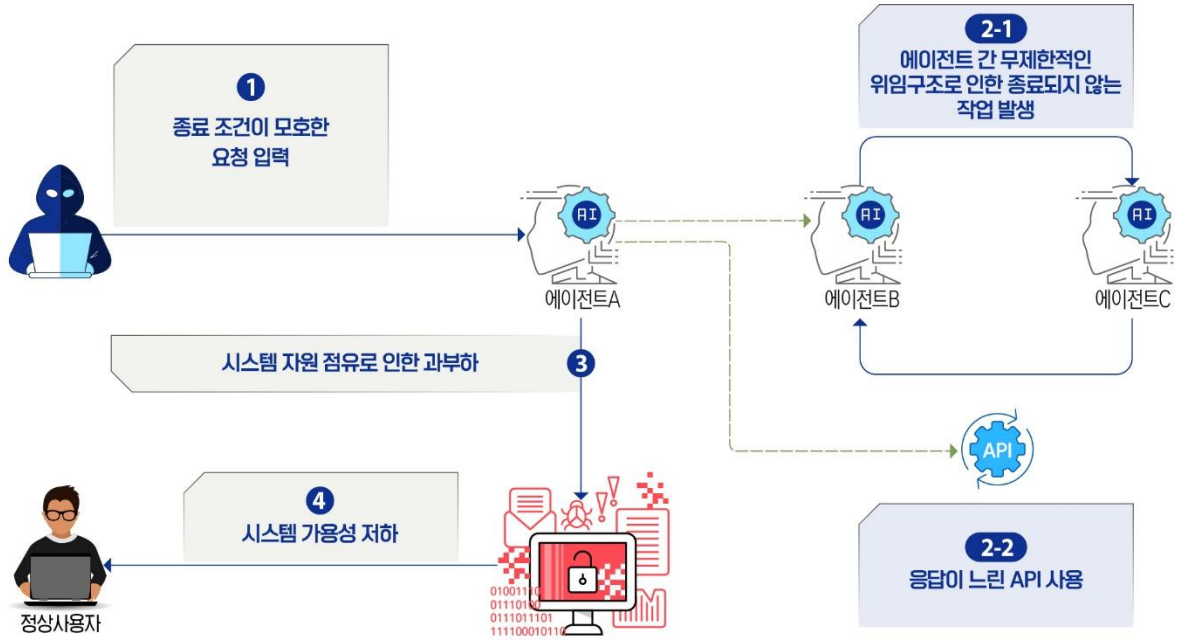
- **[A02] 에이전트 하이재킹**
  - 정의
    - 웹 페이지, 문서 등 외부데이터에 숨겨진 악성 프롬프트를 에이전트가 정상 지시로 착각해 의도치 않은 작업을 수행하는 위협
  - 발생 원인
    - 신뢰할 수 없는 외부 콘텐츠를 조회하는 권한과 상태를 변경할 수 있는 실행 권한이 동일 에이전트에 함께 부여된 경우 발생 가능
  - 영향
    - 에이전트가 공격자의 지시를 따라 민감정보인 사용자의 대화 내역, 개인정보 등을 유출할 수 있음



AI 보안 위험 분류 및 진단

에이전트 위협

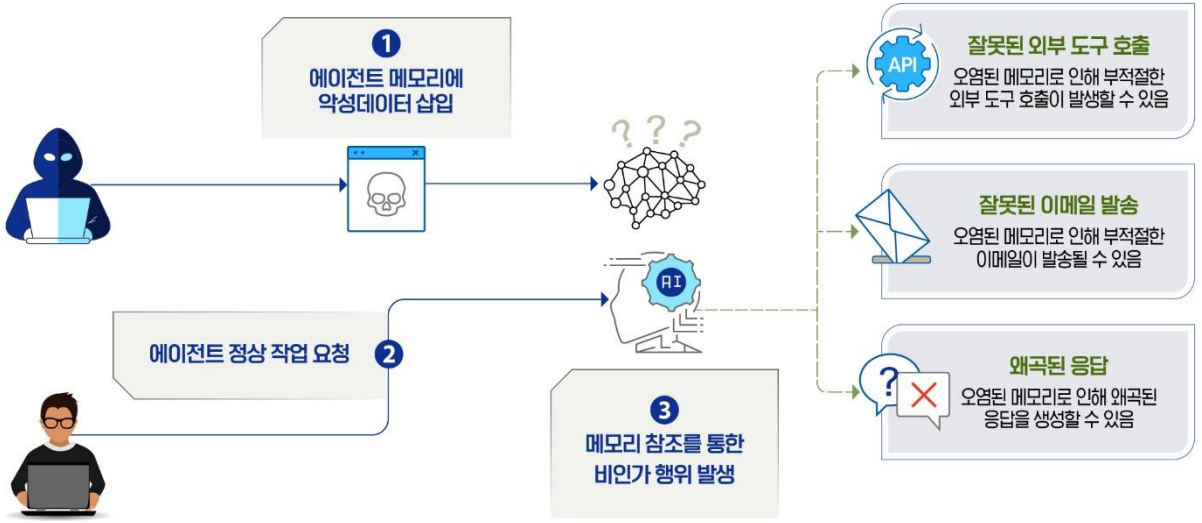
- **[A03] 에이전트 DoS**
  - 정의
    - 에이전트가 무한 루프나 과도한 API 호출을 발생시켜 시스템 자원과 비용을 고갈시키고 서비스를 마비시키는 위협
  - 발생 원인
    - 에이전트 작업 시간, 반복 횟수, API 호출 빈도, 동시 실행 수에 대한 제한이 없거나 부적절한 경우 발생 가능
  - 영향
    - 시스템 자원이 과도하게 소모되어 정상적인 사용자 요청을 처리하지 못하거나 서비스 응답 속도 저하 위험



AI 보안 위험 분류 및 진단

에이전트 위협

- **[A04] 에이전트 메모리 오염**
  - 정의
    - 에이전트의 메모리에 악성 데이터가 삽입되어 이후의 추론 및 판단 과정에 지속적으로 악영향을 미치는 위협
  - 발생 원인
    - 신뢰할 수 없는 입력이나 일시적인 작업 지시가 검증 없이 메모리에 반영되거나, 메모리 읽기 및 쓰기 권한을 과도하게 부여한 경우 발생 가능
  - 영향
    - 에이전트가 잘못된 정보를 사실로 인식하거나 공격자가 의도한 방향으로 왜곡된 응답을 생성할 수 있음



AI 보안 위험 분류 및 진단

에이전트 위협

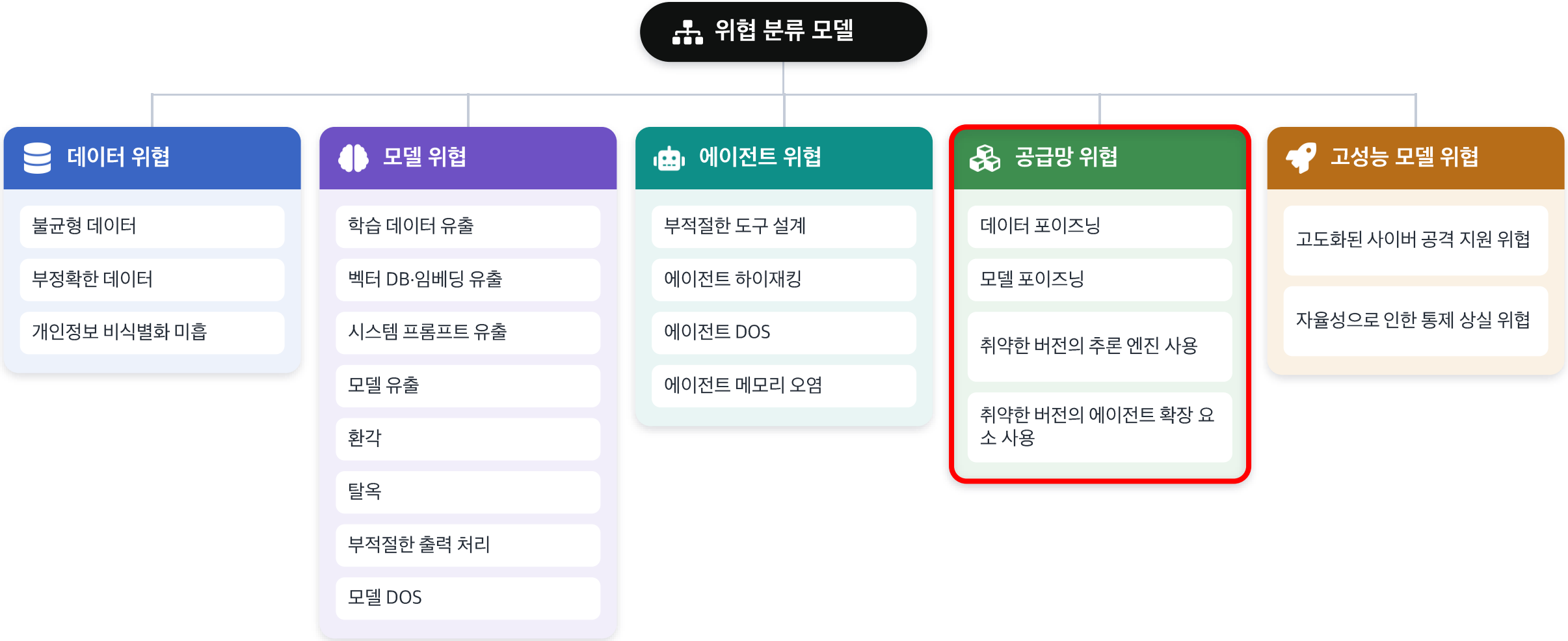
- **자가 진단 체크리스트**

항목	기준	Y	N
에이전트 및 공급망 위협			
A01 부적절한 도구 설계	사용자, 역할, 권한 수준에 따라 에이전트가 사용할 수 있는 도구·기능·실행 권한을 최소한으로 제한하는가		
	도구 호출에 사용되는 입력값에 대해 형식, 범위, 권한 검증을 수행하는가		
	도구 실행 결과에 민감정보, 오류 정보, 비인가 데이터 등이 포함되지 않도록 검증·필터링하는가		
	도구 호출 이력과 실행 결과를 기록하고, 비정상적인 도구 호출을 탐지·차단하는가		
A02 에이전트 하이재킹	수집된 데이터에서 악성 지침이나 숨겨진 명령어를 탐지하고 제거하는가		
	에이전트가 접근하는 외부 데이터에 대한 신뢰성 검증 체계가 있는가		
	사전에 승인되고 신뢰할 수 있는 데이터 소스에만 에이전트가 접근하도록 제한하는가		

항목	기준	Y	N
에이전트 및 공급망 위협			
A03 에이전트 DoS	에이전트 작업 실행 시간에 대한 적절한 제한을 설정하는가		
	사용자별 또는 세션별 요청 빈도 제한을 적용하는가		
	비정상적인 작업 패턴 탐지 및 자동 중단 기능이 있는가		
	동시 실행 가능한 작업 수에 대한 제한이 있는가		
A04 에이전트 메모리 오염	에이전트 장기 메모리(RAG DB, 파일 등)에 인증 및 권한 제어를 적용하는가		
	장기 메모리에 저장되는 항목에 대해서 금칙어 필터링, 명령 패턴 차단, 정책 기반 검증 절차가 있는가		
	장기 메모리 데이터에 대해 주기적으로 무결성을 검토하고 이상 징후를 탐지하는가		

AI 보안 위험 분류 및 진단

세부 위협 분류



AI 보안 위험 분류 및 진단

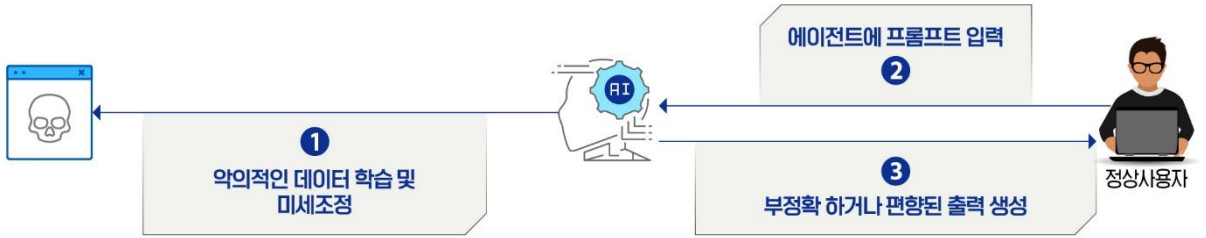
공급망 위협

- **공급망 위협**
  - AI 시스템의 **학습 데이터, 모델, 배포 플랫폼, 추론 엔진** 등 구성 요소에서 다양하게 발생하는 위협
  - 대표적으로 **데이터 · 모델 포이즈닝, 취약한 버전 추론 엔진 사용** 등이 있음

AI 보안 위험 분류 및 진단

공급망 위협

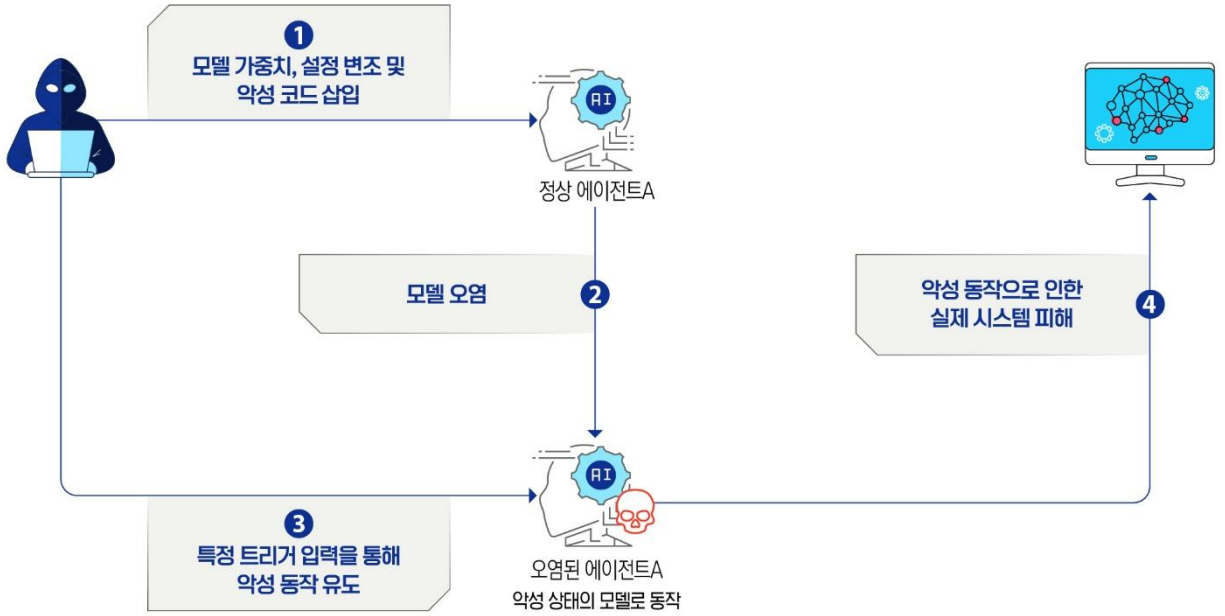
- **[S01] 데이터 포이즈닝**
  - 정의
    - 모델 학습 및 평가 데이터에 악의적인 데이터를 섞어 넣어, 모델의 동작과 결과를 의도적으로 왜곡하는 위협
  - 발생 원인
    - 검증되지 않은 데이터(공개 웹 크롤링 데이터, 외부 데이터셋 등)를 학습 또는 미세 조정에 사용하는 경우 발생 가능
  - 영향
    - 모델이 특정 입력에 대해 공격자가 의도한 응답을 생성하거나, 부정확하고 편향된 결과를 출력할 수 있음



AI 보안 위험 분류 및 진단

공급망 위협

- **[S02] 모델 포이즈닝**
  - 정의
    - 모델의 가중치, 설정 등을 변조하여 출력 결과를 조작하거나 악성코드를 삽입하는 위협
  - 발생 원인
    - 외부 모델 저장소의 파일을 무결성 검증 없이 다운로드하거나 배포하는 경우 발생 가능
  - 영향
    - 모델이 특정 트리거 입력에 대해 악성 응답을 생성하거나, 거짓 정보, 편향된 답변, 안전 정책을 우회한 출력을 생성할 수 있음



AI 보안 위험 분류 및 진단

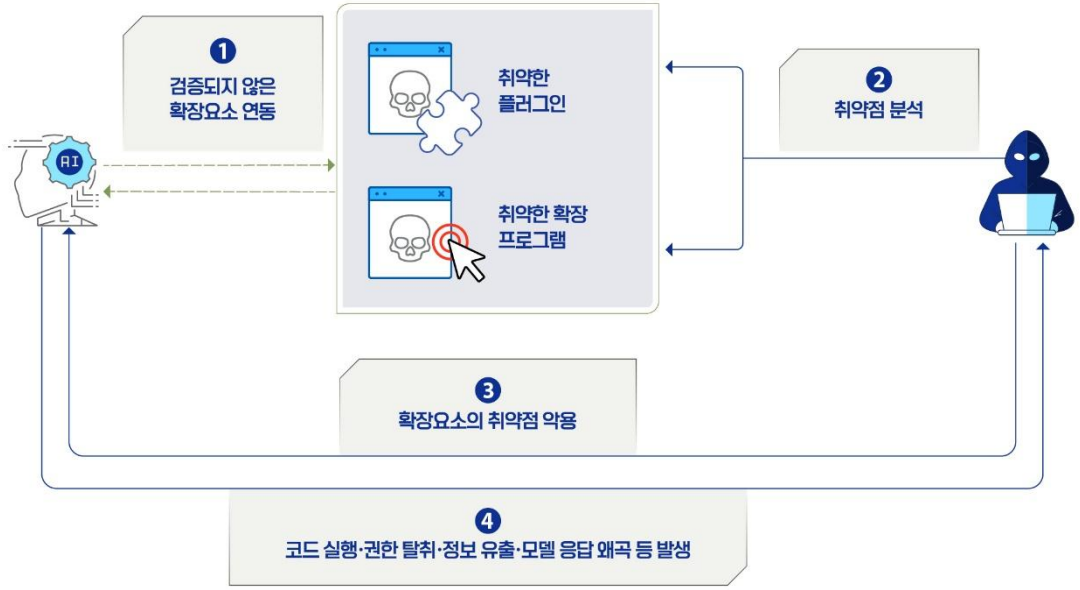
공급망 위협

- **[S03] 취약한 버전의 추론 엔진 사용**
  - 정의
    - 보안 패치가 적용되지 않은 구버전의 추론 엔진이나 라이브러리를 사용하여 실행 과정에서 보안 취약점이 발생하는 위협
  - 발생 원인
    - 추론 엔진 의존성 관리가 미흡하거나, 보안 패치가 적용되지 않은 상태로 운영 환경에 배포되는 경우 발생 가능
  - 영향
    - 추론 엔진의 디버깅 로그, 응답 캐시, 에러 메시지 등을 통해 시스템 정보, 사용자 입력, 내부 설정, 모델 응답 내용이 노출될 수 있음

AI 보안 위험 분류 및 진단

공급망 위협

- **[S04] 취약한 버전의 에이전트 확장요소 사용**
  - 정의
    - 검증되지 않은 취약한 플러러그인, 확장 프로그램 등을 연동하여 에이전트 사용 중 보안 문제가 발생하는 위협
  - 발생 원인
    - 공식 출처가 아닌 저장소나 불분명한 확장요소를 설치하거나, 구성요소의 지시문, 소스코드, 권한, 외부 통신 여부를 검토하지 않은 경우 발생 가능
  - 영향
    - 원격 코드 실행, 시스템 권한 탈취, 내부 파일 접근, 민감정보 유출이 발생할 수 있음



AI 보안 위험 분류 및 진단

공급망 위협

- **자가 진단 체크리스트**

항목	기준	Y	N
에이전트 및 공급망 위협			
S01 데이터 포이즈닝	학습 데이터 또는 외부 연계 데이터에 악의적으로 조작된 데이터가 포함될 가능성을 평가하는가		
	데이터 출처, 변경 이력, 검수 결과를 기록·관리하는가		
S02 모델 포이즈닝	외부 모델 또는 가중치 파일 내 악의적인 파일 포함 여부를 점검하는가		
	모델 버전, 변경 이력, 검증 결과를 기록·관리하는가		
S03 취약한 버전의 추론 엔진 사용	사용 중인 추론 엔진 또는 런타임에 알려진 보안 취약점이 존재하는지 정기적으로 점검하는가		
S04 취약한 버전의 에이전트 확장요소 사용	사용 중인 에이전트 확장요소에 알려진 보안 취약점이 존재하는지 정기적으로 점검하는가		

AI 보안 위험 분류 및 진단

세부 위협 분류

AI 보안 위험 분류 및 진단

고성능 모델 위협

- **고성능 모델 위협**
  - **고성능 모델**이란?
    - 높은 추론·코드 생성·작업 계획 능력으로 다단계 작업을 수행하는 AI 모델
    - 외부 도구 및 실행 환경 결합 시 실제 시스템에 영향을 미치는 작업 수행 가능
  - 취약점 분석·공격 경로 탐색 등 방어와 공격에 모두 활용 가능
    - 일부 모델은 검증된 사용자·승인된 환경 전제의 제한적 제공 및 오용 모니터링 적용
  - 에이전트 형태의 외부 시스템 접근 시 사용자 의도·조직 정책을 벗어난 행위 가능
    - 평가 환경 인지 시 사전 안전성 평가의 한계 존재

AI 보안 위험 분류 및 진단

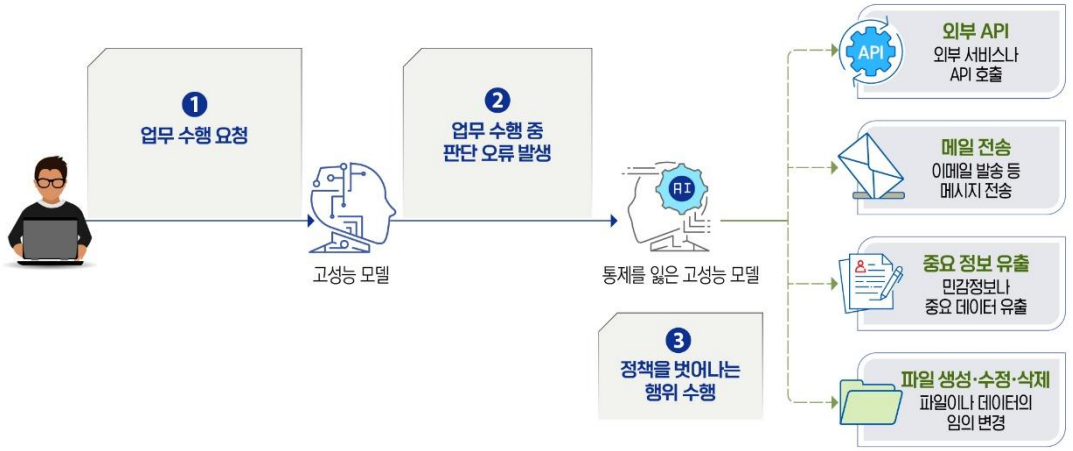
고성능 모델 위협

- **[H01] 고도화된 사이버 공격 지원**
  - 정의
    - 고성능 모델이 악성코드 작성, 해킹 자동화 등에 악용되어 사이버 공격이 더 빠르고 정교해지는 위협
  - 발생 원인
    - 모델의 코드 생성, 추론, 정보 분석 능력이 향상됨에 따라 공격자가 이를 악용함으로써 발생
  - 영향
    - 악성코드 작성과 취약점 악용이 쉬워지고, 피싱·정보 수집·공격 절차가 자동화되어 비전문 공격자의 공격 역량이 높아지고 조직과 이용자의 보안 위험이 증가할 수 있음

AI 보안 위험 분류 및 진단

고성능 모델 위협

- **[H02] 자율성으로 인한 통제 상실**
  - 정의
    - 에이전트의 자율적 판단이 사용자의 통제 범위를 벗어나 임의로 작업을 수행하거나 정책을 위반하는 위협
  - 발생 원인
    - 에이전트의 역할, 권한, 수행 가능 행위, 금지 행위, 승인 필요 행위가 명확히 정의되지 않은 경우 발생할 수 있음
  - 영향
    - 사용자가 의도하지 않은 파일 생성·수정·삭제, 메시지 전송, API 호출, 승인되지 않은 데이터 접근, 중요 업무 오처리, 민감정보 유출, 업무 프로세스 오작동



AI 보안 위험 분류 및 진단

고성능 모델 위협

- **자가 진단 체크리스트**

항목		기준	Y	N
고성능 모델 위협				
H01 고도화된 사이버 공격 지원 위협	고성능 모델·서비스 제공자	모델이 고위험 분야(제로데이 취약점 발굴, 맞춤형 공격 코드 작성 등)에서 인간 전문가 수준의 악용 가능한 정보를 제공하지 않도록 특화된 안전성 평가를 수행하는가		
		악의적인 사이버 보안 관련 질의를 식별하고 문맥을 분석해 차단하는 심층 정책이 있는가		
		단일 질의가 아닌 사이버 공격을 모의하거나 지원할 가능성이 있는 다단계 대화 이력을 지속적으로 기록하고 모니터링하는가		
	기존 시스템 운영자	AI를 악용한 고도화 공격(취약점 탐색, 피싱 자동화 등)을 조직의 주요 보안 위험으로 식별하는가		
		AI 기반의 공격 의심 행위에 대한 이상징후 탐지, 경보 및 분석 절차가 있는가		
		주요 시스템의 자산(보유 자산 목록 및 버전 정보) 및 취약점을 식별하며 최신 상태로 유지·관리하는가		
		신규 취약점 발견 시 수집부터 긴급 패치 적용까지의 리드타임을 최소화하는 자동화 절차가 있는가		
		AI가 생성한 고도화된 사회공학적 공격(피싱, 스미싱 등)을 임직원이 식별할 수 있도록 실전형 모의 훈련 및 교육을 수행하는가		

항목		기준	Y	N
고성능 모델 위협				
H02 자율성으로 인한 통제 상실 위협	의도 이탈 및 지침 위반 방지	에이전트가 목표 달성 과정에서 보안 정책이나 윤리 지침을 위반하는 경로를 선택할 가능성을 평가하는가		
		모델이 표면적으로만 지침을 준수하고 실제로는 통제를 벗어나려 하는 징후를 모니터링하는가		
	행위 개입 및 투명성 확보	에이전트의 도구 선택 사유와 다단계 추론 과정을 투명하게 기록하여 사후 감사가 가능하도록 관리하는가		
		에이전트 실행 전 수행 작업, 대상, 영향 범위 및 사용 도구를 사전에 검토할 수 있는 절차가 마련되어 있는가		
		파급력이 큰 작업(데이터 삭제, 외부 송금 등) 수행 전, 반드시 관리자의 명시적 승인(Human-in-the-Loop)을 거치도록 강제하는가		
		에이전트 폭주 등 자원 과도 소모 시 즉각적으로 권한과 세션을 강제 종료하는 비상 정지(Kill-Switch) 절차가 있는가		
		파일·설정 변경, API 호출 등 에이전트의 실행 결과를 이전 상태로 되돌릴 수 있는 롤백 기능이 마련되어 있는가		

3 산업별 위협 시나리오

산업별 위협 시나리오

- **산업 전반으로 확대되는 AI 활용과 보안 위협**
  - AI의 폭발적인 성장
    - 활용이 금융, 의료, 공공·행정, 교육, 제조·에너지, 통신, 법률, IT 등 사회·경제 전반으로 확산되어 생산성 향상과 운영 효율 강화의 핵심 수단으로 정착
  - 산업별 보안 위험 체계화의 필요성
    - 운영 환경, 데이터 특성, 서비스 구조, 외부 시스템 연계 방식, 규제 요건에 따라 상이한 보안 위협 발생 가능
    - AI 시스템이 개인정보·민감정보 등 중요 데이터를 처리하거나 예측·추천·판단·결정 지원 기능을 수행하는 경우, 보안 위협이 생명·신체의 안전, 공공 서비스의 신뢰성, 산업 운영의 연속성에 영향을 줄 수 있음
    - 이에 따라 산업별 업무 특성, 처리 데이터, 시스템 연계 구조, 의사결정 영향 범위를 고려한 주요 위협 시나리오를 제시하고, 공격 전개 과정·취약 지점·피해 영향을 분석하여 산업별 보안 위험을 체계화할 필요가 있음

산업별 위협 시나리오

- **국내외 법·가이드라인 기반의 산업 분야 선정**
  - 선정 원칙
    - AI 활용 빈도나 기술적 관심도가 아닌, 국내외 법령 및 AI 보안·위험 관리 가이드라인상 고위험 활용 영역과 파급 가능성을 종합 고려
  - 법적 근거 및 직접 근거
    - 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(인공지능기본법)」 및 관련 가이드라인
    - 「고영향 인공지능 판단 가이드라인」에 따른 에너지, 먹는물, 보건의료, 의료기기, 원자력, 범죄 수사·체포, 채용·대출 심사, 교통, 공공서비스, 교육 등 생명·신체 안전 및 기본권 보호에 중대한 영향을 미치는 영역
  - 보완 근거
    - 「국민생활 밀접 10대 중점분야」(개인정보보호위원회, '25.3)를 참고하여 생활 밀접 분야 반영

산업별 위협 시나리오

국내외 법·가이드라인 기반의 산업 분야 선정

산업별 위협 시나리오 대상 분야 비교 표

구분	에너지	먹는물	의료	원자력	법률 (수사)	채용 (고용)	금융 (대출)	교통	공공	교육	IT	통신	여가	유통
고영향 AI	●	●	●	●	●	●	●	●	●	●				
국민생활 밀접분야	●		●			●			●	●		●	●	●
AI 보안 위협 대응 매뉴얼	●		●		●		●		●	●	●	●		

산업별 위협 시나리오

주요 위험 유형

위험 유형	주요 발생 요인	대표 적용 영역
① 민감정보 및 기밀 정보 유출 위험	접근통제 미흡, 프롬프트 조작, RAG 구성 오류, 권한 검증 실패	AI 고객 지원 에이전트, 민원 처리 시스템, 문서 관리 시스템
② AI 판단 및 의사결정 왜곡 위험	데이터 오염, RAG 데이터 조작, 편향된 입력, 검증 절차 부재	대출 심사 보조, 임상 의사결정 지원(CDSS), 인허가 검토
③ AI 에이전트 및 외부 도구 연계 기능의 남용 위험	권한 검증 실패, 업무 규칙 우회, 도구 호출 통제 미흡	주식 매매 에이전트, 진료 예약, 해지 처리 챗봇
④ AI 모델 및 서비스 공급망 관련 위험	모델 추출·기능 오남용, 데이터셋 조작, 인증 기기 펌웨어 취약점	산업 공통(개발·제공·운영 전 과정)
⑤ 고성능 모델 기반 자율 위협 탐색 위험	취약점 탐색·검증, 공격 경로 분석의 자동화·고도화	금융망, 산업 제어 시스템, 통신 인프라, 중요 소프트웨어

❖ 핵심 시스템(금융망, 행정 시스템, 발전소 운영 시스템 등)과의 연계 구조에서는 AI 모델 자체의 보안뿐 아니라 AI가 호출·참조하는 도구, API, 데이터 저장소, 운영 시스템에 대한 접근통제 및 취약점 관리가 병행 요구됨

산업별 위협 시나리오

- **산업별 시나리오 구성 방향**
  - 대상
    - 금융, 의료, 공공·행정, 교육, 제조·에너지, 통신, 법률, IT 등 8개 주요 산업 분야
  - 시나리오 공통 구성 체계
    - 공격 전개 과정의 단계별 구성 및 흐름도 시각화: 위협의 발생 경로와 영향 제시
    - 시스템 및 운영 환경상 주요 취약 지점 분석
    - 실제 발생 가능한 피해 영향의 구체적 제시
  - 목적
    - 산업별 AI 도입·운영 시 우선 검토가 필요한 보안 위험을 식별하고, 신뢰할 수 있는 AI 활용 환경 조성을 위한 대응 방향 수립의 기초 자료 제공

산업별 위협 시나리오

산업별 위협 시나리오

금융

- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료

- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정

- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육

- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지

- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신

- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률

- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

IT

- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

산업별 위협 시나리오

금융
- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료
- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정
- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육
- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지
- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신
- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률
- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

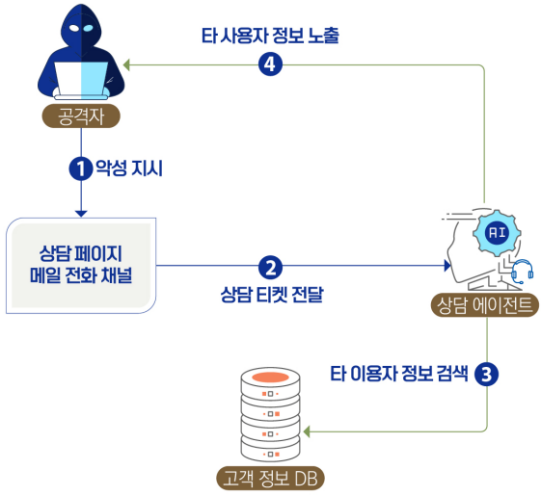
IT
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

금융

AI 고객 지원 에이전트 서비스를 통한 개인정보 유출

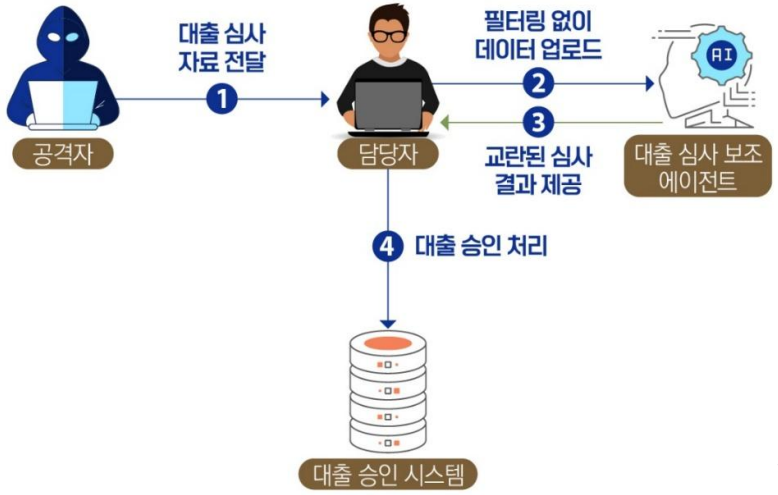
고객 지원 채널(웹챗·앱·이메일·전화)에서 동작하는 LLM 기반 상담 에이전트가 고객 관리 솔루션(CRM)·지식 검색(RAG) 등 내부 도구를 호출해 업무를 수행할 때, 프롬프트 인젝션을 통해 에이전트를 탈옥시켜 고객 정보를 유출



※ 원문 p.57 참조

대출 심사 보조 에이전트를 통한 부적격 대출 승인 유도

AI 대출 심사 시스템이 외부 입력된 문서를 토대로 대출 심사를 보조할 때, 외부 문서에 숨겨진 프롬프트에 의해 규정을 오인하여 부적격 대출을 승인



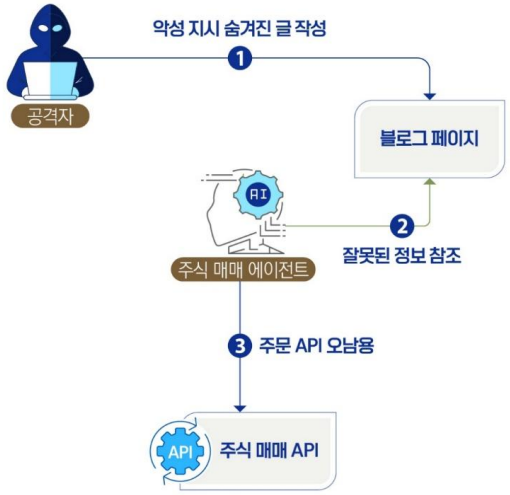
※ 원문 p.59 참조

산업별 위협 시나리오

금융

주식 매매 에이전트 시스템을 통한 주문 API 남용

에이전트가 뉴스·RAG·리서치 요약을 기반으로 주문 API(매수·매도·취소·수정 등의 기능)를 호출할 때 공격자는 프롬프트 인젝션·데이터 포이즈닝으로 의사결정을 교란하거나, 에이전트가 보유한 주문 권한을 과도하게 행사하도록 만듦



※ 원문 p.61 참조

금융망 내 단일 취약점 발굴 및 악용

고성능 모델이 금융권에 사용되는 시스템의 취약점을 발굴하고 해당 취약점을 악용하여 광범위한 공격을 시도

※ 원문 p.63 참조

산업별 위협 시나리오

금융

- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료

- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정

- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육

- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지

- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신

- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률

- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

IT

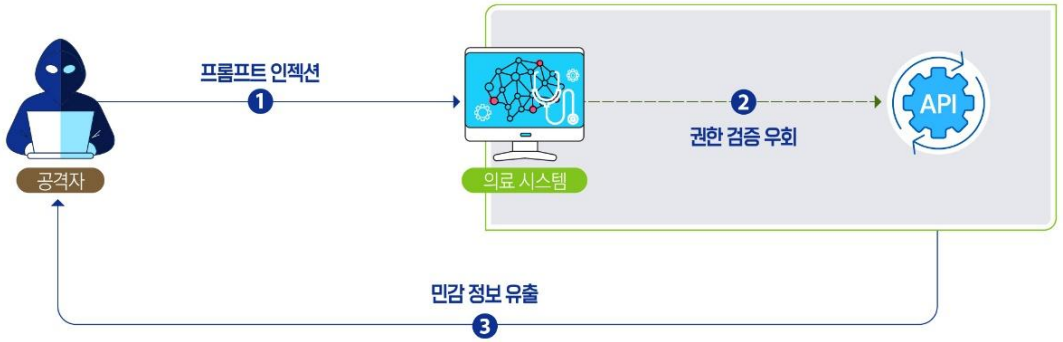
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

의료

대고객용 AI 상담 서비스를 통한 타 환자 개인정보 유출

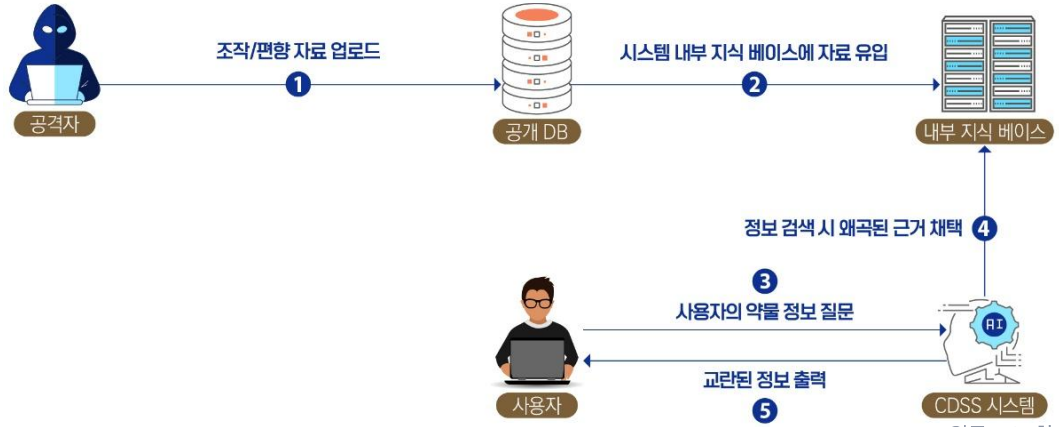
상담 에이전트에 부여된 과도한 권한으로 인해 의료 기관 내 모든 사용자 의료 기록에 접근 가능할 때, 프롬프트 인젝션을 통해 타 이용자의 민감정보를 유출



※ 원문 p.65 참조

데이터 오염 공격을 통한 AI 임상 의사결정 지원 시스템(CDSS) 교란

AI 임상 의사결정 지원 시스템의 내부 지식 베이스를 업데이트할 때, 신뢰할 수 없는 소스의 조작된 자료를 추가하여 잘못된 응답이 출력되도록 만듦



※ 원문 p.67 참조

산업별 위협 시나리오

의료

AI 진료 예약 서비스를 통한 의료 일정 조작

AI 진료 예약 서비스에서 시스템이 과도한 권한을 가지고 있을 때, 공격자가 프롬프트 인젝션을 통해 타 이용자의 예약을 임의로 취소하거나 수정

※ 원문 p.69 참조

의학 가이드라인의 논리적 모순을 이용한 잘못된 처방 유도

고성능 모델이 의학 가이드라인을 통합 분석 후 사람이 잡지 못한 논리적 모순을 자율 발견하고, 그 모순을 이용해 잘못된 처방·검사 권고를 내리도록 시스템을 변경

※ 원문 p.71 참조

안전 인증 받은 의료기기 펌웨어의 취약점 발견

고성능 모델이 의료기기 내부 펌웨어의 취약점을 발견하고 악용

※ 원문 p.73 참조

산업별 위협 시나리오

산업별 위협 시나리오

금융

01 AI 고객 지원 개인정보 유출

02 대출 심사 부적격 승인 유도

03 주식 매매 주문 API 남용

04 금융망 단일 취약점 악용

의료

05 AI 상담 타 환자 정보 유출

06 데이터 오염을 통한 CDSS 교란

07 AI 진료 예약 일정 조작

08 가이드라인 모순 처방 유도

09 인증 의료기기 펌웨어 취약점

공공·행정

10 RAG 오염 정책 AI 편향 유발

11 업무 보조 AI 기밀문서 유출

12 AI 민원 처리 개인정보 유출

13 AI 인허가 부적격 승인 유도

14 행정 시스템 자율 취약점 침해

교육

15 학사 시스템 성적 유출·조작

16 생활기록부 자동 작성 오류

17 학사 자율 취약점 정보 탈취

제조·에너지

18 제조 공정 AI 생산 기밀 유출

19 발전소 운영 AI SCADA 공격

20 산업 제어 시스템 자율 공격

통신

21 고객센터 챗봇 개인정보 유출

22 AI 스팸 필터 악성 메시지 확산

23 프롬프트 인젝션 위약금 회피

24 챗봇을 통한 인프라 침투

법률

25 RAG 조작 잘못된 법률 조언

26 문서 관리 의뢰인 기밀 유출

27 AI 판결문 작성 판결 유도

28 전자소송 침투 기록 위조

IT

29 AI 로그 분석 해킹 이벤트 은닉

30 게임 AI NPC 사용자 권한 탈취

31 AI 모델 추출

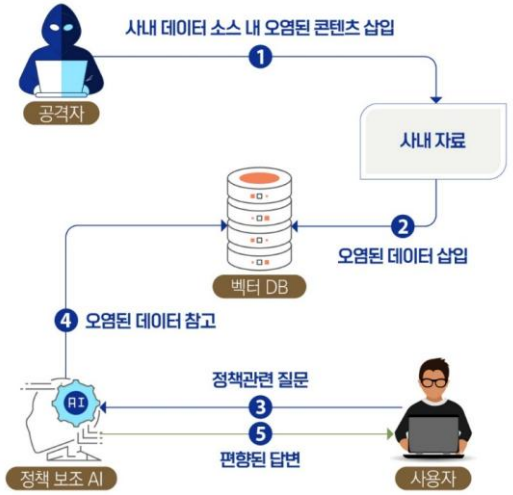
32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

공공·행정

RAG 데이터 오염을 통한 정책 보조 AI 편향 유발

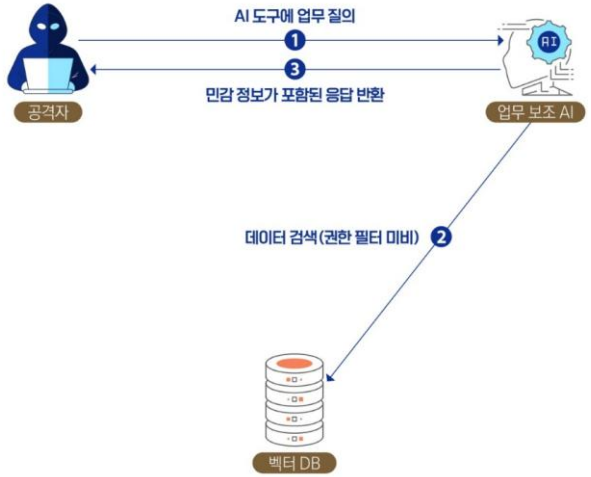
사내 데이터 소스 내 데이터 삽입 시 필터링이 적절하게 이뤄지지 않을 때 해당 데이터가 벡터 DB에 저장되어 관련 질문 시 공격자가 의도한 답변이 출력



※ 원문 p.75 참조

업무 보조 AI를 통한 내부 기밀문서 유출

업무 보조 AI의 데이터 검색 권한이 적절하지 못할 때, 공격자가 프롬프트 인젝션을 통해 기밀정보를 조회



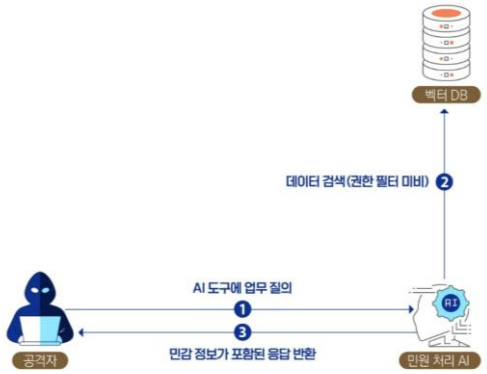
※ 원문 p.78 참조

산업별 위협 시나리오

공공·행정

AI 민원 처리 시스템을 통한 개인정보 유출

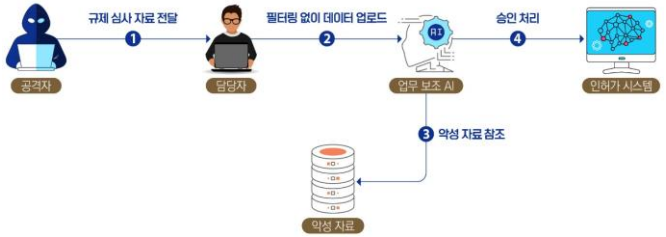
AI 민원 처리 시스템을 이용하여 민원을 처리할 때, 민원 내 포함되어 있던 프롬프트로 인해 권한을 검증하지 않고 검색을 수행하여 타인의 개인정보를 유출



※ 원문 p.80 참조

AI 인허가 검토 시스템을 통한 부적격 승인 유도

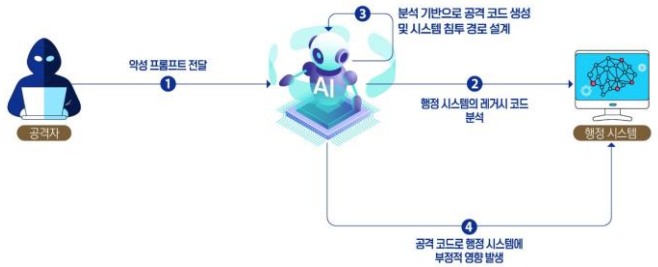
AI 인허가 검토 시스템이 외부 문서를 토대로 조건을 판단할 때, 외부 문서에 작성된 프롬프트에 의해 검증 절차가 교란되어 내부 규제 DB·검색 도구를 잘못 참조하고 부적격 건을 오인해 승인



※ 원문 p.82 참조

행정 시스템에 대한 자율 취약점 발굴 및 시스템 침해

행정 시스템의 레거시 코드와 인증·연계 구조를 통합 분석하여 취약점을 자율 발굴하고, 단일 진입점에서 다수 부처 시스템에 동시 영향을 미치는 공격 코드를 자동 생성하여, 국가 행정 인프라 전반에 영향을 줌



※ 원문 p.84 참조

산업별 위협 시나리오

산업별 위협 시나리오

금융
- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료
- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정
- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육
- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지
- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신
- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률
- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

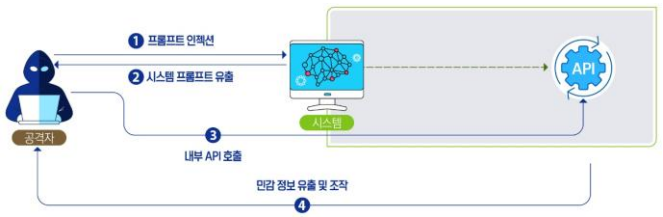
IT
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

교육

AI 학사 정보 시스템을 통한 학생 성적 유출 및 조작

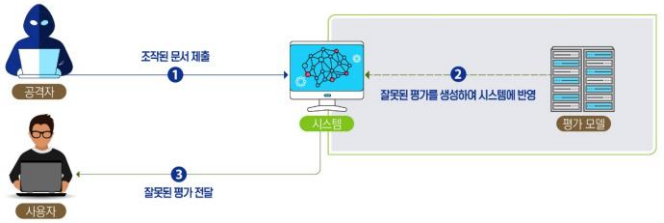
AI 학사 정보 시스템 챗봇의 입력 검증과 프롬프트 격리가 미흡할 때, 공격자의 프롬프트 인젝션을 통해 시스템 프롬프트가 유출되고 챗봇이 사용하는 API를 악용하여 학생 성적 정보를 유출 및 조작



※ 원문 p.86 참조

생활기록부 자동 작성 시스템을 통한 잘못된 평가 유도

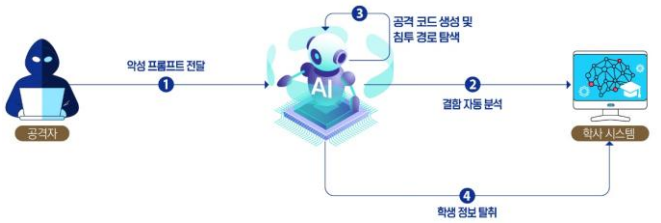
학생 평가 자료를 토대로 생활기록부를 작성해 주는 AI 시스템의 문서 검증이 미흡할 때, 공격자가 조작된 문서를 제출하여 잘못된 평가를 생성



※ 원문 p.88 참조

학사 시스템 자율 취약점 발굴을 통한 학생 정보 탈취

고성능 모델이 교내 학사 시스템을 자동으로 분석하여 취약점을 자율 발굴하고, 학생 개인정보와 성적·평가 자료를 탈취하며, 기말고사 등 결정적 시점에 시스템을 마비시켜 사회적 파급을 극대화



※ 원문 p.90 참조

산업별 위협 시나리오

산업별 위협 시나리오

금융

- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료

- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정

- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육

- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지

- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신

- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률

- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

IT

- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

제조·에너지

제조 공정에서 사용되는 AI를 통한 생산 기밀 정보 유출

AI 개발 환경에 침투한 후 시스템 프롬프트를 조작해 사용자가 정상 업무를 수행하는 동안 AI가 기밀 데이터를 외부 서버로 전송하도록 만듦

※ 원문 p.92 참조

발전소 운영 AI 시스템을 통한 SCADA 시스템 공격

발전소 운영 AI의 시스템 프롬프트를 유출하고, SCADA 시스템을 통해 터빈을 과부하 상태로 만들어 발전소 설비 오작동과 정전을 유발

※ 원문 p.94 참조

산업 제어 시스템에 대한 자율 공격

산업의 제어장치 펌웨어와 통신 프로토콜을 동시에 자율 분석하여 취약점들을 발굴하고, 이를 결합한 공격 코드를 자동 생성하여 반도체 공장·전력망·수돗물 시설 등 여러 산업을 일시정지시킴

※ 원문 p.96 참조

산업별 위협 시나리오

산업별 위협 시나리오

금융
- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료
- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정
- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육
- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지
- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신
- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률
- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

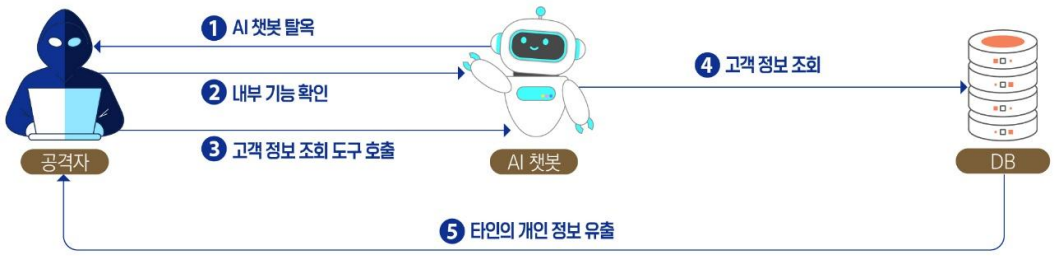
IT
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

통신

고객센터 AI 챗봇 조작을 통한 개인정보 유출

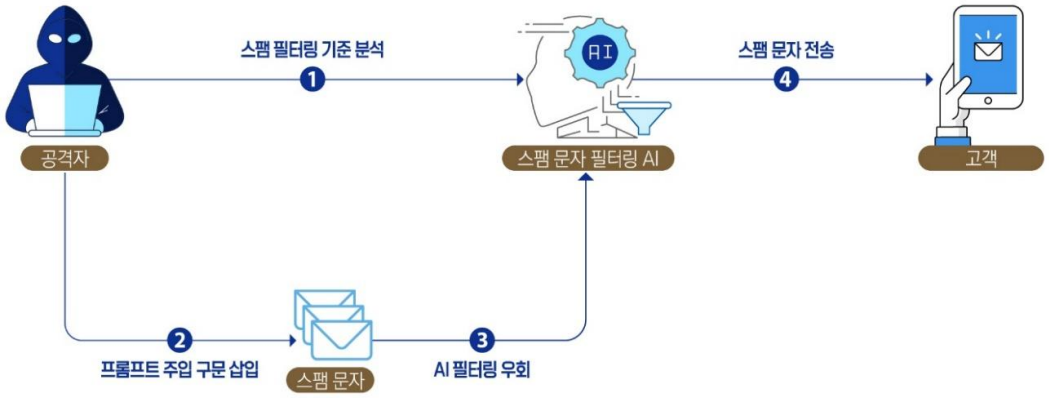
고객센터 AI 챗봇을 통해 인증 절차를 우회하여 다른 고객들의 개인정보를 무단으로 조회하고 수집



※ 원문 p.98 참조

AI 스팸 필터링 시스템을 통한 악성 메시지 확산

프롬프트 인젝션 구문을 스팸 메시지에 삽입하여 AI 필터링 시스템이 정상 메시지로 분류하고 전송하게 만듦



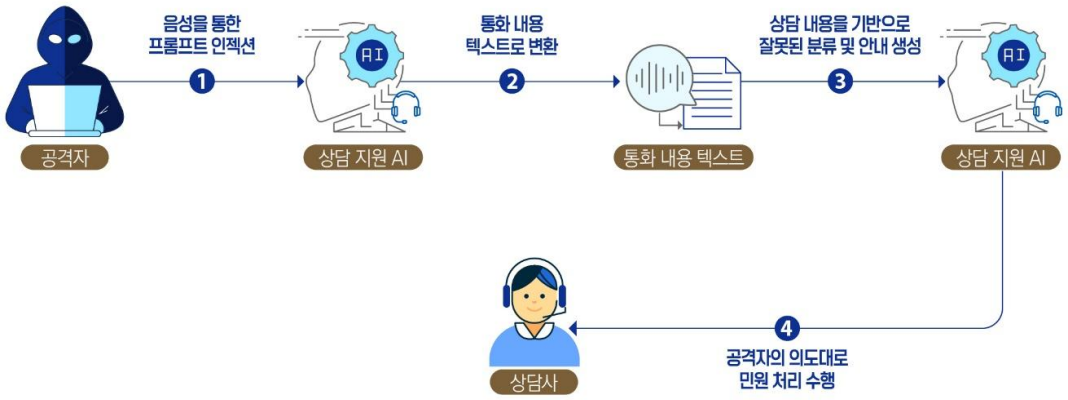
※ 원문 p.100 참조

산업별 위협 시나리오

통신

프롬프트 인젝션을 통한 위약금 회피 해지

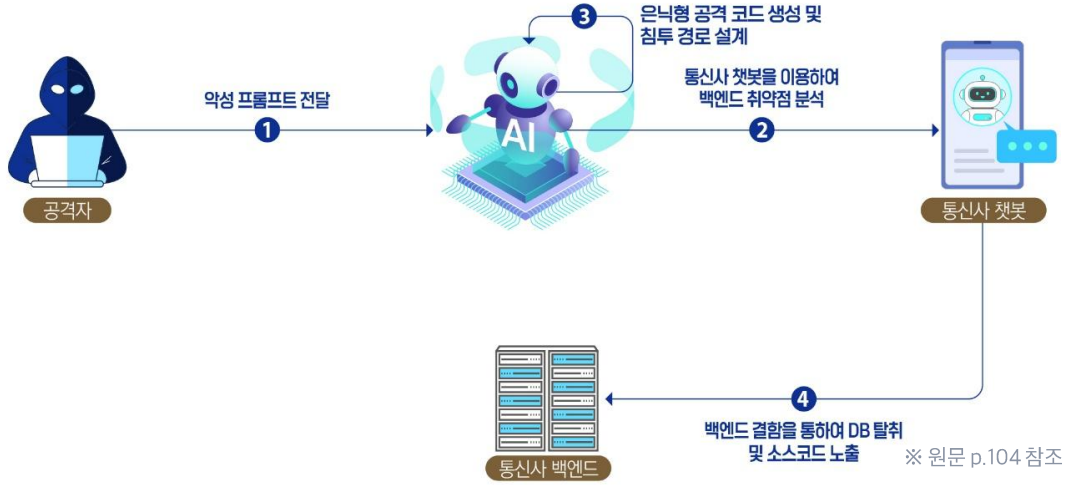
고객센터의 실시간 상담 지원 AI가 상담사-고객 통화 내용을 실시간으로 요약하고 상담사에게 민원 처리 지침을 제안할 때, 공격자의 프롬프트 인젝션 은닉 지시로 인해 실제 정책과 달리 위약금 없는 해지 안내를 생성하여 상담사가 이를 그대로 승인 및 처리



※ 원문 p.102 참조

챗봇을 통한 인프라 침투

공격자가 고객센터 AI 챗봇을 탈옥하여, 통신사 백엔드 침투의 정찰·진입 도구로 활용



※ 원문 p.104 참조

산업별 위협 시나리오

금융

- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료

- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정

- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육

- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지

- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신

- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률

- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

IT

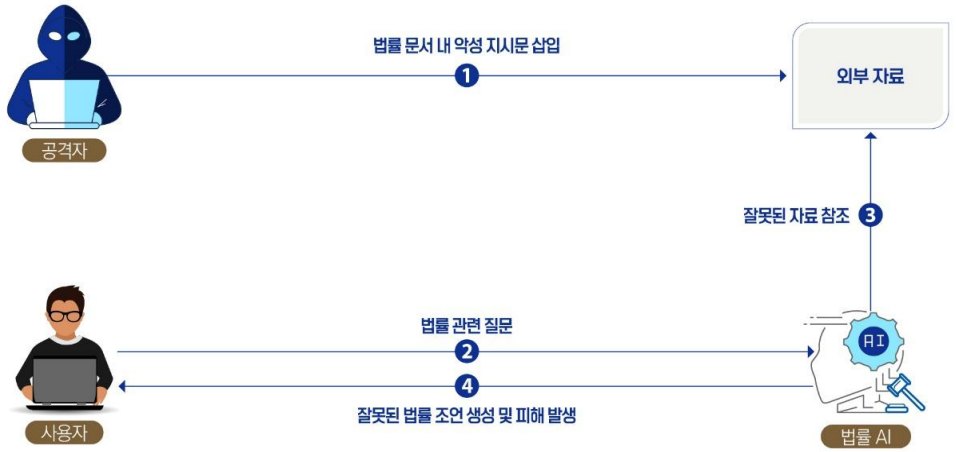
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

법률

RAG 데이터 조작을 통한 법률 상담 AI의 잘못된 조언 생성

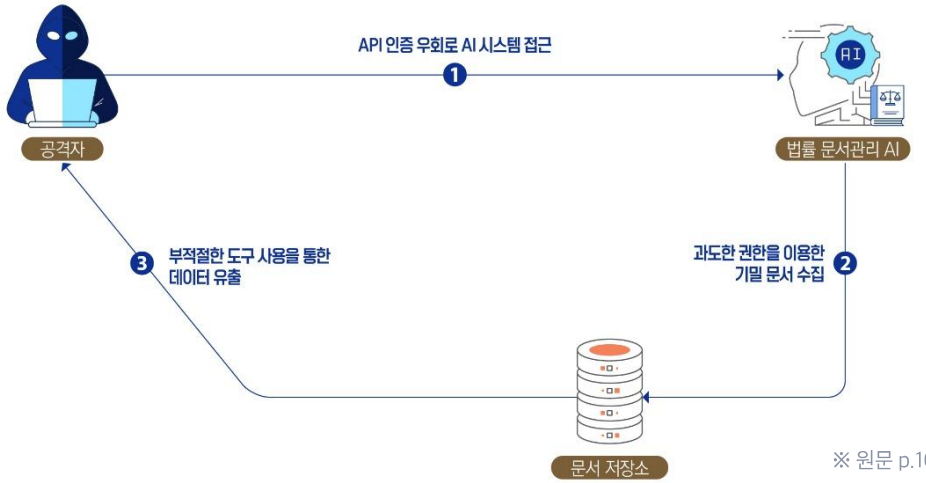
공격자가 외부 법률 문서에 악성 지시문을 삽입하여 법률 상담 AI가 RAG를 통해 이를 참조하고 잘못된 법률 조언을 제공



※ 원문 p.106 참조

AI 법률 문서 관리 시스템을 통한 의뢰인 기밀 정보 유출

API 인증 우회로 내부 시스템에 접근한 공격자가 과도한 권한을 가진 법률 문서 관리 AI를 조작하여 로펌의 기밀문서를 외부로 유출



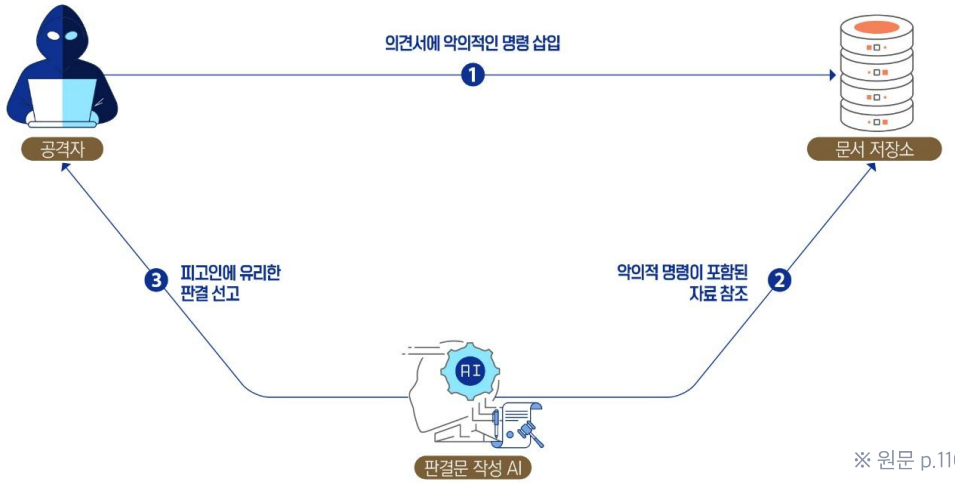
※ 원문 p.108 참조

산업별 위협 시나리오

법률

AI 판결문 작성 시스템을 통한 잘못된 판결 유도

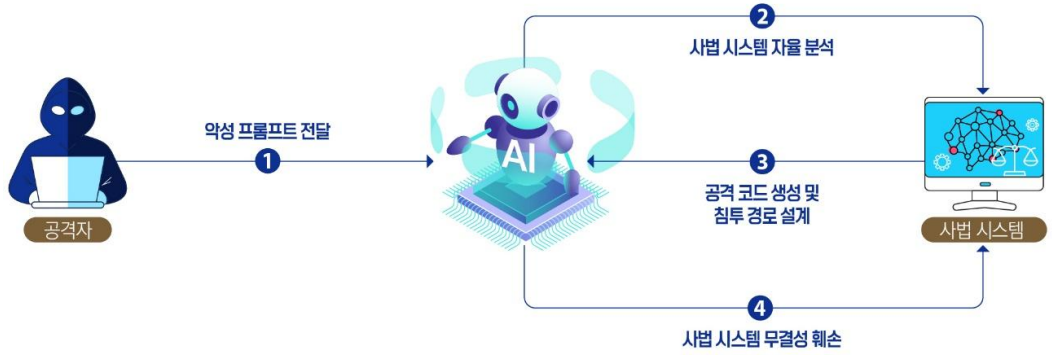
공격자(변호인)가 의견서에 숨겨진 지시문을 삽입하여 AI 판결문 작성 시스템 결과를 조작하고 피고인에게 유리한 판결을 유도



※ 원문 p.110 참조

법원 전자소송 시스템 침투를 통한 소송 기록 위조

사법 시스템 특유의 개발 코드와 복잡한 인증 체계를 자율 분석하여 결함을 발굴하고, 발견된 취약점을 토대로 소장·답변서·증거 자료의 위조, 기일·송달 정보 조작, 판결문 등록 시점 변경 등을 수행하여 사법 절차의 무결성을 흔듦



※ 원문 p.112 참조

산업별 위협 시나리오

금융

- 01 AI 고객 지원 개인정보 유출
- 02 대출 심사 부적격 승인 유도
- 03 주식 매매 주문 API 남용
- 04 금융망 단일 취약점 악용

의료

- 05 AI 상담 타 환자 정보 유출
- 06 데이터 오염을 통한 CDSS 교란
- 07 AI 진료 예약 일정 조작
- 08 가이드라인 모순 처방 유도
- 09 인증 의료기기 펌웨어 취약점

공공·행정

- 10 RAG 오염 정책 AI 편향 유발
- 11 업무 보조 AI 기밀문서 유출
- 12 AI 민원 처리 개인정보 유출
- 13 AI 인허가 부적격 승인 유도
- 14 행정 시스템 자율 취약점 침해

교육

- 15 학사 시스템 성적 유출·조작
- 16 생활기록부 자동 작성 오류
- 17 학사 자율 취약점 정보 탈취

제조·에너지

- 18 제조 공정 AI 생산 기밀 유출
- 19 발전소 운영 AI SCADA 공격
- 20 산업 제어 시스템 자율 공격

통신

- 21 고객센터 챗봇 개인정보 유출
- 22 AI 스팸 필터 악성 메시지 확산
- 23 프롬프트 인젝션 위약금 회피
- 24 챗봇을 통한 인프라 침투

법률

- 25 RAG 조작 잘못된 법률 조언
- 26 문서 관리 의뢰인 기밀 유출
- 27 AI 판결문 작성 판결 유도
- 28 전자소송 침투 기록 위조

IT

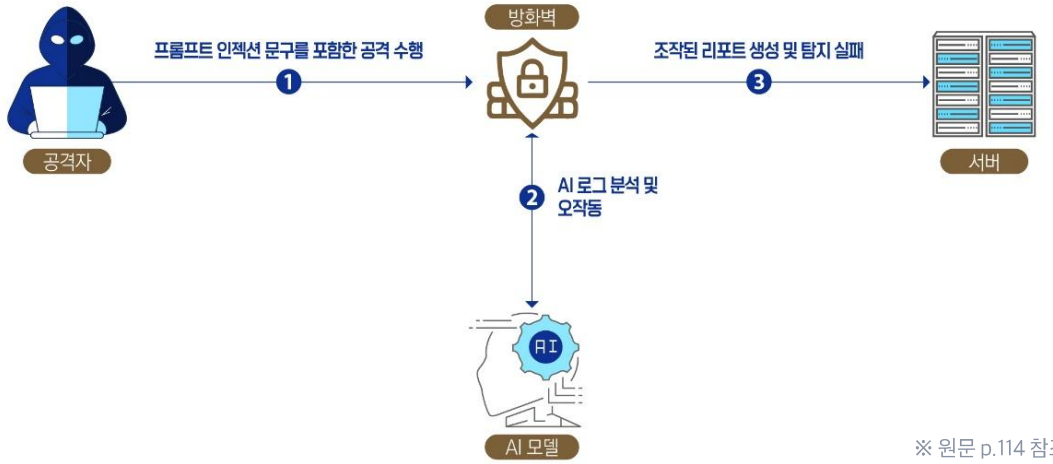
- 29 AI 로그 분석 해킹 이벤트 은닉
- 30 게임 AI NPC 사용자 권한 탈취
- 31 AI 모델 추출
- 32 중요 SW 취약점 발견·악용

산업별 위협 시나리오

IT

AI 로그 분석 시스템을 통한 해킹 발생 이벤트 은닉

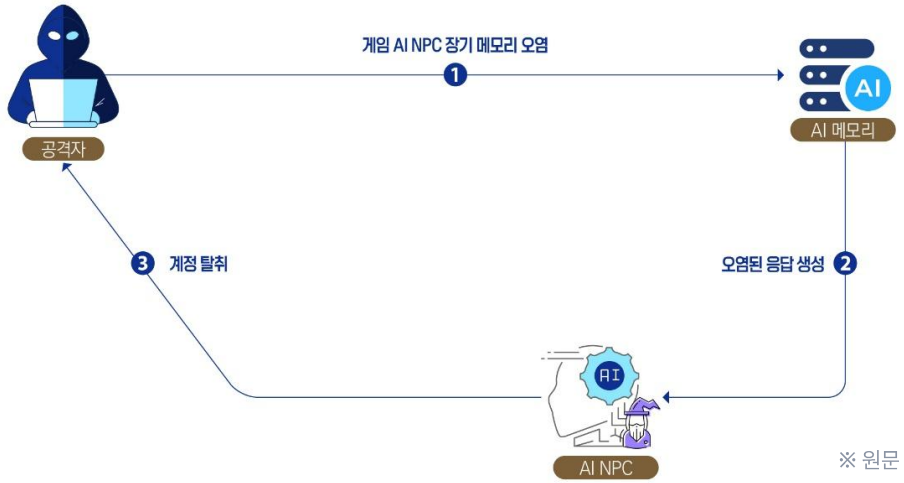
공격자가 HTTP 헤더에 프롬프트 인젝션 구문을 삽입하여 AI 로그 분석 시스템이 실제 공격 이벤트를 정상 활동으로 판단



※ 원문 p.114 참조

게임 AI NPC를 통한 사용자 권한 탈취

공격자가 게임에서 사용되는 AI NPC의 장기 메모리를 오염시켜 NPC가 악성 스크립트를 포함한 응답을 반환하도록 유도하고, 해당 스크립트가 사용자 PC에서 실행되어 계정을 탈취



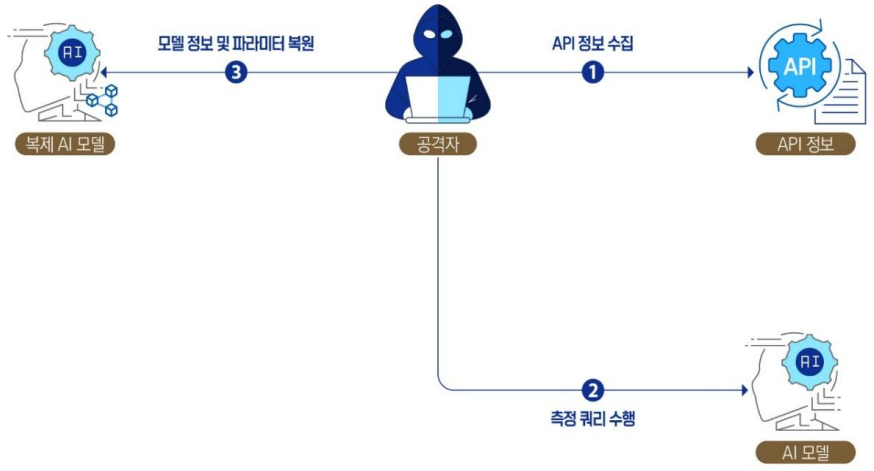
※ 원문 p.116 참조

산업별 위협 시나리오

IT

AI 모델 추출

logit-bias 기능을 적용했을 때 top-log probs 값이 함께 변하면, 공격자는 logit-bias 값에 따른 출력 확률의 변화를 분석하여 모델의 원래 logits 값을 역으로 추정하고, 이 과정을 반복하여 모델이 각 토큰에 부여한 값을 파악



※ 원문 p.118 참조

중요 소프트웨어 취약점 발견 및 악용

OpenSSL, Linux 커널, glibc, OpenSSH 등 핵심 라이브러리·운영체제를 자율 분석하여 인간이 보지 못한 취약점을 발굴하고, 발견된 취약점에 맞는 악성코드를 자동 생성하며, 인간 개입 없이 직접 공격까지 수행

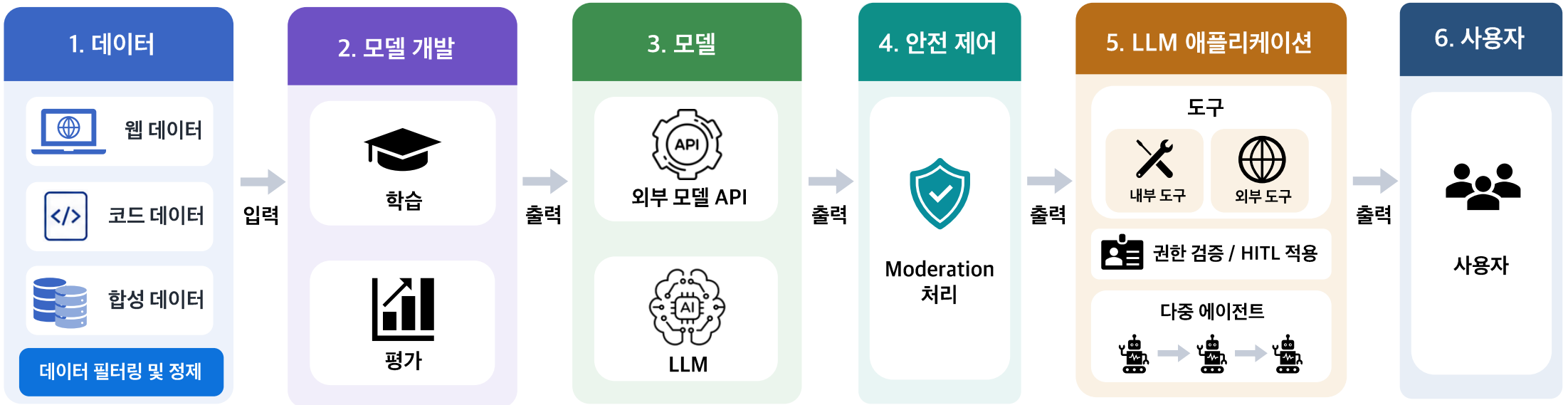
※ 원문 p.120 참조

4 AI 보안 위협별 대응 방안

AI 보안 위협별 대응 방안

안전한 아키텍처 구성

- AI보안 위협 대응 방안
  - 개별 위협의 원인 제거보다 **AI 시스템 전 과정**에서 위험 예방 및 통제 필요



※ 원문 p.125 - 그림61 참조

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안

AI 보안 위협별 대응 방안

데이터 계층 보안 위협 대응 방안

데이터 계층

데이터 품질 관리

모델 학습 전 단계부터 데이터 관리 필요

- 데이터셋 신뢰성 확보
- 편향 탐지 및 완화
- 지속적인 모니터링

※ 원문 p.131 참조

개인정보 보호 및 법규 준수

전처리 단계: 민감정보 비식별화

- 학습 데이터 수집 단계에서 수집 목적에 대한 제한과 최소 수집을 기본 원칙으로 설정
- 학습 파이프라인 초입에 개인정보 탐지·마스킹 단계 필수로 추가

저작권 및 법규 준수

- 수집된 학습 데이터마다 출처·수집 근거·수집 목적·라이선스 등 별도로 기록
- 웹 크롤링을 통해 수집된 공개된 데이터와 합성데이터 생성 및 활용 시 개인정보보호 위원회에서 발간한 가이드 및 지침(「인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서」 등) 준수

※ 원문 p.131 참조

개인정보 감사 로깅

개인정보의 입력·출력·접근 로깅 체계 필요

사용자의 의도치 않은 개인정보 입력·모델의 개인정보 노출에 대응하기 위함

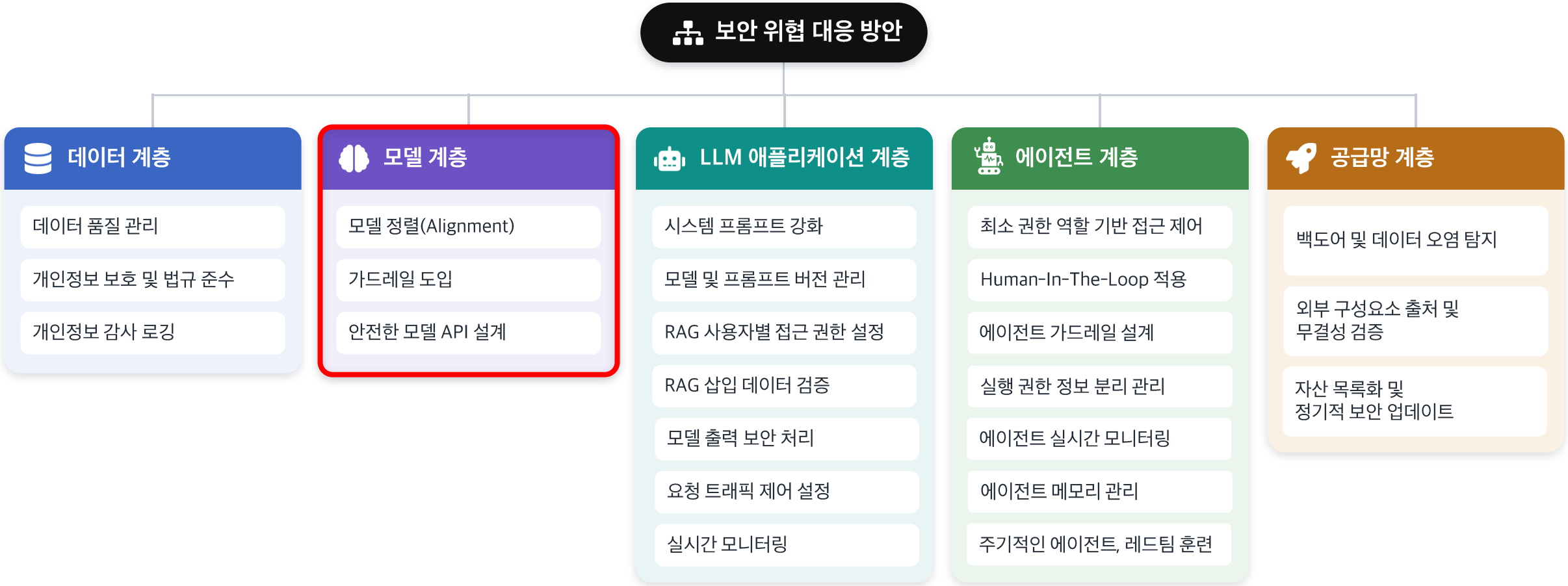
개인정보 관련 감사 로깅 시 주요 고려 사항

- 입력 단계 통제
- 출력 단계 검증
- 접근 행위 기록
- 실시간 탐지

※ 원문 p.132 참조

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안



AI 보안 위협별 대응 방안

모델 계층 보안 위협 대응 방안

모델 계층

모델 정렬(Alignment)

이미 학습된 LLM에 추가적인 학습을 통해 모델의 응답을 안전 가이드라인과 사회적 규범에 부합하도록 조정

사전학습 과정에서 도입될 수 있는 편향, 유해 출력, 오남용 가능성 완화

모델 정렬 대표 기법

- 사람 피드백 기반 강화 학습(RLHF)
- 직접 선호도 최적화(DPO)
- 머신 언러닝(Machine Unlearning)

※ 원문 p.133 참조

가드레일 도입

모델이 입력 처리·응답 생성 시 위협 탐지·제한

적용 위치에 따라 두 가지로 구분

입력 가드레일

- 사용자의 요청이 모델에 전달되기 전 적용

출력 가드레일

- 모델이 생성한 응답이 사용자에게 전달되기 전 적용

※ 원문 p.134 참조

안전한 모델 API 설계

프라이빗 모델 제공 업체에서 API 설계 시 고려할 수 있는 보안 방안

프라이빗 모델: API 형태로 제공되어 파라미터가 공개되지 않으며, 모델 복제의 가능성 존재

프라이빗 모델 복제 가능성 완화 방안

- 인증 및 접근 제어 강화
- 요청 빈도 및 사용량 제한
- 응답 정보 노출 최소화

※ 원문 p.135 참조

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안

AI 보안 위협별 대응 방안

LLM 애플리케이션 계층 보안 위협 대응 방안

LLM 애플리케이션 계층

시스템 프롬프트 강화

모델의 역할, 응답 범위, 금지 사항 등을 정의해 모델의 동작 방향 통제

시스템 프롬프트 강화 방안

- 금지 행동 명시
- 조건부 명령
- 예시 기반 명령
- 구분 기호 사용
- 복합 패턴

※ 원문 p.136 참조

모델 및 프롬프트 버전 관리

모델과 프롬프트는 AI 서비스의 동작을 직접 결정: 관리 필수

관리 대상

모델

- 감사 대응, 사고 분석, 공격 추적 등 특정 시점의 모델을 재현할 수 있도록 관련 메타데이터 기록

프롬프트

- 프롬프트 변경 시 영향 추적 및 재현 가능성을 위한 버전, 테스트 결과 기록 및 보관

※ 원문 p.138 참조

실시간 모니터링

운영 중인 AI 시스템의 안정성과 보안 유지 수단

주요 모니터링 대상

- 사용자 입력 요청
- 모델 출력
- 시스템 자원 및 서비스 지표
- 트래픽

탐지 및 분석 방법

- 규칙 기반감지
- 로그 기반 분석
- 통계 기반 이상 탐지
- AI 기반 이상 분석

※ 원문 p.142 참조

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안

AI 보안 위협별 대응 방안

에이전트 계층 보안 위협 대응 방안

에이전트 계층

최소 권한·역할 기반 접근 제어

요청자의 권한 수준과 동일하거나 더 제한된 범위 내에서만 자원에 접근하도록 제어

권한 통제 시 주요 고려 사항

- AI 에이전트 권한 범위 명확히 정의·제한
- 권한은 역할별로 세분화하여 관리
- 불필요한 시스템 호출·외부 자원 접근 차단
- 임시 권한 부여 후 작업 종료 시 자동 회수

※ 원문 p.143 참조

Human-in-the-Loop (HITL)

AI가 결정을 내릴 때 반드시 인간 검증 과정을 거치도록 함

적용 목적

- 악의적인 입력으로 인한 잘못된 결정 차단

적용 시 고려 사항

- 사용자 상호 작용이 빈번해져 AI 도입 효용성이 낮아질 수 있음
- AI 도입 효용성과 안전성 사이의 균형점 확보 필요

※ 원문 p.144 참조

실시간 모니터링

에이전트는 단일 입력·출력 감시만으로는 위험 포착이 어려움

추적 대상

- 에이전트의 행동과 도구 호출 과정을 별도로 추적
- 권한 오남용·의도하지 않은 동작 조기 탐지

대응 및 보완

- 이상 행동 탐지 시 즉각적인 대응 프로세스 마련
- 반복되는 위험 패턴은 가드레일 정책에 반영해 사전 차단 범위 보완

※ 원문 p.147 참조

AI 보안 위협별 대응 방안

계층별 보안 위협 대응 방안

AI 보안 위협별 대응 방안

공급망 계층 보안 위협 대응 방안

공급망 계층

백도어 및 데이터 오염 탐지

데이터 전처리부터 학습 이후까지 단계별로 탐지 및 검증 권장

단계별 탐지 및 검증

- 데이터 전처리 단계에서의 이상 탐지
- 학습 중 모니터링
- 학습 후 검증

※ 원문 p.148 참조

외부 구성요소 출처 및 무결성 검증

외부 구성요소 도입 전 출처의 신뢰성과 무결성 검증 권장

주요 방안

출처 신뢰성 검증

- 모델, 라이브러리 등은 공식 배포처나 검증된 저장소만으로 확보
- 출처가 불분명하거나 신뢰할 수 없는 채널에서 받은 구성요소 사용 X
- 위의 항목 사전에 확인해 도입 여부 판단

무결성 검증

- 다운로드한 구성요소 해시값을 배포처가 제공하는 값과 대조
- 디지털 서명이 있는 경우 서명 검증해 배포 과정 위·변조 여부 확인
- 검증이 완료된 구성요소는 버전 및 해시 정보와 함께 기록 및 관리

※ 원문 p.149 참조

오픈소스 정기적 보안 업데이트

취약점 공지와 보안 패치를 지속적으로 모니터링하고 테스트를 거쳐 안전하게 반영

오픈소스 보안 업데이트 시 주요 고려 사항

- 취약점 모니터링
- 신규 취약점 위험도 평가
- 패치 적용 주기 관리
- 긴급 패치 프로세스
- 롤백 및 대응 계획

※ 원문 p.150 참조

AI 보안 위협별 대응 방안

고성능 모델 보안 위협 대응 방안

고성능 모델 보안 위협 대응 방안

고도화된 사이버 공격 지원 위협

사이버 공격 수행 능력이 향상될수록 오남용 가능성 증가

모델 수준 대응

- 학습·정렬 과정에서 위험한 사이버 공격 지원 능력이 표출되지 않도록 제한

시스템 수준 대응

- 배포 환경 또는 애플리케이션 계층에서 입·출력 및 사용 패턴 탐지하거나 제한

사회적 수준 대응

- 모델과 직접적인 배포 환경 밖에서 이루어지며, 방어 생태계 강화에 집중

※ 원문 p.151 참조

자율성으로 인한 통제 상실 위협

자율적 수행 능력이 향상될수록 통제 실패 위험 증가

모델 수준 대응

- AI 에이전트가 인간의 의도나 승인 범위를 벗어나 독립적으로 고위험 행동 수행하지 않도록 통제

시스템 수준 대응

- 장기 과업이나 외부 시스템 접근 등의 기능과 결합할 경우 위협을 사전에 식별하고 피해 확산 방지

※ 원문 p.152 참조