인공지능(AI) 보안 안내서

2025. 12

과학기술정보통신부

KISA 한국인터넷진흥원

일러두기

- 본 안내서는 AI 기술 및 서비스와 관련된 보안 위협을 선제적으로 예방하고 데이터와 시스템의 안전성을 확보하며, 이를 통해 개발자, 서비스 제공자, 이용자 모두가 신뢰할 수 있는 AI 환경을 조성하는데 활용하고자 작성된 것입니다.

- 본 안내서는 과학기술정보통신부와 한국인터넷진흥원의 정책연구 사업의 연구 결과로서 내용의 무단 전재를 금합니다.

- 아울러, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국인터넷진흥원《인공지능(AI) 보안 안내서》'의 출처를 밝혀 주시기 바랍니다.

- 본 안내서는 AI 서비스 및 제품을 개발하는 과정이나 서비스 제공과정에서 참고 자료로 활용할 수 있도록 편찬 되었습니다. 본 안내서 활용 시에는 기업의 업무 환경과 상황, 모델이나 시스템 개발 목적, 서비스 내용 등을 고려하여 필요하신 내용을 취사 선택하여 활용하시기 바랍니다.

- AI와 관련된 기술은 계속적으로 발전하고 있고 AI 시스템의 취약점과 이에 따른 위협 공격도 다양해지고 있기 때문에, 「인공지능(AI) 보안 안내서」는 앞으로도 지속해서 최신 기술 동향과 정보보안 위협 동향, 침해사고 사례 등을 반영해서 보안요구사항과 검증항목 내용을 업데이트해 나갈 예정입니다.

목차

체크리스트 요약

1 AI 개발자를 위한 보안 체크리스트 ····· iv
2 AI 서비스 제공자를 위한 보안 체크리스트 ····· vii

제 1 장 개요

1. 「인공지능(AI) 보안 안내서」 개발 목적 ····· 2
2. AI 보안 위협 ····· 5
3. AI 보안 안내서의 필요성 및 적용범위 ····· 10

제 2 장 AI 개발자를 위한 보안 안내서

1. 개요 ····· 16
2. AI 개발자 대상 보안 프레임워크 ····· 18
3. AI 개발자를 위한 보안 요구사항 및 검증항목 ····· 24

제 3 장 AI 서비스 제공자를 위한 보안 안내서

1. 개요 ····· 70
2. AI 서비스 제공자 대상 보안 프레임워크 ····· 71
3. AI 서비스 제공자를 위한 보안 요구사항 및 검증항목 ····· 75

제 4 장 AI 이용자를 위한 보안 수칙

1. 개요 ····· 112
2. AI 이용자에게 발생할 수 있는 보안위협 사례 ····· 113
3. AI 서비스 이용자를 위한 보안 수칙 ····· 117

참고문헌 ····· 135

부록

1. 용어 정의 ····· 142
2. 참고자료 ····· 145
3. TTA "신뢰할 수 있는 인공지능 개발 안내서"와 비교 및 차별점 ····· 188
4. 국내 주요 AI 보안 가이드라인 비교 ····· 204
5. AI 개발자 대상 보안 프레임워크 ····· 206
6. AI 서비스 제공자 대상 보안 프레임워크 ····· 218

체크리스트 요약

1 AI 개발자를 위한 보안 체크리스트

생명주기	요구사항 및 체크리스트	Y	N	N/A
1 계획 및 설계	**(AI 개발자, AI 서비스 제공자 공통사항) 거버넌스 및 위험관리**			
	1.1 AI 보안(Security) 거버넌스 체계 구축 AI 개발자, AI 서비스 제공자 공통사항			
	1.1.1 AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?	☐	☐	☐
	1.1.2 AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?	☐	☐	☐
	1.1.3 AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?	☐	☐	☐
	1.2 AI 모델개발에 대한 위험관리 계획의 수립 AI 개발자, AI 서비스 제공자 공통사항			
	1.2.1 AI 모델 개발/서비스 제공 생명주기 및 공급망 과정에서 나타날 수 있는 위험요소를 분석·도출하고 있는가?	☐	☐	☐
	1.2.2 AI 시스템에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	☐	☐	☐
	1.2.3 AI 시스템에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	☐	☐	☐
2 데이터 수집 및 준비	2.1 데이터 수집 및 전처리			
	2.1.1 데이터 수집 시 사용되는 네트워크 프로토콜이 충분한 보안 기능을 제공하고 있는가?	☐	☐	☐
	2.1.2 수집된 데이터의 보관 및 삭제 절차가 명확하게 정의되어 있는가?	☐	☐	☐
	2.1.3 전처리 과정에서 중요 데이터를 보호하기 위해 암호화 기술을 사용하고 있는가?	☐	☐	☐
	2.2 데이터 무결성 검증 AI 개발자, AI 서비스 제공자 공통사항			
	2.2.1 데이터 저장 및 전송 시 데이터 무결성을 검증하고 있는가?	☐	☐	☐
	2.2.2 데이터 처리 과정에서 데이터 무결성을 검증하고 있는가?	☐	☐	☐
	2.2.3 데이터에 접근할 수 있는 권한을 제한하고 있는가?	☐	☐	☐
	2.3 데이터 공격에 대한 방어 AI 개발자, AI 서비스 제공자 공통사항			
	2.3.1 데이터 중독(poisoning) 공격에 대한 방어 대책을 마련하고 있는가?	☐	☐	☐
	2.3.2 데이터 회피(evasion) 공격에 대한 방어 대책을 마련하고 있는가?	☐	☐	☐
	2.3.3 데이터 유출·변조 공격을 방지하기 위한 방안을 마련하고 있는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트	Y	N	N/A
3 모델개발 (학습/ 모델링/ 검증)	3.1 학습/검증 환경에 대한 보안(Secure Training Environment)			
	3.1.1 모델 학습을 진행하는 환경이 안전하게 보안조치 되어 있는가?	☐	☐	☐
	3.1.2 학습 또는 검증 단계에서 악의적인 사용자가 허위 데이터를 삽입할 가능성을 차단하고 있는가?	☐	☐	☐
	3.1.3 연합 학습(Federated Learning)에 참여하는 장치 중 악의적인 장치가 있는지 검증하고 있는가?	☐	☐	☐
	3.2 모델 공격에 대한 방어			
	3.2.1 AI Prompt Injection 공격에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.2 적대적 예제 공격 (Adversarial Example Attacks)에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.3 모델 회피 공격(model evasion attack)에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.4 모델 오염 공격(Model Poisoning Attack)에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.5 모델 추출 공격(model extraction attack) 및 리버스 엔지니어링에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.6 반복적인 질의에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.2.7 기계 학습을 활용한 모델 공격에 대해 능동적으로 방어하고 있는가?	☐	☐	☐
	3.3 오픈소스 라이브러리 보안			
	3.3.1 오픈소스 라이브러리의 업데이트 및 취약점을 관리하고 있는가?	☐	☐	☐
	3.3.2 오픈소스 라이브러리의 소스 코드를 직접 검토하거나 사용에 대한 보안 문제를 검증하고 있는가?	☐	☐	☐
	3.3.3 오픈소스 라이브러리를 실행할 때 잠재적인 보안 위험을 제거하기 위해 격리된 환경을 이용하고 있는가?	☐	☐	☐
	3.4 LLM 보안			
	3.4.1 LLM 애플리케이션 공격에 대한 예방책을 마련하고 있는가?	☐	☐	☐
	3.4.2 LLM의 모델 서비스 거부(Model Denial of Service) 공격에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.4.3 LLM의 API 보안을 위한 방안을 수립하고 있는가?	☐	☐	☐
	3.4.4 LLM의 인터페이스 공격에 대한 예방책을 마련하고 있는가?	☐	☐	☐
	3.4.5 개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립하고 있는가?	☐	☐	☐
	3.4.6 LLM 출력결과를 정기적으로 모니터링하고 검토하고 있는가?	☐	☐	☐
	3.4.7 LLM의 Prompt Injection 공격에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐
	3.4.8 LLM의 벡터 및 임베딩 취약점에 대한 방어 방안을 수립하고 있는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Y	N	N/A
4 모델 배포	4.1	모델파일 및 배포 환경 보호			
	4.1.1	모델을 배포하기 전에 코드 및 모델을 스캔하고, 자동화된 취약점 분석을 하고 있는가?	☐	☐	☐
	4.1.2	모델파일을 암호화하여 저장하고 전송 중에도 안전하게 보호하고 있는가?	☐	☐	☐
	4.1.3	AI 모델이 배포되는 인프라(클라우드, 서버 등) 환경이 충분한 보안시스템을 갖추고 있는가?	☐	☐	☐
	4.2	API 및 인터페이스 보안			
	4.2.1	AI 모델이 배포된 후, API를 통해 외부 시스템과 상호작용하는 경우, 충분한 보안 조치 기능을 갖추고 있는가?	☐	☐	☐
	4.2.2	배포된 AI 모델이 실시간으로 데이터를 수신하고 이를 처리할 때, 중간자 공격 (Man-in-the-Middle Attack)에 대응하고 있는가?	☐	☐	☐
	4.2.3	AI 모델의 API에 대한 접근 권한을 제한하고, 강한 인증 메커니즘을 사용해 불법 접근을 방지하고 있는가?	☐	☐	☐
	4.2.4	API 사용자는 필요한 권한만 부여받도록 최소 권한 원칙(Least Privilege)을 적용하고 있는가?	☐	☐	☐
	4.2.5	AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 하고 있는가?	☐	☐	☐
5 모니터링 및 유지보수	5.1	실시간 모니터링 AI 개발자, AI 서비스 제공자 공통사항			
	5.1.1	모델의 입력 데이터, 출력 결과 등을 실시간으로 모니터링하여 비정상적인 동작을 탐지하고 있는가?	☐	☐	☐
	5.1.2	모델 응답 시간, 사용 패턴을 추적하고 분석하여 보안에 의심스러운 행동을 탐지하고 있는가?	☐	☐	☐
	5.1.3	AI 모델이 동작하는 서버 및 네트워크의 트래픽을 모니터링하여 비정상적인 요청을 탐지하고 있는가?	☐	☐	☐
	5.1.4	API 호출, 입력/출력 등 요청로그를 정기적으로 분석하여 보안에 의심스러운 동작을 탐지하고 있는가?	☐	☐	☐
	5.1.5	AI 모델과 배포 환경에 대해 모의 해킹을 수행하여 잠재적인 보안 취약점을 탐지하고 수정하고 있는가?	☐	☐	☐
	5.2	보안 패치 및 업데이트 관리 AI 개발자, AI 서비스 제공자 공통사항			
	5.2.1	모델에 대한 보안 패치 및 업데이트 관리 프로세스를 구축하고 있는가?	☐	☐	☐
	5.2.2	모델 배포 후 모델 및 라이브러리의 업데이트가 정기적으로 이루어지고 있는가?	☐	☐	☐
	5.2.3	운영 체제, 라이브러리, 프레임워크의 보안 패치를 운영 환경에 적용하기 전에 스테이징 환경에서 패치를 테스트하고 있는가?	☐	☐	☐
6 파기	6.1	파기 시 보안 AI 개발자, AI 서비스 제공자 공통사항			
	6.1.1	AI 모델이 더 이상 사용되지 않으면, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?	☐	☐	☐
	6.1.2	AI 모델에서 사용 중이던 데이터가 시스템을 폐기하거나 교체할 때 안전하게 삭제되고 있는가?	☐	☐	☐
	6.1.3	AI 모델이 더 이상 사용되지 않으면, 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?	☐	☐	☐

2 AI 서비스 제공자를 위한 보안 체크리스트

생명주기	요구사항 및 체크리스트	Y	N	N/A
1 서비스 기획 및 설계	**(AI 개발자, AI 서비스 제공자 공통사항) 거버넌스 및 위험관리**			
	1.1 AI 보안(Security) 거버넌스 체계 구축 [AI 개발자, AI 서비스 제공자 공통사항]			
	1.1.1 AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?	☐	☐	☐
	1.1.2 AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?	☐	☐	☐
	1.1.3 AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?	☐	☐	☐
	1.2 AI 서비스에 대한 위험관리 계획의 수립 [AI 개발자, AI 서비스 제공자 공통사항]			
	1.2.1 AI 모델 개발/서비스 제공 생명주기 및 공급망 과정에서 나타날 수 있는 위험요소를 분석·도출하고 있는가?	☐	☐	☐
	1.2.2 AI 서비스에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	☐	☐	☐
	1.2.3 AI 서비스에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	☐	☐	☐
	(AI 서비스 제공자) 계약관리			
	1.3 서비스 수준 계약(SLA) 관리			
	1.3.1 공급업체와 계약시, SLA에 보안요구 사항을 명확히 포함했는가?	☐	☐	☐
	1.3.2 보안 침해 발생 시를 대비하여, 대응 계획을 수립하고 있는가?	☐	☐	☐
	1.3.3 보안 침해 발생 시를 대비하여, 책임 소재를 명확히 하고 있는가?	☐	☐	☐
2 서비스 개발 및 구축	2.1 코드 취약점 점검 등 관리			
	2.1.1 정적 및 동적 코드 분석 도구를 사용하여 소스 코드의 보안 취약점을 분석하고 있는가?	☐	☐	☐
	2.1.2 코드 리뷰 프로세스를 도입하여 보안 문제가 있는 부분을 검토하고 개선하고 있는가?	☐	☐	☐
	2.2 모델 환경의 보안			
	2.2.1 모델 환경에 대한 접근 제어를 강화하고, 모델에 대한 접근 권한을 최소화하여 무단 접근을 방지하고 있는가?	☐	☐	☐
	2.2.2 모델이 악의적으로 수정되지 않도록 모델의 무결성을 보장하는 방법을 적용하고 있는가?	☐	☐	☐
	2.2.3 보안 모니터링 도구를 사용하여 모델의 비정상적인 활동을 감지하고, 실시간으로 대응할 수 있는 체계를 구축하고 있는가?	☐	☐	☐
	2.3 데이터 보안 [AI 개발자, AI 서비스 제공자 공통사항]			
	2.3.1 적대적 공격 등 데이터 공격에 대한 방어 수단을 강구하고 있는가?	☐	☐	☐
	2.3.2 데이터 저장 및 전송 시 무결성을 보호하기 위한 조치를 하고 있는가?	☐	☐	☐
	2.3.3 중요 데이터에 대한 기밀성 유지를 위해 보호 방안을 마련하고 있는가?	☐	☐	☐
	2.3.4 전송구간에서 중요정보 유출을 방지하기 위한 보호 방안을 마련하고 있는가?	☐	☐	☐
	2.3.5 데이터 유출시 책임추적을 할 수 있도록 조치를 하고 있는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Y	N	N/A
	2.4	API 및 인터페이스 보안			
	2.4.1	API 통신을 암호화하여 데이터가 전송되는 구간에서 외부 공격에 대한 방어를 하고 있는가?	☐	☐	☐
	2.4.2	모든 API 요청에 대해 인증 및 권한 관리를 강화하고, 중요 데이터에 접근할 때는 강한 인증 메커니즘을 적용하고 있는가?	☐	☐	☐
	2.4.3	API 트래픽은 암호화 기술을 사용하여 보호하고, 데이터를 안전하게 주고 받도록 보장하고 있는가?	☐	☐	☐
	2.4.4	API 호출 제한(Rate Limiting)을 설정하여 과도한 요청을 방지하고, 비정상적인 요청 패턴을 탐지하여 차단하고 있는가?	☐	☐	☐
3 서비스 제공 및 운영	3.1	로그 및 운영 데이터 보안			
	3.1.1	데이터 처리 중 접속로그 관리를 강화하고 있는가?	☐	☐	☐
	3.1.2	로그 파일 및 운영 데이터에 암호화를 적용하고, 중요정보는 별도로 관리하여 유출을 방지하고 있는가?	☐	☐	☐
	3.1.3	운영 중 발생하는 데이터를 안전하게 저장하고, 접근 제어를 통해 인증된 관리자만이 로그에 접근할 수 있도록 설정하고 있는가?	☐	☐	☐
	3.1.4	로그 데이터 접근 권한을 최소화하고, 접근 제어 및 사용자 활동 기록을 통해 비정상적인 접근을 탐지하고 있는가?	☐	☐	☐
4 서비스 유지보수 및 지원	4.1	모니터링, 업데이트 및 패치 AI 개발자, AI 서비스 제공자 공통사항			
	4.1.1	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집·관리하고 있는가?	☐	☐	☐
	4.1.2	AI 시스템이 정상적으로 작동하지 않거나 예기치 못한 오류가 발생할 경우 이를 조기에 탐지하고 대응하는 메커니즘이 있는가?	☐	☐	☐
	4.1.3	AI 시스템의 보안 패치나 모델 업데이트가 정기적으로 이루어지고 있는가?	☐	☐	☐
	4.2	성능 및 장애 관리			
	4.2.1	서비스 장애가 발생했을 때 자동으로 복구할 수 있도록 하고 있는가?	☐	☐	☐
	4.2.2	모델 성능을 지속적으로 모니터링하고, 성능 저하가 감지되면 재학습을 통해 성능을 유지하고 있는가?	☐	☐	☐
	4.2.3	실시간으로 모델 드리프트 탐지 시스템을 운영하여 모델 성능이 저하될 경우 즉시 대응할 수 있는 체계를 마련하고 있는가?	☐	☐	☐
	4.2.4	정기적으로 모델 재학습 및 업데이트를 수행하여 새로운 데이터 패턴을 반영하고, 성능을 개선하고 있는가?	☐	☐	☐
	4.2.5	AI 서비스에 대해 다중화(HA) 및 백업 시스템을 구축하여 장애 발생 시에도 서비스가 연속적으로 제공될 수 있도록 하고 있는가?	☐	☐	☐
	4.2.6	침입차단시스템 등을 통해 외부에서 발생하는 DoS/DDoS 공격을 방어하고, 실시간 모니터링 시스템을 운영하여 장애를 빠르게 감지하고 대응하고 있는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Y	N	N/A
5 Feedback (환류) 및 개선	5.1	사용자 피드백 관리			
	5.1.1	사용자 피드백 시스템에 입력 검증 및 필터링을 적용하여 악성 코드나 비정상적인 데이터 입력을 차단하고 있는가?	☐	☐	☐
	5.1.2	피드백을 자동으로 분석하기 전에 사전 검증 절차를 마련하여 피드백 데이터의 무결성을 확인하고 있는가?	☐	☐	☐
	5.1.3	최소 권한 원칙(Least Privilege)을 적용하여 피드백 및 개선 과정에서 접근할 수 있는 권한을 최소화하고 있는가?	☐	☐	☐
	5.1.4	피드백 처리 및 개선 과정에서 이루어진 모든 접근 및 변경 사항을 감사 로그로 기록하고, 정기적으로 검토하여 무단 접근을 탐지하고 있는가?	☐	☐	☐
6 파기	6.1	파기 시 보안 AI 개발자, AI 서비스 제공자 공통사항			
	6.1.1	모델 파기 시, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?	☐	☐	☐
	6.1.2	시스템을 폐기하거나 교체할 때 AI 모델에서 사용 중이던 관련 파일 및 데이터를 안전하게 삭제하고 있는가?	☐	☐	☐
	6.1.3	AI 모델이 더 이상 사용되지 않으면 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?	☐	☐	☐

1

개요

01 「인공지능(AI) 보안 안내서」 개발 목적

본 안내서에서는 AI 보안(Security)을 “외부의 사이버 공격·침해행위 등으로 인한 AI 시스템 손상, 탈취 등 무결성, 기밀성, 가용성을 저해하는 행위로부터 AI 시스템을 보호하는 것”으로 정의하였다. 또한 NIST AI RMF 1.0, OWASP Top 10 for LLM Applications 등 국내외 AI 보안 정책, 연구 자료들을 분석하여 AI 보안 위협을 구분하고, AI 보안 위협 예방·대응을 위한 보안 요구사항들을 마련하는데 참고하였다.

< AI 보안 위협 분류 및 주요 내용 >

구분	주요 내용
AI 모델 취약점 공격	▸ AI 모델과 시스템의 취약점을 공격하여 서비스 중단, 결과물 조작, 데이터 탈취 등을 일으키는 행위
	① 프롬프트 공격 ② 회피 공격, ③ 오염 공격, ④ 데이터 추출 공격, ⑤ 모델 추출 공격 등
AI 서비스 공격	▸ AI 기반 서비스를 대상으로 서비스를 마비시키거나 서비스社가 보유하고 있는 데이터 탈취 등을 하기 위한 일련의 행위
	⑥ DoS/DDoS, ⑦ 피싱 메일 등을 통한 악성코드 배포, ⑧ AI 서비스 대상 해킹 등
AI 악용 보안 위협	▸ AI 서비스의 뛰어난 성능을 해킹, 침해사고 등에 악용하는 행위
	⑨ 악성코드 생성, ⑩ 피싱/스미싱 메일 등 생성, ⑪ 범죄 지식 학습
데이터 유출	▸ 이용자가 실수로 민감정보를 AI 서비스에 입력하거나, AI 서비스의 오류로 민감정보를 출력하는 행위
	⑫ 민감정보(기업 기밀정보 등) 입력, ⑬ 개인정보 등 출력

AI와 관련된 보안은 AI 시스템 자체의 보안(Security for AI)과 AI 기술을 활용하여 다른 시스템의 보안을 강화하는 역할(Security by AI)로 나눌 수 있고, 두 개념은 AI와 보안의 상호작용에 대한 서로 다른 관점을 다룬다.

- “Security for AI”(AI 시스템에 대한 보안)는 AI 시스템 자체를 보호하는 것을 의미한다. AI 시스템이 외부 위협이나 악의적인 공격으로부터 안전하게 운영되도록 체계적인 보안 조치를 취하는 것이 핵심이다. AI 시스템의 모든 생애 주기 단계에서 보안성을 보장하는 것을 목표로 하며, 데이터 보호, 모델 보호, 운영 및 인프라 보안 등이 포함된다.
- “Security for AI“의 주요 내용은 다음과 같다.
  - **(데이터 보안)** AI 시스템에서 사용되는 데이터가 유출되거나 변조되지 않도록 보호하는 것으로, 특히, 훈련 데이터와 실시간 데이터의 기밀성, 무결성, 가용성을 보장하는 것이 중요하다.

- **(모델 보안)** AI 모델 자체가 도난, 역설계(Reverse Engineering), 적대적 공격(adversarial attacks)에 노출되지 않도록 보호하는 것으로, 이를 통해 AI 모델이 안전하게 운영되도록 신뢰성을 유지해야 한다.
- **(운영 보안)** AI 시스템이 배포된 후에도 외부 공격이나 성능 저하에 대응할 수 있는 운영 보안 조치를 마련하는 것으로, API 보안, 모델 드리프트(model drift) 탐지, 패치 및 업데이트 적용 등이 포함된다.

"Security by AI"(AI에 의한 보안)은 AI 기술을 사이버 보안에 적용하여 보안 위협을 자동으로 탐지하고 대응하는 것을 의미한다. 즉, AI가 보안 강화 도구로 활용되는 것이며, 이를 통해 사이버 보안의 효율성을 크게 향상시킬 수 있다.

- "Security by AI"의 주요 내용은 다음과 같다.
  - **(위협 탐지 및 분석)** AI는 대규모 데이터 분석을 통해 네트워크 상의 이상 징후를 감지하거나 보안 위협을 탐지할 수 있다. AI 모델은 정상 패턴과 비정상 패턴을 학습하고, 이를 기반으로 실시간으로 사이버 공격을 식별하는 데 효과적이다.
  - **(자동화된 보안 대응)** AI는 발견된 보안 위협에 대해 자동으로 대응하는 메커니즘을 제공한다. 예를 들어, AI 시스템이 특정 공격을 탐지하면 해당 IP를 차단하거나 보안 설정을 자동으로 변경할 수 있다.
  - **(사이버 공격 예측)** AI는 과거의 보안사고 데이터를 분석하여 향후 발생할 수 있는 공격을 예측할 수 있다. 이로 인해 사전에 보안 조치를 취해 공격을 방어할 수 있다.
  - **(사용자 인증 및 접근 제어)** AI 기반 시스템은 이상 행동 감지를 통해 사용자 인증 절차를 강화하고, 비정상적인 접근 시도를 차단할 수 있다.

표 1-1 "Security for AI"와 "Security by AI" 비교

구분	Security for AI	Security by AI
주요 목적	AI 시스템 자체를 보호	AI 기술을 활용해 보안 시스템을 강화
적용 대상	AI 모델, 데이터, 인프라 등 AI 시스템 자체	침입 탐지, 위협 예측, 자동 대응 등의 보안 시스템
주요 위협	데이터 포이즈닝, 모델 탈취, 적대적 예제 공격	사이버 공격, 네트워크 침입, 악성 소프트웨어
활용 기술	데이터 암호화, 모델 보호, API 접근 제어	AI 기반 위협 탐지, 이상행동 탐지, 자동화된 보안 대응
대응 대상	AI 시스템에 대한 외부 위협	일반적인 네트워크, 시스템, 사용자에 대한 보안 위협
예시	AI 모델 기밀성 보호, 데이터 무결성 유지	AI 기반 침입탐지시스템 및 피싱 탐지 및 차단

- 그 동안 많은 기관에서 인공지능 신뢰성 확보를 위한 원칙과 지침, 안내서를 발간하였으나, 주로 신뢰성 관점에서 다루거나 보안 관점에서 상세한 가이드를 제시하고 있지 않아 아쉬운 점이 많았다.
  - 이에 따라 본 안내서는 AI 제품 및 서비스 개발 현장에서 모델 개발자, 서비스 제공자, 서비스 이용자 등 이해관계자들이 정보보안 측면에서 참고할 수 있는 기준을 마련하고자 추진하였다.
  - 본 안내서에서는 AI 시스템의 전반적인 안전성과 무결성을 보장하기 위한 다양한 보안 방책을 의미하는 “Security for AI”에 초점을 맞추고자 한다. “Security for AI”는 AI 시스템의 모든 생애주기(데이터 수집, 모델 개발, 학습, 배포, 유지보수, 폐기)에서 발생할 수 있는 다양한 위협에 대해 AI 시스템을 보호하는 것을 의미한다. 이는 데이터 보안, 모델 보안, 배포 및 운영 보안 등의 영역을 포함하며, 기밀성(Confidentiality), 무결성(Integrity), 가용성(Availability) 등을 적용하여 AI 시스템을 대내외적인 위협으로부터 보호하는데 목적이 있다.

- 본 안내서는 AI 시스템의 설계, 개발, 운영, 배포, 유지보수, 폐기 등 전 단계에서 발생할 수 있는 다양한 보안 위험을 관리하기 위한 방안을 제시하였다. 또한 AI 시스템의 생애 주기에 걸쳐 기술적 고려가 필요한 요구사항 및 검증항목을 기밀성, 무결성 및 가용성을 기준으로 하여 세부적으로 다루었다. 따라서 AI 개발자 및 서비스 제공자가 AI 시스템의 보안(Security) 문제를 해결하는데 「인공지능(AI) 보안 안내서」가 실질적인 도움이 될 것으로 기대된다.

- 다만, AI 신뢰성 보장을 위해서는 보안(Security) 측면 외에도 윤리, 편향성, 개인정보보호와 같은 법·제도적 측면도 함께 고려되어야 한다. 「인공지능(AI) 보안 안내서」는 이러한 범용성을 갖추는 것보다는 보안 영역에 초점을 맞추었기 때문에 AI의 윤리적 측면, 신뢰성 측면, 개인정보보호 측면에 대해서는 타 자료를 병행하여 참고할 필요가 있다.
  - 따라서 AI 개발자 및 서비스 제공자는 기업 내부의 기술 역량, 제품 서비스 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 「인공지능(AI) 보안 안내서」가 다루지 않는 부분에 대해서는 기업에서 제공 중인 서비스 분야 및 환경에 맞게 국내 AI 관련 안내서를 참고 자료로 활용하길 바란다.

- 본 안내서는 예측형 AI, 생성형 AI, LLM 등 다양한 산업, 경제, 사회 분야에서 활용되는 AI를 대상으로 공통적으로 발생할 수 있는 보안 위협과 침해사고를 선제적으로 예방·대응하기 위한 보안 요구사항과 검증항목을 제시하고 있다. 따라서 본 안내서는 AI 개발 및 활용 환경에서 발생할 수 있는 보안 위협에 대응하는데 초점을 맞추어 개발되었다. 향후 의료, 교통, 에너지 등 각 산업 분야 內 AI 모델, 시스템의 특징과 AI 활용 현황 등에 따라 개별 산업 별 AI 보안 위협, AI 보안 요구사항 등을 연구하여, 산업 별 보안 안내서 개발도 추진할 예정이다.

02 AI 보안 위협

01 Prompt Injection

공격자는 정교하게 만든 입력을 통해 LLM을 조작하여 공격자의 의도를 실행하게 할 수 있다. 이는 시스템 프롬프트를 적대적으로 유도하거나 조작된 외부 입력을 통해 간접적으로 수행할 수 있으며, 잠재적으로 데이터 유출 등의 문제로 이어질 수 있다.

- Direct Prompt Injections (일명 jailbreaking): 악의적인 사용자가 민감한 정보를 추출하기 위해 프롬프트를 삽입
- Indirect Prompt Injections: 사용자가 웹페이지 프롬프트를 통해 민감한 데이터를 요청
- 플러그인을 통한 사기: 웹사이트가 플러그인을 악용하여 사기 등 범죄목적을 위한 기망행위를 함

참고 사례

- 인공지능(AI) 음성비서 서비스인 아마존 알렉사와 구글 어시스턴트 등도 해킹에 취약하였다. 스마트 스피커에 레이저를 쏴 이들에게 명령을 내릴 수 있음이 알려졌기 때문이다.[1] 미시간대와 일본 전기통신대(UEC) 연구진에 따르면 “구글, 차고 문을 열어줘” 같은 명령어가 암호화되어 입력된 빛을 스마트 스피커의 마이크에 비춰 음성 명령을 암호화해 빛에 실어 보내면, 이 빛이 스마트 스피커에 내장된 진동판에 부딪쳐 마치 사람이 음성 명령을 말했을 때와 똑같이 이 진동판이 떨리면서 스마트 스피커에 명령을 내린다는 것이다. 연구진은 실험에서 이런 취약점을 이용해 스마트 차고 문을 열거나 현재 시간을 묻는 등의 작업을 수행하였다. 연구진은 이와 같이 빛을 이용한 해킹에 취약한 기기들은 구글 홈, 구글 네스트 캠 IQ, 아마존 에코·에코 닷·에코 쇼, 페이스북의 포털 미니, 아이폰 XR, 6세대 아이패드 등이었다고 밝혔다.
- 또한, 이러한 인공지능(AI) 음성 비서는 제3자가 대화 내용을 엿듣거나 해킹 프로그램을 설치해 민감한 정보를 넘겨주도록 유도할 수도 있다.[2] 아마존과 구글이 애플리케이션 업그레이드를 위해 앱 개발자들에게 제공하는 접근법을 해커들이 악용하여 자신이 원하는 대로 음성비서의 응답을 유도하는 명령을 내릴 수 있음이 알려졌다. 예를 들어, 사용자를 가장한 해커가 음성 명령을 통해 앱을 열고 그 앱이 실행되지 않는다고 말한다. AI 음성 비서는 대답하지 않지만, 사용자에게 알리지 않은 채 계속 백그라운드를 실행하고, 몇 분 후 AI 음성 비서는 회사 업데이트가 있었다면서 사용자에게 비밀번호를 말해달라고 요청한다.

1 연합뉴스, https://www.yna.co.kr/view/AKR20191106004500091, 2019.11.06

2 머니투데이, https://news.mt.co.kr/mtview.php?no=2019102214138230917, 2019.10.22

02 민감 정보 노출(Sensitive Information Disclosure)

LLM 애플리케이션은 실수로 민감한 정보 또는 기밀 데이터를 공개하여 무단 액세스, 지적 재산권 도용 및 개인정보 침해로 이어질 수 있다.

- 의도치 않은 노출: 잘못된 해석이나 스크러빙 부족으로 인한 데이터 유출
- Complete Filtering (완료 필터링) 오류: 모델이 생성한 텍스트(완료 output)를 필터링하는 과정에서 위험하거나 부적절한 응답이 필터링을 우회해 노출될 수도 있음
- Overfitting (과잉적합) 오류: LLM이 훈련 데이터에 지나치게 적합하여, 훈련 시에는 잘 작동하지만 새로운 상황에서는 비효율적이거나 학습 데이터 내 포함된 민감정보를 기억해 출력할 수도 있음.

공격 시나리오

- **(의도치 않은 노출)** 사용자 A가 다른 사용자 데이터에 노출됨

의심하지 않는 합법적인 사용자 A는 LLM 애플리케이션과 악의적이지 않은 방식으로 상호 작용할 때 LLM을 통해 특정 다른 사용자 데이터에 노출된다.

참고 사례

- 오픈AI의 챗GPT가 사용자의 로그인 정보와 개인정보를 유출하는 일이 발생하였다.[3] 정보 유출 피해를 입었다고 주장하는 제보자가 제출한 스크린샷에는 약국 처방약 포털의 직원이 사용하는 지원 시스템에 연결된 제보자 이름과 비밀번호가 포함돼 있었다. 이 외에도 앱의 이름과 문제가 발생한 스토어 번호 등이 포함돼 있는 것이 나타났다. 2023년 3월에는 챗GPT의 버그로 인해 사이트에서 한 활성 사용자의 채팅 기록이 관련 없는 사용자에게 표시되는 현상이 발생한 바 있다. 이러한 유형의 시스템 오류는 종종 발생할 수 있으며, 정확한 원인은 인시던트마다 다르지만, 프론트엔드 디바이스와 백엔드 디바이스 사이에 있는 미들박스 디바이스와 관련된 경우가 많은 것으로 알려져 있다. 미들박스는 성능 향상을 위해 최근에 로그인한 사용자의 자격 증명을 비롯한 특정 데이터를 캐시하는 기능을 말한다. 이때 불일치가 발생하면 한 계정의 정보를 다른 계정에 매핑하는 등의 문제가 발생할 수 있다.
- 2023년 구글 딥마인드와 대학 공동 연구진은 챗GPT에 단순한 프롬프트 공격으로 개인정보를 비롯한 1만 개 훈련 데이터를 추출할 수 있었다는 논문을 airXiv에 게재하였다.[4]

3 디지털투데이, https://www.digitaltoday.co.kr/news/articleView.html?idxno=504025, 2024.1.30

4 Nasr 외(2023.11), Scalable Extraction of Training Data from (Production) Language Models, arXiv:2311.17035v1

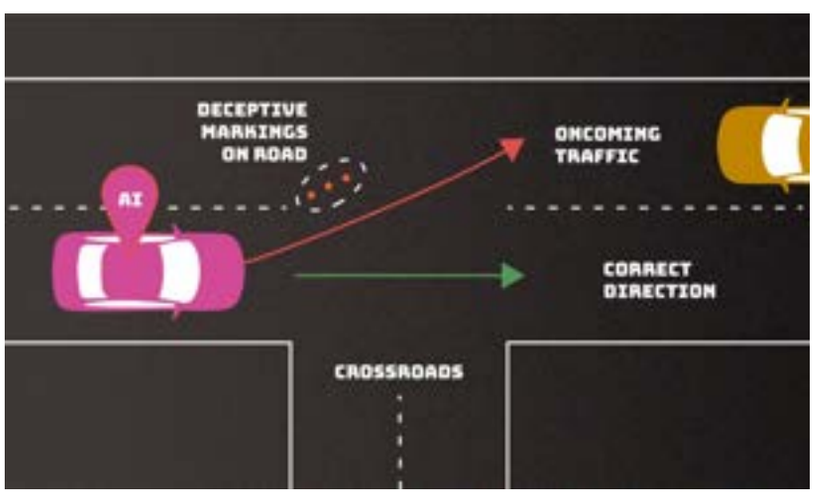
03 적대적 예제(Adversarial Example) 공격

3.1 회피 공격(Evasion attack)

- 회피 공격(Evasion Attack)은 입력데이터에 노이즈를 추가하여 모델을 속이는 공격이다. 이미지 데이터에 노이즈를 추가하게 되면 인간의 눈으로는 큰 차이가 없지만, 모델은 다른 데이터로 인식하여 행동하게 된다.

- 참고 사례
  - 정지표지판에 표시를 추가하여 자율주행 차량이 이를 속도제한표지판으로 잘못 인식하게 하거나, 차선 표시를 혼동하게 만들어 차량이 도로를 벗어나도록 하는 등의 공격이 이에 해당한다. 아래의 예에서는 도로의 잘못된 표시가 무인 자동차를 잘못 인도하여 마주 오는 차량으로 방향을 틀게 할 수 있는 것으로 나타났다.[5]

그림 1-1 회피공격의 예



출처=N. Hanacek/NIST

5 AEM, https://www.autoelectronics.co.kr/article/articleView.asp?idx=5496, 2024.1.15.

- 2017년 워싱턴대학 연구팀은 자율주행차를 오작동시키는 시연을 한 바 있는데, 그때 쓰인 방법이 회피 공격의 사례이다. 연구팀은 STOP 표지판에 스티커를 붙여 인공지능을 교란함으로써, 자율주행차가 '정지' 표시판을 '속도제한' 표시판으로 오인하도록 만들었다. 이는 사이버 공간이 아닌 물리적 공간에서도 아주 간단한 방법으로 적대적 공격이 가능함을 보여준다.
- 2018년 구글 리서치 그룹은 논문을 통해 이미지 인식 머신러닝 알고리즘을 오작동 시킬 수 있는 스티커를 발표했다. 적대적 스티커(Adversarial patch)라고 불리는 이 스티커를 바나나 옆에 붙이면 이미지 인식 앱이 바나나를 100% 확률로 토스터 기기로 인식했다. 이 스티커는 누구나 쉽게 인쇄해 사용할 수 있고 악의적인 공격인지 쉽게 발견하기 어려워서, 악용되는 경우 큰 위험을 가져올 수도 있다.

3.2 오염 공격(Poisoning attack)

오염 공격(Poisoning attack)은 학습데이터에 오염된 데이터를 추가하여 모델을 망가뜨리는 공격 방법이다. 공격자가 AI 모델의 학습 단계에서 의도적으로 악의적인 데이터를 주입하여 발생시키는 것으로, 예를 들어, 챗봇이 부적절한 발언을 하도록 악의적인 행위자에 의해 학습되어 욕설, 인종차별 발언을 남발하도록 하는 것이다.

참고 사례

- 2024년 중국에서 제조된 로봇청소기 '에코백스 디봇 X2s(Ecovacs Deebot X2s)'가 해킹을 당해 미국 가정집에서 욕설과 인종차별적 발언을 하는 일이 발생하였다.[6] 로봇청소기가 마이크 기능을 통해 인종차별적인 욕설을 하거나 개를 쫓아다닌 일도 있었다. 조사 결과 해커가 제조사의 보안 조치를 우회해 카메라, 마이크, 이동 제어 기능을 탈취한 것으로 파악되었다. 제조사 측이 보안 조사를 실시한 결과, 이용자의 계정과 비밀번호가 도용되면서 이런 일이 발생했고, 제조사 기술팀이 범인의 IP 주소를 파악해 계정으로의 추가 접근을 막았다고 밝혔다. 특히 보안에 취약했던 부분은 4자리의 PIN코드였던 것으로 확인되었다. 보안 전문가들 사이에서 이 제조사의 보안 취약성에 대해서는 이미 지적이 나온 바 있었다. 2024년 8월 미국 라스베이거스에서 열린 '데프콘 해킹 콘퍼런스'에서 보안 연구원들이 에코백스 제품을 분석한 결과 블루투스로 로봇을 해킹하거나 원격으로 마이크와 카메라를 몰래 켜는 데 악용할 수 있다고 하였다. 특히 약 130m 떨어진 곳에서 블루투스를 활용해 로봇을 해킹하고 원격으로 기기를 제어할 수 있는 것으로 밝혀졌다.
- 마이크로소프트는 2016년 인공지능 챗봇 'Tay'를 출시했다. 그러나 일부 사용자들이 악의적인 메시지와 인종차별적 발언을 반복적으로 입력하여 학습 데이터를 오염시켰고, 그 결과 Tay는 부적절한 발언을 생성하게 되었다. 이 사건으로 인해 마이크로소프트는 출시 16시간 만에 Tay의 운영을 중단해야 했으며, AI 시스템의 학습 데이터에 대한 보안과 검증의 중요성이 부각되었다.

6 조선일보, https://www.chosun.com/international/international_general/2024/10/21/CCCTWF5ERRCQRDOF6C7MUEQSUU/, 2024.10.21

3.3 탐색적 공격

AI 데이터 추출 공격

- AI 모델의 학습에 사용했던 데이터 자체를 탈취하는 공격 기법이며, 이때 데이터 추출을 위해 활용한 공격을 '전도 공격(Inversion Attack)'이라고도 한다. 인공지능이 훈련한 원래 데이터를 찾아내면 악의적인 의도에 맞게 학습 결과를 유도할 수 있게 된다.
- 데이터 추출 공격은 인공지능에 하는 질의 횟수를 조정하는 방식으로 대응이 가능하다. 예컨대, 하루 동안 한 명이 질의할 수 있는 횟수를 작게 제한함으로써 데이터가 유출되더라도 피해를 최소화할 수 있다

AI 모델 추출(Model Extraction) 공격

- 머신러닝 모델에 수많은 쿼리를 던진 후, 산출된 결과값을 분석해 모델 학습을 위해 사용된 데이터를 추출하는 공격을 말한다. 이 공격 방법은 얼굴인식 머신러닝 모델의 학습을 위해 사용한 얼굴 이미지 데이터를 복원할 수도 있다. 70초 동안 650번 쿼리만으로도 아마존 머신러닝 모델과 유사한 모델을 만들어내는 것이 가능하다는 연구 결과도 있다.
- 이 공격 방법은 유료 러닝모델 서비스(MLaas : Machine Learning as a Service)를 탈취하거나 Inversion attack, Evasion attack과 같은 2차 공격에 활용하기 위해 사용될 수 있다. 머신러닝 모델을 훈련시키는 학습 데이터 안에 개인정보, 민감정보 등이 포함되어 있는 경우에는 이러한 공격에 의해 유출될 가능성이 있다.
- 공격 대상이 되는 AI 모델에 대한 공격자의 사전 지식(AI 모델의 구조, 학습 데이터 등)이 양에 따라 블랙박스 AI 모델 탈취 공격과 그레이박스 AI 모델 탈취 공격으로 분류한다.
  - 그레이박스 AI 모델 탈취 공격: 공격자가 공격 대상이 되는 AI 모델에 대하여 부분적인 사전 지식을 갖고 있을 때 수행되는 공격이다.
  - 블랙박스 AI 모델 탈취 공격: 공격자가 공격 대상이 되는 AI 모델에 대하여 어떠한 사전 지식도 갖지 않을 때 수행되는 공격이다.

03 AI 보안 안내서의 필요성 및 적용범위

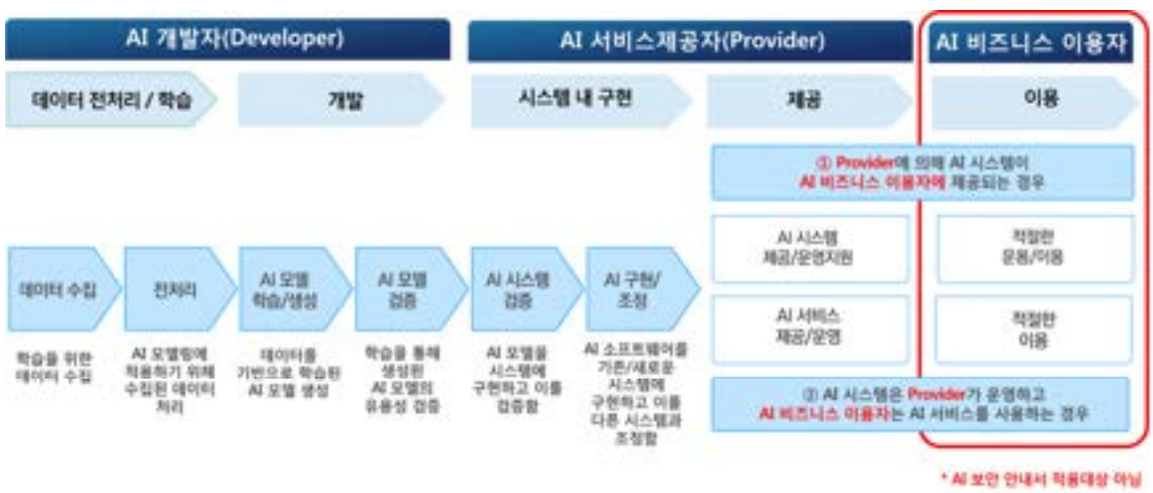
- 본 안내서는 미국, 유럽, 일본 등 해외기관에서 발표한 원칙 및 프레임워크 등을 참조하였고, 국내자료로는 과학기술정보통신부/한국정보통신기술협회(TTA)에서 발간한 「신뢰할 수 있는 인공지능 개발 안내서: 일반분야」, 국가정보원/국가보안기술연구소에서 발간한 「챗GPT 등 생성형 AI 활용 보안 가이드라인」, 금융보안원에서 발간한 「금융분야 AI 보안 가이드라인」 등도 참고하였다. 본 「인공지능(AI) 보안 안내서」의 특징은 다음과 같다.

- 첫째, 신뢰성, 프라이버시 등에 관한 항목을 모두 제외하고 보안의 목표인 기밀성・무결성・가용성 보장을 위한 요구사항에 초점을 맞추었다.

- 둘째 **개발자 뿐만 아니라 서비스 제공자** 및 **이용자**를 대상으로 적용대상을 확대하였다.
  - 「인공지능(AI) 보안 안내서」는 AI 모델을 개발하는 개발자와 개발조직, 모델을 활용하여 서비스를 제공하는 사업자, 해당 서비스를 이용하여 결과물을 생성한 이용자 모두를 대상으로 하여 각 주체별로 AI 보안에 대한 안내서를 마련하였다. 이를 위해 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(인공지능 기본법, '26. 1. 22 시행예정), EU AI Act 등을 참고하여 개발자, 서비스 제공자, 이용자의 개념을 아래와 같이 정의하였다.
    - "**개발자**"라고 하면 주로 소프트웨어 개발자(Developer 혹은 Engineer) 또는 개발 조직(기업)을 지칭하며, 이들은 시스템 분석가의 요구에 맞게 컴퓨터 프로그래밍을 하거나 시스템 설계를 하는 사람 또는 조직(기업)을 말한다. 안내서에서 개발자는 요구사항 및 검증항목에 따라 조직의 구성원 개인일 수도 있고 팀(조직) 또는 회사가 될 수도 있다. 따라서 개발자가 실제 이 안내서를 참고할 때 해당 내용이 개발자 개인에 관한 것인지 아니면 조직 또는 회사가 주도적으로 해야 할 것인지 여부에 대한 혼란이 있을 수 있다. 그래서 「인공지능(AI) 보안 안내서」에서는 요구사항 별로 수행주체를 명시적으로 표시하였다.
    - "**서비스제공자**"는 업으로서 AI 서비스 또는 AI 부수 서비스를 타인에게 제공하는 자를 말한다. "업으로" 한다는 것은 같은 행위를 계속하여 반복하는 것을 의미하고, 여기에 해당하는지 여부는 단순히 그에 필요한 인적 또는 물적 시설의 구비 여부와는 관계없이 행위의 반복·계속성 여부, 영업성의 유무, 그 행위의 목적이나 규모·횟수·기간·태양 등의 여러 사정을 종합적으로 고려하여 사회통념에 따라 판단하여야 한다. 이 안내서에서 서비스제공자는 영리를 목적으로 AI 서비스를 제공하는 법인(회사 등)을 말한다. 그러나 실제 업무 수행 시에는 임직원 개인이 해야 할 것인지 아니면 조직 또는 회사가 해야 할 것인지 불명확한 경우가 있을 수 있다. 따라서 「인공지능(AI) 보안 안내서」에서는 이를 명확히 하고자 요구사항 별로 수행주체를 표기하였다.

- "**이용자**"는 AI 서비스 또는 AI 부수 서비스를 타인에게 제공하지 않고 AI 서비스 또는 AI 부수 서비스를 이용하는 사람을 말한다. 이러한 "이용자" 개념에는 업으로서 AI 시스템 또는 AI 서비스를 이용하는 사람(이하 "AI 비즈니스 이용자"라고 함)이 포함될 수도 있으나, 본 「AI 이용자를 위한 보안 수칙」의 적용 대상은 AI 서비스를 이용하는 일반 국민을 대상으로 작성하였다.

그림 1-2 AI 서비스 관련 사업 활동의 주체



셋째, 보안 측면에서 중요한 **"파기"** 단계를 추가하여 AI 서비스의 생명주기를 차별화했다.

- 인공지능 서비스 생명주기를 **"파기"** 단계를 추가하여 총 6단계로 구분하였으며, 생명주기 단계별 세부 내용도 차별화하였다. 「인공지능(AI) 보안 안내서」에서 정의한 각 단계별 목표와 주요 활동은 다음과 같다.

그림 1-3 인공지능 서비스의 생명주기

1. 계획 및 설계
2. 데이터 수집 및 준비
3. 모델 개발(학습/모델링/검증)
4. 모델 배포
5. 모니터링 및 유지보수
6. 파기

표 1-2 인공지능 생명주기별 주요활동(AI 보안 안내서)

생명주기	목표	주요 활동
1. 계획 및 설계	AI 시스템이 해결할 목표 및 성공 지표를 정의	• AI가 해결할 수 있는 비즈니스 및 기술적 목표를 정의 • AI 시스템 관리 감독 조직 및 방안 마련 • AI시스템 위험요소 분석 및 대응 방안 마련
2. 데이터 수집 및 준비	AI 모델을 학습하고 개발하는 데 사용할 데이터를 수집하고, 사전 처리 및 모델 개발에 적합한 형식으로 변환	• 데이터 소스(구조화된 데이터 및 구조화되지 않은 데이터, 센서 데이터, 과거 데이터 세트 등)를 정의함 • 데이터 사용과 관련된 보안 정책 및 법적 제약을 고려함 • 누락된 데이터를 처리하고 중복을 제거하고 데이터 일관성 보장 • 데이터 세트를 학습, 검증 및 테스트 세트로 분할함
3. 모델 개발	AI 모델을 구축하고, 학습하여 성능 평가를 통해 필요한 지표를 충족	• 적절한 기술(예: 머신 러닝, 딥 러닝, 자연어 처리), 알고리즘(예: 의사 결정 트리, 신경망, SVM 등)과 모델 아키텍처 선택 • 준비된 데이터를 사용하여 모델 학습, 하이퍼파라미터 조정 • 모델을 검증 또는 보이지 않는 테스트 데이터 세트에서 테스트하여 정확도와 견고성을 확인함
4. 모델 배포	학습된 AI 모델을 실제 애플리케이션에서 예측할 수 있는 프로덕션 환경에 통합	• 클라우드 서비스, 에지 장치 또는 내부 서버 내에서 모델을 패키징하여 배포 • 모델이 실시간 또는 일괄 모드에서 다른 시스템이나 서비스와 상호 작용할 수 있는지 확인
5. 모니터링 및 유지보수	배포된 모델의 성능을 지속적으로 모니터링하고 시간이 지남에 따라 유지 관리	• 시간 경과에 따른 모델 성능 추적 • 배포 후에 나타나는 보안취약성을 감지하고 완화함 • 모델 및 기반 인프라에 대한 업데이트 및 패치를 구현
6. 파기	더 이상 유용하지 않거나 교체해야 할 때 AI 모델을 안전하게 폐기	• 모델을 폐기하기 전에 중요한 데이터와 로그를 백업 • 잔여 데이터나 지적 재산이 유출되지 않도록 함 • 폐기 사유와 향후 모델을 위해 얻은 교훈을 문서화

넷째, **예측형 AI(Pred AI)와 생성형 AI(Gen AI)에 맞게 구별**하여 검증항목을 제시했다.

- 요구사항별 검증항목에서 예측형 AI(이하, Pred AI)와 생성형 AI(Gen AI)를 구별하여 제시하였다. 예측형 AI(Pred AI)와 생성형 AI(이하, Gen AI)는 두 기술의 목적, 작동 방식, 위험 요소가 서로 다를 수 있으므로, 각 기술의 특성과 관련된 위협을 명확히 이해하고 이에 적합한 보안 대책을 수립하는 것이 중요하다. 따라서 본 「인공지능(AI) 보안 안내서」에서는 이를 구별하여 AI 개발자와 AI 서비스 제공자 대상 보안요구사항과 검증항목을 제시하였다.
  - Pred AI는 과거 및 현재 데이터를 사용하여 패턴을 식별하고 해당 정보를 기반으로 추론한다. 이는 주로 통계 알고리즘과 ML(기계학습)에 사용한다. 반면에 Gen AI는 한 단계 더 나아가 딥러닝을 사용하여 학습된 데이터를 기반으로 새로운 콘텐츠를 생성한다.
  - 이러한 기술적 특성으로 인해 데이터 관련 문제라 하더라도 위험 요소가 서로 다를 수 있다. 예를 들어 Pred AI는 정확한 예측을 지원할 수 있도록 고품질 데이터와 라벨링이 필요할 것이고, Gen AI는 모델의 학습 기반이 된 데이터를 제공하는 오픈소스 모델을 안전하게 사용하는 것에 초점을 맞추는 것이 중요하므로 이에 맞는 보안 검증항목이 필요하다.

- 따라서 개발자나 서비스 제공자 등은 발생한 위험이 어떠한 AI 유형과 관련이 있는지 사전에 파악하는 것이 매우 중요하고, 이를 반영한 검증항목이 필요할 것으로 예측된다. 이에 본 「인공지능(AI) 보안 안내서」에서는 요구사항 및 검증항목 별로 이를 구분하여 제시하였다. 예측형 AI와 생성형 AI에 따라 구별하여 점검하면, 각 기술에 필요한 보안 조치를 적절하게 파악하고 자원을 효율적으로 분배할 수 있어, 과도하거나 불필요한 보안 비용을 줄일 수 있을 것으로 기대된다.

2

AI 개발자를 위한 보안 안내서

01 개요

개발자 대상 「인공지능(AI) 보안 안내서」의 특징

- AI에 대한 공격 스펙트럼은 넓고 빠르게 진행되고 있으며, 설계 및 구현에서 학습 및 테스트, 실제 배포 및 파기에 이르기까지 라이프사이클의 모든 단계를 포함하고 있다. 따라서 개발자 대상 「인공지능(AI) 보안 안내서」도 생명주기에 따라 6개의 섹션(계획 및 설계 → 데이터 수집 및 준비 → 모델 개발 → 모델 배포 → 모니터링 및 유지보수 → 파기)으로 구성하였다. 특히 모델 파기 시 보안 사항은 다른 가이드라인에서는 없는 내용으로 본 「인공지능(AI) 보안 안내서」에서 이를 추가하였다.

AI 보안 안내서 생명주기
1. 계획 및 설계
2. 데이터 수집 및 준비
3. 모델 개발(학습/모델링/검증)
4. 모델 배포
5. 모니터링 및 유지보수
6. 파기

- AI 모델의 크기가 계속 커짐에 따라 많은 기업에서 직접 사용하거나 새로운 데이터 세트로 미세 조정하여 다양한 작업을 수행할 수 있는 사전 학습된 모델에 의존하는 경향이 커지고 있다. 그런데 이는 공격자가 모델 가용성을 손상시키거나 악성코드를 삽입하여 사전 학습된 모델을 악의적으로 수정할 수 있는 더 큰 기회를 제공해 주기도 한다. 따라서 「인공지능(AI) 보안 안내서」에서는 개발자들이 참고할 수 있도록 '**오픈소스 라이브러리 보안**'에 대한 요구사항 및 검증항목을 반영하였다.
- 생성형 AI와 검색 증강 생성(RAG: Retrieval-Augmented Generation) 기반의 LLM 도입이 증가하면서, 기업에서는 업무 효율성 향상, 창의성 증진, 고객 만족도 향상 등 다양한 비즈니스 혁신이 이루어지고 있다. 이처럼 LLM 기술이 활발하게 사용되고 있지만 동시에 고유한 보안 위협에 노출될 가능성을 내포하고 있다. 「인공지능(AI) 보안 안내서」는 개발자들에게 보안 관점에서 도움을 주고자 '**LLM 보안**'에 대한 요구사항 및 검증항목을 추가하였다. 이를 활용하여 개발자는 LLM의 안정성, 데이터 보호, 사용자 신뢰성을 유지하고, 적절한 방어 조치를 통해 비즈니스 지속성을 확보할 수 있을 것으로 기대된다.

개발자 대상 「인공지능(AI) 보안 안내서」의 활용 방안

- 「인공지능(AI) 보안 안내서」는 AI 제품 및 시스템 개발 현장에서 보안(Security) 관련 공격을 받거나 받을 우려가 있는 경우 데이터 과학자, 모델 개발자 등이 실제로 취해야 할 조치 사항으로 활용할 수 있다.
- 「인공지능(AI) 보안 안내서」의 요구사항 및 검증항목 점검주체는 조직의 구성원 개인을 의미하는 경우도 있으나, 조직의 부서(예컨대, 개발팀) 내지 조직 자체(예컨대, AI 소프트웨어 개발회사)를 가리키는 경우도 있다. 예를 들어, 모델 학습을 진행하는 환경이 안전하게 보안조치 되어 있는지 여부를 파악하기 위해서는 물리적 보안 외에도 방화벽·VPN 등 네트워크 보안도 함께 봐야 한다.(3.1.1. 참고) 이 경우 해당 보호조치는 개발자 개인이 아니라 전사적 차원에서 수행해야 할 업무이다. 반면에 개발자는 직무 관련 교육을 연간 100시간 이상 받아야 한다면 이 업무는 개발자 개인이 1차적인 수행주체가 된다. 「인공지능(AI) 보안 안내서」에서는 요구사항 및 검증항목에 대한 개인과 조직의 역할을 분명하게 인식할 수 있도록 해당 항목의 이행주체를 명시하였다.
- **다만, 본 안내서는 대외적으로 법적 효력을 가지는 것이 아니므로 본문의 기술 방식('~해야 한다.' '필수적이다' 등)에도 불구하고 참고로만 활용하기 바란다.**

안내서 작성 과정 및 참고 자료

- 2024년 6월부터 「AI 보안 정책 포럼」을 구성하여 운영하였고, 그 외에도 다양한 의견을 수렴하는 과정을 거쳤다. 초안 작성한 후, 학계 및 산업계 전문가 등의 의견수렴을 거쳐 〈AI 개발자를 위한 보안 요구사항 및 검증항목〉 최종본을 마련하였다.
- 미국, 유럽, 일본 등 해외기관에서 발표한 원칙 및 프레임워크 등을 참조하였고, 국내자료로는 과학기술정보통신부/한국정보통신기술협회(TTA)에서 발간한 「신뢰할 수 있는 인공지능 개발 안내서」, 국가정보원/국가보안기술연구소에서 발간한 「챗GPT 등 생성형 AI 활용 보안 가이드라인」, 금융보안원에서 발간한 「금융분야 AI 보안 가이드라인」 등도 참고하였다.

※ 주요 해외 참고자료

주관	명칭	특징
(미국) NIST	AI Risk Management Framework(AI RMF 1.0), 2023.1	글로벌 표준
(미국) ISO/IEC	ISO/IEC 42001	인증
	ISO/IEC 23894	AI에 특화
(미국) OWASP	OWASP Top 10 for LLM Applications (Ver 1.1)	개발 보안
(미국) Google	Google SAIF	개발 보안
EU	EU AI ACT	규제
OECD	OECD AI 원칙	정책 권고
(싱가폴) 정보통신부	Model AI Governance Framework, 2020	생성형 AI
(영국) NCSC	Guidelines for secure AI system development, 2023.11	가이드라인
(일본) 총무성·경제산업성	AI 사업자 가이드라인(제1.0판), 2024. 6	

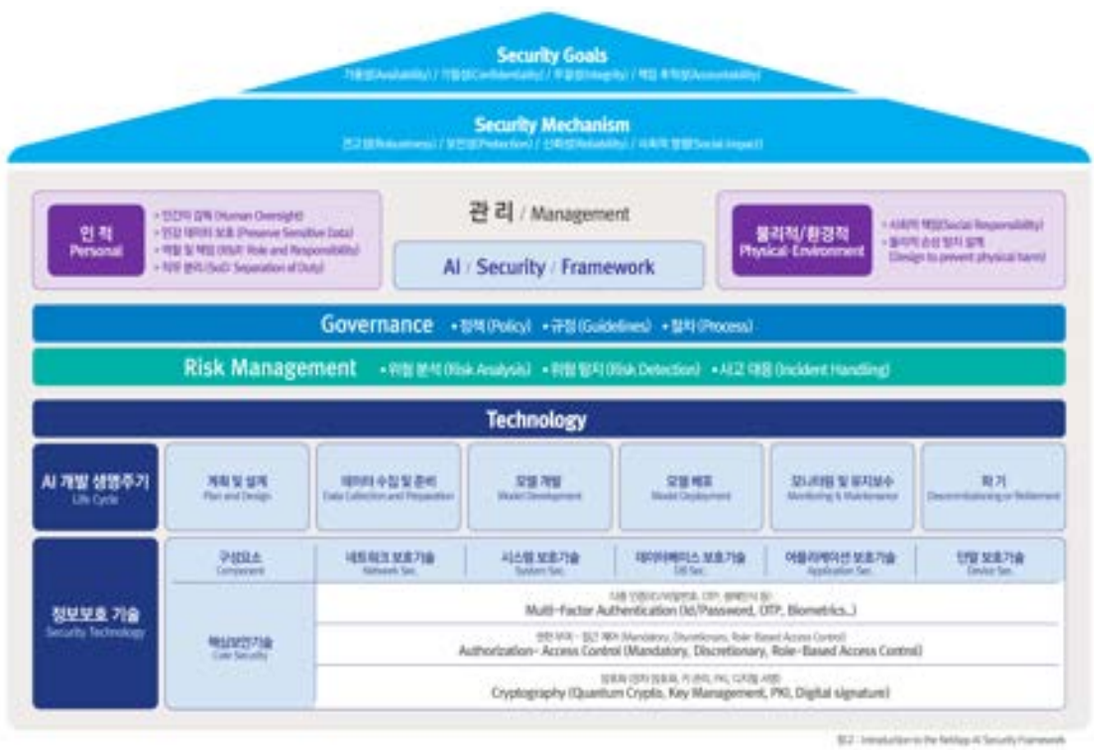
02 AI 개발자 대상 보안 프레임워크

01 AI 시스템 보안(Security) 목표

AI 시스템 보안(Security) 목표는 AI 기술을 안전하게 보호하고 신뢰할 수 있는 방식으로 운영되도록 보장하는 데 있다. 이 목표는 AI 시스템이 외부 공격, 데이터 유출, 시스템 오작동 등 다양한 보안 위협에 대응할 수 있도록 설계되고 유지되는 것을 의미한다.

- AI 시스템에서 보안의 주요 목표는 전통적인 정보보호의 3대 요소인 기밀성(Confidentiality), 무결성(Integrity), 가용성(Availability)을 기본으로 하고, AI 모델·시스템 등 개발 생애 주기에서의 책임성 확보 및 검증을 위한 Accountability(책임 추적성)를 추가하였다.

그림 2-1 AI 개발자 대상 보안 프레임워크(Security Framework)



1.1 가용성(Availability)

가용성(Availability)은 AI 보안 전략에서 중요한 요소 중 하나로, 서비스와 데이터가 항상 접근 가능하고 사용할 수 있는 상태를 유지하는 것을 목표로 한다. 이는 AI 시스템의 연속성과 신뢰성을 보장하기 위해 필수적이다.

1.2 기밀성(Confidentiality)

기밀성(Confidentiality)은 AI 보안 전략에서 중요한 요소 중 하나로, 정보가 승인된 사람만 접근할 수 있도록 보호하는 것을 목표로 한다. 이는 데이터 유출과 같은 보안 사고를 방지하고, 민감한 정보가 보호되도록 하는 데 필수적이다.

1.3 무결성(Integrity)

무결성(Integrity)은 AI 보안 전략의 중요한 요소 중 하나로, 데이터와 시스템이 허가되지 않은 변경 없이 정확하고 일관된 상태를 유지하는 것을 목표로 한다. 이는 AI 시스템이 신뢰할 수 있는 결과를 제공하고, 데이터의 변조나 손상을 방지하기 위해 필수적이다.

1.4 책임 추적성(Accountability)

책임 추적성(Accountability)은 AI 보안 전략에서 중요한 요소 중 하나로, AI 시스템의 행동과 결정에 대한 책임을 명확히 하고, 문제가 발생했을 때 원인을 추적할 수 있도록 하는 것을 목표로 한다. 이는 AI 시스템의 투명성을 높이고, 신뢰성을 보장하며, 윤리적이고 법적 기준을 준수하기 위해 필수적이다.

02 AI 시스템 보안을 위한 거버넌스(Governance)

AI 시스템 보안을 위한 조직 내 규정과 정책, 그리고 거버넌스 체계는 AI 모델이 조직에서 안전하게 운영되고, 보안 리스크를 관리하는 데 중요한 역할을 한다. 효과적인 AI 보안 거버넌스는 기술, 사람, 프로세스를 통합하여 AI 시스템의 무결성, 기밀성, 가용성을 유지하고, 법적 규제와 윤리적 기준을 준수하도록 설계되어야 한다. 이를 위해 조직은 명확한 정책과 절차를 수립하고, 보안 거버넌스를 통해 AI 시스템이 지속적으로 안전하게 운영될 수 있도록 관리해야 한다. AI 보안 거버넌스의 핵심 요소는 정책(Policy), 규정(Guidelines), 절차(Processes) 등을 포함한다.

2.1 정책(Policy)

- AI 보안 정책 수립: 조직은 AI 시스템을 안전하게 운영하기 위한 포괄적인 보안 정책을 수립해야 한다. 정책에는 AI 시스템의 설계, 개발, 배포 및 운영에 필요한 보안 기준과 절차를 명확히 정의해야 한다.
- 접근 권한 관리 정책: AI 시스템에 대한 접근 권한을 관리하기 위한 정책을 수립하여, 민감한 데이터와 시스템에 대한 무단 접근을 방지해야 한다. 역할 기반 접근 제어(RBAC)를 통해 AI 시스템에 접근할 수 있는 사용자와 권한을 제한해야 한다.
- 데이터 보호 정책: AI 시스템이 다루는 데이터의 기밀성과 무결성을 보장하기 위해 데이터 암호화, 데이터 저장소 보호에 대한 규정을 마련해야 한다.

2.2 규정(Guidelines)

- AI 모델 관리 규정: AI 모델의 학습, 배포 및 사용에 대한 구체적인 지침을 제공해야 한다. 이 규정은 AI 모델의 검증 및 테스트 절차, 안전한 데이터 사용, 모델 업데이트 및 폐기 절차를 포함한다.
- AI 시스템 모니터링 규정: AI 시스템이 보안 위협에 대응할 수 있도록 실시간 모니터링 및 로그 분석 규정을 마련해야 한다. 이를 통해 시스템 내 이상 활동이나 보안 위협을 조기에 감지하고 대응할 수 있어야 한다.
- 리스크 관리 규정: AI 시스템에 내재된 보안 리스크를 식별하고 관리하기 위한 규정을 수립해야 한다. 이는 잠재적인 보안 위협을 사전에 평가하고 완화하는 절차를 포함한다.

2.3 절차(Process)

- AI 보안 평가 절차: AI 시스템 개발 초기부터 보안 평가 절차를 구축하여, 보안 위험을 사전에 감지하고 해결해야 한다. 이 절차는 보안 취약점 분석, 침투 테스트 및 코드 리뷰를 포함할 수 있다.
- AI 시스템 업데이트 및 패치 절차: AI 시스템과 모델이 최신 보안 위협에 대응할 수 있도록 정기적인 업데이트 및 패치 절차를 수립해야 한다. 이 절차는 새로운 보안 위협에 대한 대응 방안을 포함하며, 보안 업데이트가 원활하게 이루어지도록 해야 한다.
- 비상 대응 계획: 보안 사고가 발생했을 때 즉각적으로 대응할 수 있는 비상 대응 계획을 마련해야 한다. 여기에는 사고 보고, 대응 팀 구성, 시스템 복구 및 데이터 손실 방지 절차가 포함된다.

03 AI 시스템 보안을 위한 위험 관리(Risk Management)

위험 관리(Risk Management)는 AI를 활용한 서비스에서 필수적인 요소로, AI 시스템이 직면할 수 있는 잠재적 위험을 식별하고, 이를 효과적으로 관리하고 대응하기 위한 체계적인 접근을 포함한다. 위험 관리는 위험 분석, 위험 감지, 사고 대응의 세 가지 주요 단계로 나눌 수 있다.

3.1 위험 분석(Risk Analysis)

- 위험 분석(Risk Analysis)은 AI 시스템이 직면할 수 있는 잠재적 위험을 식별하고, 그 심각성과 발생 가능성을 평가하는 과정이다. 이 과정은 위험 관리의 첫 단계로, 체계적인 접근을 통해 위험 요소를 사전에 파악하고 대응 전략을 마련하는 것을 목표로 한다.
- 주요 활동
  - 위험 식별 (Risk Identification): AI 시스템과 관련된 모든 잠재적 위험을 식별하고, 기술적·운영적·윤리적·법적 측면에서 발생할 수 있는 다양한 위험을 고려한다.
    - 예시 데이터 손실, 시스템 오류, 보안 침해, 윤리적 문제 등
  - 위험 평가 (Risk Assessment): 식별된 위험의 심각성과 발생 가능성을 평가하고, 각 위험 요소의 영향을 분석하여 우선순위를 정한다.
  - 위험 대응 계획 (Risk Mitigation Plan) 수립: 평가된 위험에 대한 대응 계획을 수립하고, 위험을 줄이기 위한 예방 조치와 대응 전략을 마련한다.
    - 예시 보안 강화, 데이터 백업, 긴급 대응 절차 마련 등

3.2 위험 감지(Risk Detection)

- 위험 감지(Risk Detection)는 AI 시스템의 운영 중 발생하는 이상 징후나 위험 요소를 실시간으로 감지하고, 이를 신속하게 보고하는 과정이다. 이는 위험 발생 시 신속한 대응을 위해 중요한 단계이다.
- 주요 활동
  - 모니터링 시스템 구축 (Establish Monitoring Systems) : AI 시스템의 성능과 안전성을 실시간으로 모니터링하는 시스템을 구축한다.
    - 예시 네트워크 트래픽 모니터링, 시스템 로그 분석, 사용자 활동 추적 등

- 이상 징후 탐지 (Anomaly Detection) : 정상적인 패턴에서 벗어난 이상 징후를 자동으로 감지할 수 있는 알고리즘과 기술을 도입한다.

  예시 머신러닝 기반 이상 탐지 알고리즘, 실시간 경고 시스템 등

- 자동화된 경고 (Automated Alerts) : 이상 징후나 위험 요소 감지 시 즉각적으로 관련 담당자에게 경고를 보내는 자동화된 시스템을 운영한다.

  예시 이메일 알림, SMS 경고, 대시보드 알림 등

- 정기적 검토 (Regular Reviews) : 모니터링 결과와 경고 로그를 정기적으로 검토하여, 새로운 위험 요소나 패턴을 식별하고 대응 방안을 업데이트한다.

  예시 주간/월간 보고서 작성, 경고 로그 분석 회의 등

3.3 사고 대응(Incident Handling)

- 사고 대응(Incident Handling)은 실제로 위험이 발생했을 때 이를 효과적으로 대응하고 해결하는 과정이다. 이는 신속한 대응과 문제 해결을 통해 피해를 최소화하고, 재발 방지를 위한 교훈을 도출하는 것을 목표로 한다.
- 주요 활동
  - 사고 대응 절차 수립 (Establish Incident Response Procedures) : 사고 발생 시 신속하고 체계적으로 대응할 수 있는 절차와 계획을 마련한다.

    예시 사고 대응 매뉴얼 작성, 비상 연락망 구축, 역할 및 책임 정의 등

  - 사고 대응 팀 구성 (Form Incident Response Team) : 사고 발생 시 대응할 전담 팀을 구성하고, 각 팀원의 역할과 책임을 명확히 한다.

    예시 보안 담당자, 데이터 과학자, IT 지원팀 등으로 구성

  - 초기 대응 및 완화 (Initial Response and Mitigation) : 사고 발생 시 초기 대응을 통해 피해를 최소화하고, 추가 피해를 방지하기 위한 조치를 취한다.

    예시 시스템 격리, 데이터 복구, 보안 패치 적용 등

  - 사고 분석 및 보고 (Incident Analysis and Reporting) : 사고의 원인을 분석하고, 사고 발생 과정과 대응 결과를 상세히 기록하여 보고한다.

    예시 사고 원인 분석 보고서 작성, 대응 결과 리뷰 등

  - 사후 조치 및 재발 방지 (Post-Incident Actions and Prevention) : 사고 종료 후 사후 조치를 취하고, 재발 방지를 위한 교훈을 도출하여 시스템과 절차를 개선한다.

    예시 시스템 업데이트, 보안 정책 강화, 교육 프로그램 운영 등

- 예시 시나리오
  - AI 모델 오류 사고 대응
    - ▸위험 분석: AI 모델의 예측 오류 가능성을 평가하고, 모델 검증 및 테스트 계획 수립
    - ▸위험 감지: AI 모델의 실시간 성능 모니터링 시스템 도입, 예측 오류 발생 시 경고 시스템 운영
    - ▸사고 대응: AI 모델의 예측 오류 발생 시 즉시 모델 사용 중지 및 수정 작업, 오류 원인 분석 및 보고, 수정된 모델의 검증 및 테스트 후 재배포, 재발 방지를 위한 모델 검증 절차 강화

03 AI 개발자를 위한 요구사항 및 검증항목

체크리스트 요약

생명주기	요구사항 및 체크리스트	개발자		AI 유형	
		담당자	조직	Pred AI	Gen AI
1 계획 및 설계	**(AI 개발자, AI 서비스 제공자 공통사항) 거버넌스 및 위험관리**				
	1.1 AI 보안(Security) 거버넌스 체계 구축 AI 개발자, AI 서비스 제공자 공통사항				
	1.1.1 AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?		○	○	○
	1.1.2 AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?		○	○	○
	1.1.3 AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?		○	○	○
	1.2 AI 모델개발에 대한 위험관리 계획의 수립 AI 개발자, AI 서비스 제공자 공통사항				
	1.2.1 AI 모델 개발/서비스 제공 생명주기 및 공급망 과정에서 나타날 수 있는 위험요소를 분석·도출하고 있는가?	○		○	○
	1.2.2 AI 시스템에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	○		○	○
	1.2.3 AI 시스템에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	○		○	○
2 데이터 수집 및 준비	2.1 데이터 수집 및 전처리				
	2.1.1 데이터 수집 시 사용되는 네트워크 프로토콜이 충분한 보안 기능을 제공하고 있는가?	○		○	○
	2.1.2 수집된 데이터의 보관 및 삭제 절차가 명확하게 정의되어 있는가?	○	○	○	○
	2.1.3 전처리 과정에서 중요 데이터를 보호하기 위해 암호화 기술을 사용하고 있는가?	○		○	○
	2.2 데이터 무결성 검증 AI 개발자, AI 서비스 제공자 공통사항				
	2.2.1 데이터 처리 과정에서 데이터 무결성을 검증하고 있는가?	○		○	○
	2.2.2 데이터에 접근할 수 있는 권한을 제한하고 있는가?	○	○	○	○
	2.3 데이터 공격에 대한 방어 AI 개발자, AI 서비스 제공자 공통사항				
	2.3.1 데이터 중독(poisoning) 공격에 대한 방어 대책을 마련하고 있는가?	○		○	○
	2.3.2 데이터 회피(evasion) 공격에 대한 방어 대책을 마련하고 있는가?	○		○	○
	2.3.3 데이터 유출·변조 공격을 방지하기 위한 방안을 마련하고 있는가?	○		○	○

생명주기	요구사항 및 체크리스트		개발자		AI 유형	
			담당자	조직	Pred AI	Gen AI
3 모델개발 (학습/ 모델링/ 검증)	3.1	학습/검증 환경에 대한 보안(Secure Training Environment)				
	3.1.1	모델 학습을 진행하는 환경이 안전하게 보안조치 되어 있는가?	○	○	○	○
	3.1.2	학습 또는 검증 단계에서 악의적인 사용자가 허위 데이터를 삽입할 가능성을 차단하고 있는가?	○		○	○
	3.1.3	연합 학습(Federated Learning)에 참여하는 장치 중 악의적인 장치가 있는지 검증하고 있는가?	○		○	○
	3.2	모델 공격에 대한 방어				
	3.2.1	AI Prompt Injection 공격에 대한 방어 방안을 수립하고 있는가?	○			○
	3.2.2	적대적 예제 공격 (Adversarial Example Attacks)에 대한 방어 방안을 수립하고 있는가?	○		○	○
	3.2.3	모델 회피 공격(model evasion attack)에 대한 방어 방안을 수립하고 있는가?	○		○	○
	3.2.4	모델 오염 공격(Model Poisoning Attack)에 대한 방어 방안을 수립하고 있는가?	○		○	○
	3.2.5	모델 추출 공격(model extraction attack) 및 리버스 엔지니어링에 대한 방어 방안을 수립하고 있는가?	○		○	○
	3.2.6	반복적인 질의에 대한 방어 방안을 수립하고 있는가?	○			○
	3.2.7	기계 학습을 활용한 모델 공격에 대해 능동적으로 방어하고 있는가?	○		○	○
	3.3	오픈소스 라이브러리 보안				
	3.3.1	오픈소스 라이브러리의 업데이트 및 취약점을 관리하고 있는가?	○		○	○
	3.3.2	오픈소스 라이브러리의 소스 코드를 직접 검토하거나 사용에 대한 보안 문제를 검증하고 있는가?	○		○	○
	3.3.3	오픈소스 라이브러리를 실행할 때 잠재적인 보안 위험을 제거하기 위해 격리된 환경을 이용하고 있는가?	○		○	○
	3.4	LLM 보안				
	3.4.1	LLM 애플리케이션 공격에 대한 예방책을 마련하고 있는가?	○			○
	3.4.2	LLM의 모델 서비스 거부(Model Denial of Service) 공격에 대한 방어 방안을 수립하고 있는가?	○			○
	3.4.3	LLM의 API 보안을 위한 방안을 수립하고 있는가?	○			○
	3.4.4	LLM의 인터페이스 공격에 대한 예방책을 마련하고 있는가?	○			○
	3.4.5	개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립하고 있는가?	○	○		○
	3.4.6	LLM 출력결과를 정기적으로 모니터링하고 검토하고 있는가?	○			○
	3.4.7	LLM의 벡터 및 임베딩 취약점에 대한 방어 방안을 수립하고 있는가?	○			○

생명주기	요구사항 및 체크리스트		개발자		AI 유형	
			담당자	조직	Pred AI	Gen AI
4 모델 배포	4.1	모델파일 및 배포 환경 보호				
	4.1.1	모델을 배포하기 전에 코드 및 모델을 스캔하고, 자동화된 취약점 분석을 하고 있는가?	○		○	○
	4.1.2	모델파일을 암호화하여 저장하고 전송 중에도 안전하게 보호하고 있는가?	○		○	○
	4.1.3	AI 모델이 배포되는 인프라(클라우드, 서버 등) 환경이 충분한 보안시스템을 갖추고 있는가?	○		○	○
	4.2	API 및 인터페이스 보안				
	4.2.1	AI 모델이 배포된 후, API를 통해 외부 시스템과 상호작용하는 경우, 충분한 보안 조치 기능을 갖추고 있는가?	○	○	○	○
	4.2.2	배포된 AI 모델이 실시간으로 데이터를 수신하고 이를 처리할 때, 중간자 공격(Man-in-the-Middle Attack)에 대응하고 있는가?	○		○	○
	4.2.3	AI 모델의 API에 대한 접근 권한을 제한하고, 강한 인증 메커니즘을 사용해 불법 접근을 방지하고 있는가?	○		○	○
	4.2.4	API 사용자는 필요한 권한만 부여받도록 최소 권한(Least Privilege) 원칙을 적용하고 있는가?	○		○	○
	4.2.5	AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 하고 있는가?	○		○	○
5 모니터링 및 유지보수	5.1	실시간 모니터링 AI 개발자, AI 서비스 제공자 공통사항				
	5.1.1	모델의 입력 데이터, 출력 결과 등을 실시간으로 모니터링하여 비정상적인 동작을 탐지하고 있는가?	○		○	○
	5.1.2	모델 응답 시간, 사용 패턴을 추적하고 분석하여 보안에 의심스러운 행동을 탐지하고 있는가?	○		○	○
	5.1.3	AI 모델이 동작하는 서버 및 네트워크의 트래픽을 모니터링하여 비정상적인 요청을 탐지하고 있는가?	○		○	○
	5.1.4	API 호출, 입력/출력 등 요청로그를 정기적으로 분석하여 보안에 의심스러운 동작을 탐지하고 있는가?	○		○	○
	5.1.5	AI 모델과 배포 환경에 대해 모의 해킹을 수행하여 잠재적인 보안 취약점을 탐지하고 수정하고 있는가?	○		○	○
	5.2	보안 패치 및 업데이트 관리 AI 개발자, AI 서비스 제공자 공통사항				
	5.2.1	모델에 대한 보안 패치 및 업데이트 관리 프로세스를 구축하고 있는가?	○		○	○
	5.2.2	모델 배포 후 모델 및 라이브러리의 업데이트가 정기적으로 이루어지고 있는가?	○		○	○
	5.2.3	운영 체제, 라이브러리, 프레임워크의 보안 패치를 운영 환경에 적용하기 전에 스테이징 환경에서 패치를 테스트하고 있는가?	○		○	○
6 파기	6.1	파기 시 보안 AI 개발자, AI 서비스 제공자 공통사항				
	6.1.1	AI 모델이 더 이상 사용되지 않으면, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?	○		○	○
	6.1.2	AI 모델에서 사용 중이던 데이터가 시스템을 폐기하거나 교체할 때 안전하게 삭제되고 있는가?	○		○	○
	6.1.3	AI 모델이 더 이상 사용되지 않으면, 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?	○		○	○

01 계획 및 설계

1.1 AI 보안(Security) 거버넌스 체계 구축

AI 개발자, AI 서비스 제공자 공통사항

- AI 모델을 개발하는 기업과 AI 서비스를 제공하는 기업에서는 AI 보안(Security) 거버넌스 체계를 구축하기 위해 AI 보안 정책 및 절차 정의, AI 보안 조직 및 역할 정의 등과 같은 AI 보안 거버넌스 프레임워크를 수립하고, AI 보안 위험 분석 및 관리 체계 등을 구축해야 함

1.1.1 AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?

YES ☐ NO ☐ N/A ☐

- AI 시스템은 보안와 관련된 문제가 발생할 수 있다는 위험 요소가 존재하므로 다양한 위험 요소를 인식하고 관련 규정을 마련하여 이를 실행할 수 있도록 관리 및 감독하는 조직이 필요함
- AI 보안(Security) 거버넌스를 위한 조직을 구성하기 위해서는 조직 내 역할과 책임(Roles and Responsibilities)도 명확하게 정의해야 함
  - 역할 분담: 보안 정책 수립, 기술 구현, 규제 준수 등 각 부서의 책임을 명확히 정의
  - 독립성: 보안 거버넌스 조직은 개발팀과 독립적으로 운영되어 객관성을 유지해야 함
  - 리더십 확보: 보안 관련 결정을 내릴 수 있는 충분한 권한을 부여

(예시) 명확한 역할과 책임

- CISO(Chief Information Security Officer): 보안 전략 및 거버넌스의 총괄 책임자로서 정책 수립과 이행 감독
- 데이터 보안팀: 데이터 암호화, 접근 제어, 민감 데이터 보호 관리
- AI 개발팀과 협업: AI 모델 개발 단계부터 보안 요소를 반영하도록 보안 관점에서 협업 필요
- 규제 준수 담당자: 글로벌 및 로컬 보안 규제 등에 대한 모니터링 및 준수 확인

1.1.2 AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델을 개발하는 기업과 AI 서비스를 제공하는 기업이 AI 보안(Security) 거버넌스 정책 수립을 위해서는 다음과 같은 내용을 수행해야 함
  - AI 보안 원칙 및 목표 정의
    - AI 모델 및 서비스의 보안, 데이터 무결성, 서비스 신뢰성을 보장하기 위한 핵심 원칙 설정
    - AI 보안 비전과 목표를 명확히 수립
  - AI 보안 정책 문서화 및 실행
    - AI 보안 가이드라인 작성(데이터 보호, 모델 보안, API 보안, 서비스 보호 등)
    - 보안 거버넌스 조직 구성(CISO, AI 보안팀, 데이터 보호 책임자 등 역할 정의)
    - AI SW, HW 공급망 보안 관리 체계 구축(AI 모델 알고리즘 정보, 소스코드 정보, AI 칩 정보, 보안 패치 적용 및 이력 등)

- AI 보안 규제 및 컴플라이언스 준수
  - GDPR, CCPA, EU AI Act, NIST AI RMF, ISO/IEC 27001 등 글로벌 규제 및 표준 준수
  - 법적 리스크 평가 및 보안 정책과의 정합성 검토

• AI 모델을 개발하는 기업과 AI 서비스를 제공하는 기업은 AI 보안(Security) 거버넌스를 위해 AI 보안 감사 및 지속적인 보안 거버넌스 운영 체계도 갖춰야 함

- AI 보안 거버넌스 체계 지속 점검 및 개선
  - 정기적인 AI 보안 감사(Security Audit) 및 리스크 평가 수행
  - AI 보안 정책 및 절차의 지속적인 개선 및 업데이트
- AI 보안 교육 및 인식 제고
  - 개발자 및 운영자를 위한 AI 보안 교육 프로그램 운영
  - AI 보안 사고 사례 공유 및 대응 훈련 수행
- AI 보안 사고 대응 및 위기관리 프로세스 구축
  - AI 보안 사고 대응 계획 수립 및 시뮬레이션 테스트 진행
  - 보안 사고 발생 시 보고 및 대응 절차 운영

AI 거버넌스 체계 구축 관련 요구사항 예시

- AI와 관련된 법률 및 규제 요구 사항을 이해하고 관리하며 문서화함
- 조직의 위험 허용 범위를 기반으로 필요한 위험 관리 활동 수준을 결정하기 위한 프로세스, 절차 및 관행이 마련되어야 함
- 위험 관리 프로세스와 그 결과는 조직의 위험 우선 순위에 따라 투명한 정책, 절차 및 기타 통제를 통해 설정
- 위험 관리 프로세스와 그 결과에 대한 지속적인 모니터링과 정기 검토가 계획되고, 정기 검토 빈도 결정을 포함하여 조직의 역할과 책임이 명확하게 정의
- 위험을 증가시키거나 조직의 신뢰성을 저하시키지 않는 방식으로 AI 시스템을 안전하게 폐기하고 단계적으로 폐지하기 위한 프로세스와 절차 마련

NIST, AI Risk Management Framework 참조

1.1.3 AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 보안 거버넌스 전문인력(Security Governance Specialists)은 AI 모델 및 서비스의 보안성을 보장하고, 보안 정책을 수립하며, 규제 준수를 이행하는 역할을 수행함. 이들의 업무는 크게 정책 수립, 보안 리스크 분석, 데이터 보호, AI 모델 및 인프라 보안 강화, 법적 준수, 보안 모니터링 및 대응으로 나눌 수 있음
  - AI 보안 정책 및 거버넌스 체계 구축 - AI 보안 원칙, 정책, 운영 가이드라인 수립 등
  - AI 보안 리스크 분석 및 위협 모델링 수행 - AI 서비스의 보안 취약점 평가 등
  - 컴플라이언스 및 법적 규제 준수 - GDPR, CCPA, AI 윤리 원칙 준수 등
  - AI 보안 사고 대응 및 위기 관리 운영 - 보안 사고 대응 계획, 보안 로그 모니터링 등
  - AI 보안 교육 및 내부 인식 강화 - AI 보안 교육 제공, 보안 가이드 문서화 등
- AI 모델 보안을 담당하는 전문 인력은 AI 기술과 보안 기술 모두에 대한 깊은 이해를 갖추고 있어야 하며, 동시에 윤리적 기준과 법적 규제를 준수하는 능력도 중요함. 기술적·윤리적·관리적 역량을 통합적으로 갖춘 인력이 기업의 AI 보안 거버넌스 성공의 핵심임
  - 기술적 역량: AI 및 머신러닝 관련 기술, 데이터 보안, 네트워크 및 시스템 보안, 보안 자동화 이해 등
  - 윤리적 역량: AI 윤리 및 공정성, 규제 및 법적 지식 등
  - 관리적 역량: 위험 관리, 사고 대응, 거버넌스 정책 설계, 협업 능력 등 소프트 스킬 등
- 거버넌스 담당자는 AI 시스템 생명주기에 따라 조직이 보안 관련 내부 규정을 준수함을 확인·감독해야 함. 또한, 정의된 규정을 실행하고 관리하기 위해 각 담당자에게 관련 교육을 충분히 제공해야 함

1.2 AI 모델개발/서비스 제공에 대한 위험관리 계획의 수립 AI 개발자, AI 서비스 제공자 공통사항

- AI 시스템에 대한 보안 측면에서의 위험관리 계획은 모델 도난, 해킹, 기밀 데이터 유출, 서비스 마비 등의 위협을 방지하기 위해 필수적임. 모델개발/서비스 운영 생명주기에 걸쳐 나타날 수 있는 위험요소를 분석하고, 지속적인 보안 점검, 접근 통제, 데이터 보호, 모델 무결성 유지 등의 전략을 통해 AI 시스템에 대한 위험요소를 제거·완화해야 함

1.2.1 AI 모델 개발/서비스 제공 생명주기 및 공급망 과정에서 나타날 수 있는 위험요소를 분석·도출하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델개발/서비스 제공 생명주기 전반에서 보안 위험 요소를 분석·도출하는 것이 중요한 이유는 다음과 같음
  - 데이터 및 모델의 보안 취약점으로 인한 피해 예방
  - AI 모델의 신뢰성과 무결성 유지, AI 서비스의 악용 방지
  - AI 시스템의 안정성과 지속 가능성 확보
  - 개발/서비스 기업의 경제적 손실 방지 및 지속적적인 성장과 경쟁력 확보
- 따라서 AI 모델개발/서비스 제공 기업에서 보안 위험을 사전에 분석하고 대응 전략을 마련하는 것은 AI 기술/서비스의 안정성과 지속 가능성을 보장하는 필수적인 과정임
- AI 모델개발/서비스 제공 생명주기 전반에서 보안 위험 요소를 도출하려면 각 단계별로 발생 가능한 위협을 분석하고, 보안 정책 및 기술적 대응 방안을 마련하는 것이 필수적임. 이를 위해 보안 테스트, 위협 모델링, 침투 테스트, 지속적인 모니터링 및 AI 보안 거버넌스 구축을 적극적으로 수행해야 함

예방·대응 필요한 보안 관련 취약점 예시

1. 데이터와 관련된 취약점
   - 데이터 무결성 문제: AI 모델이 학습하는 데이터셋이 조작되거나 변조되면, 모델이 왜곡된 결과를 생성할 수 있음
2. 알고리즘과 모델의 취약점
   - 모델 탈취: 공격자가 AI 모델에 접근해 모델의 내부 구조와 학습 데이터를 복제하거나 악용할 수 있음
   - 모델 업데이트의 취약점: 지속적으로 업데이트되는 AI 시스템이 악의적인 코드나 데이터를 통해 오염될 위험도 존재할 수 있음
   - 소스코드 및 라이브러리의 보안 취약점으로 인한 서비스 마비, 데이터 유출 등 보안 위협 발생 우려
   - Prompt Injection, 악의적인 질의, 우회된 질의 등을 악용한 AI 탈옥(AI Jailbreak)으로 인해 AI 행동 유도, 유해한 콘텐츠 생성, 범죄 악용 등이 발생할 수 있음
3. AI 시스템 인프라의 취약점
   - API 및 인터페이스 보안: AI 시스템의 API가 부적절하게 보호되면, 공격자가 시스템을 오용하거나 데이터를 유출할 수 있음
   - AI 시스템, 서비스에 대한 공격(DDoS 등)으로 인한 서비스 장애, 서비스 오류 등이 발생할 수 있음

- 또한 AI 모델 개발/서비스 제공하는 과정에서 AI와 관련된 SW, HW 등 공급망 보안 관리 체계를 구축하고 관리하여야 함.
  - AI 모델개발 및 서비스 제공과정에서의 개발, 제조, 장비 증 구매, 배포, 통합 운영 및 유지 보수 또는 폐기 등 공급망 요소에서 모델과 서비스의 보안 취약점 등을 악용하여 AI 보안 위험이 발생할 수 있음

1.2.2 AI 시스템에 대한 위협 모델링 및 위험 평가를 수행하고 있는가? YES ☐ NO ☐ N/A ☐

- 위협 및 위험 평가(TRA: Threat and Risk Assessment)는 시스템에 대한 다양한 위협과 취약성을 식별하고 이러한 시스템이 노출되는 위험 수준을 결정하며 적절한 보호 수준을 권장하는 체계적인 프로세스임
  - 위협 및 위험 평가를 수행하려면 먼저 시스템의 보안 분류를 결정하는 것이 중요함
  - 보안 분류는 위험을 평가할 때 위협 및 취약성 정보와 함께 사용됨
- TRA의 목적은 위험을 최소화하면서 기밀성, 무결성 및 가용성 보호를 극대화하는 것이며, 일반적으로 TRA에는 다음의 사항이 포함됨; ①기능 요구사항 사양 검토, ②위협 및 취약점 식별, ③위험 식별, 분석 및 평가, ④적절한 보안 통제에 대한 권장 사항
- AI 위험을 다음과 같이 평가함; ①위험 분석 결과를 위험 기준과 비교함, ②위험 처리를 위해 평가된 위험의 우선순위를 정함
- 조직은 AI 위험 평가 프로세스에 대해 문서화된 정보로 관리/유지해야 함
- (예시) 공격 벡터 식별
  - 공격 표면 매핑: 시스템이 외부와 상호작용하는 지점을 분석하여 공격 가능성을 확인
    - ▸데이터 입력 및 출력 채널, API 엔드포인트, 모델 업데이트 또는 재훈련 과정
  - 잠재적 위협 목록 작성
    - ▸데이터 수준 위협: 데이터 오염, 데이터 유출, 적대적 데이터 공격 등
    - ▸모델 수준 위협: 모델 도용, 적대적 샘플 공격, 모델 역공학, AI 탈옥 등
    - ▸시스템 수준 위협: 인증 우회, API 악용, 서비스 거부(DoS) 공격 등

1.2.3 AI 시스템에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가? YES ☐ NO ☐ N/A ☐

- AI를 구현하는 이해관계자는 위험 요소에 대한 대응 방안을 마련하고, 위험이 제거 및 완화되었는지 확인해야 함
- 대응 방안이란, 구현 및 운영 방식 등의 절차, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등 기술적으로 적용할 수 있는 모든 방법을 의미하고, 위험 요소의 분석 과정에서 평가한 파급효과가 가장 큰 위험 요소를 우선적으로 대응해야 함
- 대응 방안이 적용된 이후에는 파급효과를 재평가함으로써 위험 요소가 실제로 제거, 방지 혹은 이의 영향이 완화되었는지 확인해야 함
- (예시) 설계 및 개발 단계에서의 보안 강화
  - 보안 중심 설계(Security by Design): AI 시스템을 설계할 때부터 보안을 우선시하는 접근법을 적용
  - AI 탈옥, 모델 해킹 등 방지 : AI 모델에 대한 적대적 공격, 프롬프트 인젝션 등의 공격을 방지하기 위한 적대적 훈련을 실시 및 검증
  - 데이터 무결성 검증: 학습 데이터의 품질과 신뢰성을 확인하고, 데이터 중복 또는 오염을 방지
  - 개발 소스코드, 라이브러리 등의 보안 취약점 점검 및 시큐어코딩 적용 등 개발 보안 가이드라인 및 SW 공급망 보안 가이드라인 등 준수

02 데이터 수집 및 준비

2.1 데이터 수집 및 전처리

• AI 모델은 중요 데이터를 보호하고, 승인되지 않은 접근이나 유출을 방지하여야 함. 이는 사용자의 신뢰를 유지하고, 법적 및 규제 요구 사항을 준수하며, 기밀성/무결성을 보장하기 위해 필수적임.

2.1.1 데이터 수집 시 사용되는 네트워크 프로토콜이 충분한 보안 기능을 제공하고 있는가?

YES ☐ NO ☐ N/A ☐

• 데이터 수집 시 사용되는 네트워크 프로토콜이 충분한 보안 기능을 제공하지 않으면, 데이터가 전송 중에 공격자에게 탈취되거나 변조될 수 있음
• 데이터 전송 중에 발생하는 공격으로 인해 데이터 유출이나 변조가 발생하면, AI 모델에 잘못된 데이터가 제공되거나 중요한 정보가 손실될 수 있음
  - 데이터 수집 과정에서 공격자가 데이터를 조작하거나 변조할 경우, 신뢰할 수 없는 데이터가 AI 모델에 전달됨. 특히 센서나 IoT 장치에서 수집된 데이터는 외부 요인에 의해 쉽게 변조될 수 있음
  - 변조된 데이터는 AI 시스템의 예측 정확도를 크게 떨어뜨리며, 특히 의사결정이 중요한 분야(의료, 금융)에서 큰 문제를 야기할 수 있음
• (예시) 사용 중인 프로토콜 파악
  - 프로토콜 유형 식별: HTTP/HTTPS, FTP/SFTP, MQTT, WebSocket 등 사용 중인 프로토콜 확인
  - 전송 계층 확인: TCP, UDP 등 기본 전송 계층 프로토콜도 파악
• (예시) 보안 표준 준수 여부 검토
  - 암호화 프로토콜 확인: HTTPS, SFTP, VPN 등
  - 인증 메커니즘: 프로토콜에서 다중 인증(Multi-Factor Authentication, MFA) 지원 여부

2.1.2 수집된 데이터의 보관 및 삭제 절차가 명확하게 정의되어 있는가? YES ☐ NO ☐ N/A ☐

- 수집된 데이터의 보관 및 삭제 절차가 명확하지 않거나, 삭제 정책이 적용되지 않으면 오래된 데이터나 필요 없는 데이터가 계속 남아 있어 보안 위협이 될 수 있음
  - 보유 기간이 만료된 데이터가 지속적으로 저장되면, 불필요한 위험이 증가하며, 해커가 공격 시 사용할 수 있는 표적이 될 수 있음
- 데이터 완전 삭제 및 파기 : 데이터가 저장된 모든 장치에서 데이터를 완전히 삭제하고, 중요 데이터는 물리적으로 파기함. 이를 위해 데이터 완전 삭제(secure erase)나 장치 파쇄 등의 방법을 사용함
- (예시) 보관 절차 정의
  - 데이터 암호화: 저장 시 안전한 암호화 방식 채택
  - 접근 통제: 권한이 있는 사용자만 데이터에 접근할 수 있도록 설정
  - 백업 정책: 필요 시 데이터 복구를 위한 백업 절차 마련
- (예시) 삭제 절차 정의
  - 삭제 방법: 완전 삭제(예: 디지털 삭제 기술, 영구 삭제), 익명화(비식별화)
  - 자동화된 삭제 시스템: 보관 기간 종료 시 자동으로 삭제되는 메커니즘 구축
  - 감사 로그 유지: 삭제 기록을 추적하고 검증 가능하게 유지

2.1.3 전처리 과정에서 중요 데이터를 보호하기 위해 암호화 기술을 사용하고 있는가? YES ☐ NO ☐ N/A ☐

- 데이터 수집 및 처리 과정에서 암호화가 적용되지 않으면, 민감한 정보가 외부 공격자에게 쉽게 노출될 수 있음.
- 암호화되지 않은 데이터는 네트워크 스니핑, 중간자 공격(MITM) 등에 취약하며, 이는 데이터 유출로 이어질 수 있음
- 중요 데이터 보호
  - 기밀 정보 보호: 전처리 단계에서는 데이터를 수집, 정리, 변환하는 과정에서 원본 데이터가 노출될 가능성이 높음. 암호화를 사용하면 중요 데이터가 보호됨
- (예시) 암호화 적용 방법
  - 저장 데이터 암호화 (At-Rest Encryption)
    - ▸ 데이터가 저장될 때 암호화 적용
    - ▸ 파일 수준 또는 데이터베이스 암호화 사용
    - ▸ 기술 예시: AES(Advanced Encryption Standard) 256비트, TDE(Transparent Data Encryption) for databases

- 전송 데이터 암호화 (In-Transit Encryption): 네트워크를 통해 데이터를 전송할 때 보호
  ▸ 기술 예시: HTTPS/TLS, VPN, SSH
- 처리 중 데이터 암호화 (In-Use Encryption): 전처리 단계에서 데이터를 처리할 때도 보호
  ▸ 기술 예시: 동형 암호화 (Homomorphic Encryption), 암호화 메모리(Intel SGX, AMD SEV)

• (예시) 암호화 단계별 사용
  - 데이터 수집 후 암호화: 데이터를 수집한 즉시 암호화하여 원본 데이터를 보호
  - 전처리 중 암호화 유지: 동형 암호화나 암호화 연산 라이브러리 사용
  - 전처리 후 암호화된 출력 저장: 전처리 후 데이터 결과도 암호화하여 저장
• (예시) 적용 가능한 암호화 기술
  - AES (Advanced Encryption Standard): 대칭 키 암호화 방식, 빠르고 효율적이며 저장 및 전송 데이터 암호화에 적합
  - RSA (Rivest-Shamir-Adleman): 비대칭 키 암호화 방식, 키 관리 및 소량의 데이터 암호화에 적합
  - 동형 암호화 (Homomorphic Encryption): 암호화 상태에서 데이터 연산이 가능
  - 해시 함수: 데이터 식별을 위해 비가역적 해싱 적용, 예: SHA-256, bcrypt
• (예시) 암호화 키 관리
  - 키 생성: 안전한 키 생성 방식 사용
  - 키 저장: 하드웨어 보안 모듈(Hardware Security Module, HSM) 또는 키 관리 서비스(KMS) 활용
  - 키 교체 및 폐기: 키 교체 주기와 폐기 절차 명확화

2.2 데이터 무결성 검증

AI 개발자, AI 서비스 제공자 공통사항

• AI 모델 개발 시 데이터 처리 과정에서 데이터 무결성 검증의 목적은 데이터가 정확하고, 완전하며, 변조되지 않았음을 보장하여 신뢰할 수 있는 결과를 도출하는 데 있음

2.2.1 데이터 처리 과정에서 데이터 무결성을 검증하고 있는가?

YES ☐ NO ☐ N/A ☐

• 데이터 무결성의 검증은 공격자의 침해 영향 최소화에 도움이 됨
  - 안전한 학습 환경 조성: 무결성을 검증하면 외부 공격으로부터 AI 모델의 학습 환경을 보호할 수 있음
  - 사고 대응 용이성 제공: 무결성 검증 기록은 보안 사고 발생 시 문제를 빠르게 식별하고 대응하는 데 도움을 줌
• (예시) 데이터 전송 시 무결성 검증 방법
  - 메시지 인증 코드(MAC):
    ▸ 데이터를 전송할 때 메시지 인증 코드를 함께 전송
    ▸ 수신 측에서 재계산한 MAC과 비교하여 데이터 무결성 검증
    ▸ HMAC(Hash-based Message Authentication Code): 암호화 키와 해시를 결합하여 안전한 인증 코드 생성
  - 데이터 패킷의 체크섬
    ▸ 데이터 전송 시 각 패킷에 체크섬을 포함하여 전송
    ▸ 수신 측에서 체크섬을 계산하여 무결성 확인
    ▸ TCP/UDP 프로토콜의 기본 기능으로 지원

2.2.2 데이터에 접근할 수 있는 권한을 제한하고 있는가?

YES ☐ NO ☐ N/A ☐

- 데이터 수집 및 처리 과정에서 접근 제어가 제대로 설정되지 않으면, 권한이 없는 사용자가 기밀 데이터에 접근할 수 있음
- 권한 없는 접근은 데이터 유출뿐만 아니라 데이터 무결성을 해치거나 악의적으로 수정할 위험을 초래할 수 있음
- 내부 위협 방지
  - 내부자 위험 관리: 내부 직원이나 협력자가 데이터를 무단으로 접근하거나 외부로 유출할 가능성이 있음. 권한 제한은 내부 위협을 줄이는 데 필수적임
  - 역할 기반 접근 제어(RBAC): 필요한 작업에만 접근 권한을 부여하면, 데이터 남용과 실수를 예방할 수 있음
- 보안 사고의 범위 최소화
  - 사고 시 영향 제한: 데이터 접근이 제한되면, 보안 사고 발생 시 피해 범위를 줄일 수 있음. 권한이 없는 사람은 데이터에 접근할 수 없으므로 유출 위험이 감소함
  - 접근 로그 및 추적성 확보: 제한된 권한으로 접근이 이루어질 경우, 어떤 사용자가 데이터를 조회했는지 명확히 파악할 수 있어 사고 발생 시 원인 분석이 용이함
- 과도한 데이터 접근으로 인한 리소스 낭비
  - 불필요한 데이터 활용 방지: 권한이 제한되지 않으면, AI 개발자나 팀원이 불필요한 데이터에 접근하여 리소스를 낭비할 가능성이 있음
  - 효율적인 데이터 관리: 필요하지 않은 데이터에 대한 접근을 제한하면, 데이터 관리를 효율적으로 수행할 수 있음
- (예시) 역할 기반 접근 제어 (Role-Based Access Control, RBAC)
  - 사용자에게 특정 역할(Role)을 할당하고, 역할에 따라 데이터 접근 권한을 제한
  - 예) 관리자: 모든 데이터에 접근 가능, 데이터 분석가: 비식별화된 데이터만 접근 가능, 일반 사용자: 제한된 데이터에만 접근 가능
- (예시) 원칙 기반 접근 제어 (Policy-Based Access Control, PBAC)
  - 접근 정책을 정의하여 조건에 따라 데이터 접근 허용 여부를 결정
  - 예) 특정 시간대에만 접근 허용, 회사 내부 네트워크에서만 접근 가능
- (예시) 다중 인증(Multi-Factor Authentication, MFA)
  - 데이터 접근 시 추가적인 인증 단계를 요구
  - MFA 요소:
    - ▸무엇을 알고 있는가: 비밀번호, PIN
    - ▸무엇을 소유하고 있는가: OTP, 인증 앱
    - ▸무엇인가: 생체 인식(지문, 얼굴 인식)
  - MFA를 IAM 시스템과 통합하여 강화된 보안을 제공

2.3 데이터 공격에 대한 방어

AI 개발자, AI 서비스 제공자 공통사항

- 인공지능 서비스 개발 또는 운영 과정에서 의도적으로 학습 데이터를 변질시키거나 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 공격에 노출될 수 있으므로, 이를 대처할 방안을 검토 및 적용하는 것이 바람직함

2.3.1 데이터 오염(poisoning) 공격에 대한 방어 대책을 마련하고 있는가?

YES ☐ NO ☐ N/A ☐

- 적대적 공격을 방어하고 AI 서비스의 강건성을 높이기 위한 다양한 방어 기법이 존재함. 특히 데이터 수집 및 준비 단계에서의 오염 공격 방어를 위한 대표적 기법으로는 적대적 학습(adversarial training), Gradient Masking, Feature Squeezing 등이 있음
- (예시) 데이터 오염(Poisoning) 공격에 대한 방어 대책
  ❶ 데이터 수준 방어
  - 데이터 검증 및 정제: 데이터 수집 및 학습 전 단계에서 품질 검증 및 노이즈 제거
  - 이상치 탐지 기술을 활용해 의심스러운 데이터를 식별(예: Isolation Forest, PCA)
  - 신뢰할 수 있는 데이터 소스: 신뢰할 수 있는 출처에서만 데이터를 수집하여 악의적인 데이터 유입 방지
  - 데이터 감사: 정기적으로 데이터셋을 검토하여 의도적으로 삽입된 비정상 데이터를 탐지

  ❷ 모델 수준 방어
  - 로버스트 학습 알고리즘: 적대적 공격에 견고한 알고리즘을 사용(예: Krum, Bulyan 등 집계 전략)
  - 안정적 훈련(Stable Training): 데이터 가중치를 재조정하여 이상치가 모델 학습에 미치는 영향을 완화

적대적 공격에 대한 방어 기법

방어기법	방어기법 내용
적대적 학습 (adversarial training)	모델을 학습시킬 때, 적대적 사례로 활용할 수 있는 모든 경우의 수를 미리 고려하여 학습 데이터셋에 포함시키는 방법. 충분한 수와 다양성이 보장된 적대적 데이터를 생성하는 과정 없이는 그 성능을 보장할 수 없음
Gradient Masking (Distillation)	대부분의 공격은 모델 추론 과정에서의 경사(gradient)를 보고 이루어지므로 학습 모델의 경사가 그대로 노출되는 것을 방지하거나 gradient masking, 정규화 방법 등을 통해 경사가 두드러지지 않게 하여 적대적 공격에 방어할 수 있는 방법(distillation)들이 제안됨
Feature Squeezing	본래의 학습 모델과 별도로, 주어진 입력이 적대적 사례인지 아닌지를 판단하는 학습 모델을 추가하는 방법. 그 외에 다수의 학습 모델을 조합하여 시스템을 구성하면 특정 모델에 대한 화이트박스 공격을 피할 수 있으며, 특정 모델에 적용되는 적대적 공격이 불가능해짐

2.3.2 데이터 회피(evasion) 공격에 대한 방어 대책을 마련하고 있는가? YES ☐ NO ☐ N/A ☐

- (예시) 데이터 회피(evasion) 공격에 대한 방어 대책
  - 입력 데이터 검증 및 전처리 강화
    - ▸데이터 정규화: 입력 데이터를 일정한 범위로 정규화하여 이상값을 감지
    - ▸이상 탐지 모델: 데이터 분포에서 벗어난 입력을 감지하고 차단
      예) Autoencoder, Isolation Forest 등을 사용
    - ▸입력 필터링: 입력 이미지나 데이터를 변조 탐지 알고리즘으로 사전 처리
  - 모델 예측 확률 제한
    - ▸출력 확률 경계 설정: 모델의 출력 확률이 특정 범위를 벗어나면 의심스러운 입력으로 간주
    - ▸확률 분포 점검: Softmax 출력 분포의 평탄화 정도를 분석하여 이상 여부 판단
  - 방어용 알고리즘 사용
    - ▸Defensive Distillation: 모델을 학습시키기 전에 출력 확률 분포를 부드럽게 만들어 적대적 공격의 효과를 감소, 높은 온도 매개변수를 사용하여 Softmax를 부드럽게 학습
    - ▸Randomization: 입력 데이터나 모델의 일부를 무작위화(randomize)하여 공격자가 예측하기 어렵게 구성. 예) 입력 크기 변형, 픽셀 섞기 등

2.3.3 데이터 유출·변조 공격을 방지하기 위한 방안을 마련하고 있는가? YES ☐ NO ☐ N/A ☐

- 데이터 수집 및 처리 단계에서 기밀 데이터를 보호하지 못하면 데이터 유출이 발생할 수 있음.
  - 암호화되지 않은 데이터는 네트워크 스니핑, 중간자 공격(MITM) 등에 취약하며, 이는 데이터 유출로 이어질 수 있음
  - 데이터 전송 중에 발생하는 공격으로 인해 데이터 유출이나 변조가 발생하면, AI 모델에 잘못된 데이터가 제공되거나 중요한 정보가 손실될 수 있음
- 액세스 제어
  - 역할 기반 접근 제어(RBAC): 데이터를 처리하거나 접근할 수 있는 사용자와 시스템에 권한을 제한
  - 최소 권한 원칙(Least Privilege): 사용자와 프로세스가 필요한 데이터에만 접근할 수 있도록 권한을 최소화
  - 다중 인증(MFA): 데이터를 다룰 때 추가적인 인증 단계를 요구하여 보안을 강화
- 데이터 접근 및 사용 추적
  - 로그 기록 및 모니터링: 데이터에 대한 접근 기록을 보관하고 정기적으로 모니터링하여 이상 활동 탐지
  - 데이터 사용 감사: 데이터 사용 현황을 주기적으로 검토하고, 비정상적인 패턴이나 의심스러운 행동을 분석

03 모델개발 (학습/모델링/검증)

3.1 학습/검증 환경에 대한 보안(Secure Training Environment)

- 학습·검증을 진행하는 환경이 안전하지 않을 경우 공격자가 학습 프로세스를 방해하거나 모델을 악의적으로 수정할 수 있음

3.1.1 모델 학습을 진행하는 환경이 안전하게 보안조치 되어 있는가? YES ☐ NO ☐ N/A ☐

- 모델 학습을 진행하는 환경이 안전하지 않거나, 외부에서 무단 접근이 가능할 경우 공격자는 학습 중인 모델에 접근하여 모델의 매개변수를 조작하거나 학습 데이터를 변조하여 시스템에 악성 코드를 삽입할 수 있음
- (예시) AI 모델 학습 환경을 안전하게 보안 조치하기 위한 주요 방법

❶ 물리적 보안

- 접근 제한: 모델 학습이 진행되는 서버룸이나 데이터센터에 물리적 접근 제어 시스템(예: 카드키, 생체 인증) 도입
- 감시 시스템: 학습 환경의 물리적 보안을 강화하기 위해 CCTV 및 경고 시스템 설치
- 보안 네트워크 분리: 학습 서버가 외부 네트워크와 분리되도록 설계하여 불법 접근 차단

❷ 네트워크 보안

- 방화벽(Firewall): 외부 및 내부 네트워크로부터 불법 트래픽을 차단
- VPN 사용: 학습 환경에 접근할 때 VPN을 사용하여 암호화된 통신 경로 제공
- IDS/IPS 시스템: 침입 탐지 및 방지 시스템을 설치하여 의심스러운 네트워크 활동 탐지 및 차단
- 네트워크 분리: 학습 환경을 공용 네트워크에서 분리하고, 필요한 경우 최소한의 포트를 통해 통신

3.1.2 학습 또는 검증 단계에서 악의적인 사용자가 허위 데이터를 삽입할 가능성을 차단하고 있는가?

YES ☐ NO ☐ N/A ☐

- 학습 또는 검증 단계에서 악의적인 사용자가 허위 데이터를 삽입(Fake Data Injection)할 경우, 모델이 이러한 데이터를 학습하면서 잘못된 결과를 도출할 수 있음. 이는 학습 데이터가 공개 출처일 때 더욱 취약함
- 허위 데이터로 학습된 모델은 신뢰할 수 없는 결과를 생성할 가능성이 높으며, 이는 모델 성능 저하 및 보안 취약점으로 이어짐
- 허위 데이터를 삽입하려는 시도를 차단하려면 데이터 수집, 전처리, 학습, 검증, 배포 등 모든 단계에서 종합적인 보안 조치를 적용해야 함. 신뢰할 수 있는 데이터 소스 사용, 디지털 서명 및 인증, 이상치 탐지, 권한 관리 등 기술적 조치와 함께 조직 차원의 교육과 정책적 조치가 병행되어야 함
- (예시) 악의적인 사용자가 허위 데이터를 삽입할 가능성을 차단할 수 있도록 사용자 접근 관리
  ❶ 인증 및 권한 관리
  - 다중 인증(Multi-Factor Authentication): 학습 환경 및 데이터에 접근하려면 추가 인증 단계를 요구
  - 권한 기반 접근 제어(RBAC): 데이터에 접근할 수 있는 사용자 및 애플리케이션을 제한

  ❷ 활동 모니터링
  - 사용자 로그 기록: 학습 환경에서 데이터와 상호작용하는 사용자 활동을 기록
  - 비정상 활동 탐지: 의심스러운 활동이 감지되면 즉시 차단 및 경고

3.1.3 연합 학습(Federated Learning)에 참여하는 장치 중 악의적인 장치가 있는지 검증하고 있는가?

YES ☐ NO ☐ N/A ☐

- 연합 학습(Federated Learning)은 여러 분산된 장치에서 데이터를 수집하여 학습을 진행하는 방법이며, 참여하는 장치 중 악의적인 장치가 있으면 전체 모델 학습에 악영향을 미칠 수 있음
- 악의적인 참여자가 학습 데이터나 업데이트를 조작해 전체 모델에 영향을 미치면, 연합 학습의 신뢰성이 저하되고 보안 위협이 발생할 수 있음
- (예시) 연합 학습(Federated Learning)에 참여하는 장치 중 악의적인 장치가 있는지 검증하는 방법

❶ 모델 업데이트 유효성 검증

- 비교 검증: 제출된 업데이트를 다수의 장치로부터 수집된 정상 업데이트와 비교하여 편향성이나 비정상적인 패턴 탐지
- 검증 데이터 기반 평가: 중앙 서버에서 검증 데이터셋을 사용해 각 참여자의 업데이트를 평가하고, 성능이 현저히 낮거나 의심스러운 경우 차단

❷ 신뢰 기반 접근으로 참여 장치 검증

- 디지털 인증: 참여 장치에 디지털 인증서를 부여하여 인증된 장치만 학습에 참여
- 장치 프로파일링: 장치의 과거 참여 기록과 행동 패턴을 기반으로 신뢰 점수를 부여하고 악의적 활동을 탐지

❸ 장치 격리로 참여 장치 검증

- 의심 장치 격리: 악의적 의심이 드는 장치를 임시적으로 학습 과정에서 제외하고 추가 검증
- 샌드박스 테스트: 의심 장치의 업데이트를 별도의 환경에서 테스트하여 안전성 확인

3.2 모델 공격에 대한 방어

- 인공지능 모델은 적대적 의도를 가진 사용자에 의해 학습 데이터 및 기능을 도용당하거나 다른 방식의 공격으로 인한 AI 탈옥, AI 모델 탈취 등이 발생할 수 있으므로 이를 방지 또는 완화하기 위한 대책을 수립함

3.2.1 AI Prompt Injection 공격에 대한 방어 방안을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- “Prompt Injection”은 공격자가 정교하게 설계된 입력을 통해 AI의 동작이나 출력을 조작하여 의도하지 않은 결과를 유도함. 유형으로 ▲직접 프롬프트 인젝션 ▲간접 프롬프트 인젝션, ▲다중 모달 및 언어 기반 인젝션 등이 있음
- “Prompt Injection”은 출력을 조작하여 의도하지 않은 결과 유도, AI 탈옥 또는 모델 탈취 등을 발생시킬 수 있음
- “Prompt Injection”은 AI 애플리케이션의 보안을 위협할 수 있는 가장 흔한 공격으로, 다양한 애플리케이션에서 발생 가능하며, 방어가 까다롭기 때문에 지속적인 모니터링과 보안 대책이 필수

“Prompt Injection”에 대한 예방책

방어 기법	세부 내용
모델 동작 제한	모델의 역할, 기능, 한계에 대한 구체적인 지침을 설정. 응답을 특정 작업이나 주제로 제한하며, 핵심 지침을 수정하려는 시도를 무시하도록 지시
예상 출력 형식 정의 및 검증	명확한 출력 형식을 지정하고 상세한 근거와 출처를 요구
입력 및 출력 필터링 구현	민감한 내용 감지 및 차단을 위한 규칙을 설정. RAG Triad를 활용해 ▲컨텍스트 관련성, ▲정보 근거성, ▲질문/응답 관련성을 평가하며 잠재적으로 악성일 수 있는 출력을 식별
권한 제어 및 최소 권한 접근 시행	애플리케이션에 자체 API 토큰을 제공하여 확장 가능한 기능을 구현하고 모델 대신 코드에서 처리되도록 설계. 모델의 접근 권한을 필요 최소한의 수준으로 제한
고위험 작업에 대한 사용자 승인 요구	민감한 작업(예: 고급 권한이 필요한 작업)을 위한 사용자 승인 절차(human-in-the-loop controls)를 구현하여 비인가된 행동을 방지
외부 콘텐츠 분리 및 식별	신뢰할 수 없는 콘텐츠를 분리하고 이를 명확히 표시하여 사용자 프롬프트에 미치는 영향을 제한
적대적 테스트 및 공격 시뮬레이션 수행	모델을 신뢰할 수 없는 사용자로 취급하여 침투 테스트와 침해 시뮬레이션을 정기적으로 수행하여 신뢰 경계 및 접근 제어의 효과를 평가

3.2.2 적대적 예제 공격(Adversarial Example Attacks)에 대한 방어 방안을 수립하고 있는가? YES ☐ NO ☐ N/A ☐

- 모델 학습이나 검증 과정에서 적대적 공격을 고려하지 않으면, 공격자가 고의적으로 생성한 적대적 예제(Adversarial Examples)에 의해 모델이 잘못된 결정을 내리게 할 수 있음. 적대적 예제는 인간이 식별하기 어려운 미세한 변화를 통해 모델을 속임
- 적대적 공격으로 인해 AI 모델이 이미지 분류, 음성 인식, 자율주행 등에서 오작동을 일으킬 수 있으며, 이는 안전성과 직결된 문제를 야기할 수 있음
- (예시) 적대적 예제 공격(Adversarial Example Attacks) 방어 방법: 적대적 훈련(Adversarial Training)
  - ❶ 적대적 샘플 포함 학습
  - 적대적 샘플 생성 및 추가: 학습 데이터에 적대적 샘플을 포함시켜 모델이 이러한 입력에 대해 견고하게 학습
  - 공격 기법에 따라 다양화: FGSM(Fast Gradient Sign Method), PGD(Projected Gradient Descent)와 같은 다양한 공격 기법으로 생성된 샘플을 포함
  - ❷ 정기적 업데이트
  - 새로운 공격 기법에 대해 정기적으로 모델을 재훈련하여 적대적 샘플 방어 능력 강화
- (예시) 적대적 예제 공격(Adversarial Example Attacks) 방어 방법: 모델 설계 및 학습 과정 개선
  - ❶ 모델 견고성 강화
  - Dropout 및 정규화: 학습 과정에서 Dropout과 정규화를 사용해 과적합 방지
  - Gradient Masking: 모델이 내부 경사 정보를 공격자에게 노출하지 않도록 설계
  - ❷ Robust Optimization
  - 학습 과정에서 적대적 샘플을 포함해 최적화 과정을 강화하는 알고리즘 사용
  - ❸ 다중 모델 앙상블
  - 여러 개의 모델을 조합해 하나의 예측값을 생성하여 특정 모델 공격의 성공률 감소

3.2.3 모델 회피 공격(model evasion attack)에 대한 방어 방안을 수립하고 있는가? YES ☐ NO ☐ N/A ☐

- 모델 회피 공격은 입력 데이터에 최소한의 변조를 가해 인공지능 모델을 속이는 기법임. 특히 이미지 메인은 약간의 변화가 발생해도 사람의 눈에 잘 띄지 않으므로 적대적 공격(adversarial attack)에 취약함
- 모델 회피 공격에 대한 방어는 다중 방어 전략을 통해 이루어져야 함. 적대적 훈련, 입력 전처리, 모델 앙상블과 같은 기술적 방어뿐만 아니라, 지속적 평가와 커뮤니티 협업을 통해 AI 모델의 강건성/안전성을 강화하는 것이 필수임
- (예시) 적대적 훈련 (Adversarial Training)
  - 개념: 공격자가 사용할 수 있는 적대적 예제(Adversarial Examples)를 모델 학습 데이터에 포함시켜 훈련
  - 방법: 적대적 샘플을 생성하여, 원본 데이터와 적대적 데이터 모두로 모델을 훈련
  - 효과: 모델이 공격에 노출되었을 때도 강건성(Robustness)이 향상됨
- (예시) 탐지 및 모니터링 시스템 구축
  - 개념: 입력이 정상적인 데이터 분포에서 벗어나면 이를 탐지해 방어
  - 방법:
    - ▸ Anomaly Detection: 이상 데이터 탐지 기법을 활용
    - ▸ Input Certification: 입력 데이터가 허용 가능한 분포 내에 있는지 확인
  - 효과: 회피 공격을 조기에 탐지하고 대응할 수 있음
- (예시) 방어적 디코딩 (Defensive Decoding)
  - 개념: 입력 데이터를 처리하기 전에 추가적인 필터링과 디코딩을 수행해 모델의 취약점을 차단
  - 방법:
    - ▸ 이미지 데이터에 대해 복원 네트워크를 사용하여 노이즈를 제거
    - ▸ 텍스트 데이터의 경우 문장 구조를 재확인하고 오류를 수정
  - 효과: 모델에 전달되는 데이터의 안전성 향상

3.2.4 모델 오염 공격(Model Poisoning Attack)에 대한 방어 방안을 수립하고 있는가? YES ☐ NO ☐ N/A ☐

- 공격자가 훈련 데이터를 조작하여 AI 모델이 악의적인 방향으로 학습하도록 유도하는 공격임. 주로 훈련 데이터에 악성 데이터를 삽입해 모델이 왜곡된 학습을 하게 함
- 모델이 공격자의 의도에 따라 편향된 결과를 도출할 수 있으며, 정상적인 상황에서 예측 오류가 증가할 수 있음. 이는 AI 시스템의 신뢰성을 저하시키고, 악의적인 목적에 사용될 위험이 있음
- (예시) 모델 오염 공격(Model Poisoning Attack) 방어 방법: 데이터 수준에서의 방어

  ❶ 데이터 검증 및 정제

  - 데이터 무결성 검사: 데이터 수집 단계에서 해시 값을 비교하여 데이터 무결성을 검증
  - 이상치 탐지: 학습 데이터 내에서 비정상적인 패턴을 보이는 데이터를 식별
    예) Isolation Forest, PCA 기반 탐지
  - 샘플링 검토: 학습 데이터의 하위 샘플링을 무작위로 점검하여 의도적으로 조작된 데이터를 제거

  ❷ 데이터 필터링

  - 노이즈 필터링: 데이터에 노이즈를 제거하거나 정규화를 수행하여 포이즈닝 효과를 완화

  ❸ 신뢰할 수 있는 데이터 소스 사용

  - 데이터 수집 시 신뢰할 수 있는 출처에서 데이터를 획득하며, 외부 데이터 검증 프로세스 도입

3.2.5 모델 추출 공격(model extraction attack) 및 리버스 엔지니어링에 대한 방어 방안을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 예측 결과를 분석하고 분류 기준을 추출하여 서비스 중인 학습 모델과 유사한 성능의 대체 모델을 구성하는 방법과 학습된 모델의 입력 데이터, 모델의 초매개변수(hyperparameter) 정보, 계층 구조 등을 추출하는 공격 방식이 존재함
- 인공지능 모델 공격에 대한 주요 완화 방법에는 특정 시간 간격당 인공지능 서비스에 대한 질의 수를 제한, 의심스러운 질의에 대한 탐지 및 경고, 예측 결과의 난독화(obfuscation) 등이 있음

인공지능 모델 추출 공격에 대한 방어 기법 (예시)

방어 기법 분류	방어 기법 내용
질의 횟수 제한	특정 기간 내에 수행할 수 있는 질의 수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하는 기법
학습기반 모니터링	기계학습을 활용하여 모델 공격에 대해 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등 능동적으로 방어하는 기법
결과 난독화	예측 결과가 결정경계에 가까운 경우 예측 결과의 정확도를 임의로 낮춰 모델의 세부 속성에 대한 추출을 방해하는 기법

- (예시) 리버스 엔지니어링 방어 방법: 모델 출력 제어
  - ❶ 출력 제한
  - 결과 일반화: 모델 출력값(예: 확률 분포)을 과도하게 자세히 제공하지 않고, 최상위 클래스나 간략화된 정보를 제공
  - 탑-k 제한: 모델이 반환하는 결과를 상위 k개의 결과로 제한하여 상세한 출력 제공 방지
  - 예: 분류 모델에서 확률값 대신 가장 높은 확률의 k개 클래스만 반환
  - ❷ 출력 노이즈 추가
  - 차등정보보호(Differential Privacy): 모델 출력에 소량의 노이즈를 추가하여 원본 모델 구조와 매개변수를 추론하기 어렵게 제공
  - 적응형 노이즈: 특정 질의 패턴에서만 노이즈를 추가해 공격자를 혼란시킴
  - ❸ 출력 왜곡
  - 랜덤화: 출력값의 순서나 범위를 무작위로 변경하여 제공함으로써 탈취 시도 무력화
  - 결과 변형: 특정 질의 패턴에서는 변형된 결과를 제공

3.2.6 반복적인 질의 공격(Repetitive Query Attack)에 대한 방어 방안을 수립하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 모델 자체는 지적 재산으로서 보호가 필요하며, 적대적 공격이나 모델 탈취로부터 안전해야 함. 특히, Model Denial of Service는 공격자가 LLM과 상호작용할 때 예외적으로 많은 양의 리소스를 소모할 때 발생함. 이는 다른 사용자의 서비스 품질을 저하시킬 수 있으며, 잠재적으로 높은 리소스 비용이 발생할 수 있음
- 위와 같은 모델 공격에 대해서는 특정 기간 내에 개별 사용자 또는 특정 IP가 수행할 수 있는 질의 수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하는 방법이 있음
- (예시) 질의 제한 및 속도 조절
  - ❶ 요청 속도 제한 (Rate Limiting)
  - – 일정 시간 내 허용되는 질의 수를 제한하여 반복적이고 과도한 요청을 방지
  - ❷ 요청 빈도 감소 (Throttling)
  - – 동일한 사용자가 짧은 시간 내 다수의 요청을 보내는 경우 응답을 지연시켜 공격의 효율성 감소
  - ❸ 동적 속도 제한
  - – 사용자의 행동 패턴에 따라 속도 제한 정책을 동적으로 조정

3.2.7 기계 학습을 활용한 모델 공격에 대해 능동적으로 방어하고 있는가? YES ☐ NO ☐ N/A ☐

- 기계 학습을 활용한 모델 공격에 대해서는 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등 능동적으로 방어할 필요가 있음
  - – DoS 공격을 나타낼 수 있는 비정상적인 급증이나 패턴을 식별하기 위해 LLM의 리소스 사용률을 지속적으로 모니터링함
  - – 요청 또는 단계당 리소스 사용량을 제한하여 복잡한 부분을 포함하는 요청을 더 느리게 실행
- (예시) 기계 학습을 활용한 모델 공격 방어 방법: 공격 탐지 및 모니터링
  - ❶ 이상 탐지 기법
  - – ML 기반 탐지: 요청 데이터와 사용 패턴을 학습하여 정상 사용자와 공격자의 행동 차이를 식별. Isolation Forest, DBSCAN, Autoencoder 등을 활용해 비정상적인 요청 탐지
  - – 행동 패턴 분석: 반복적이고 대량의 요청, 유사한 질의 패턴을 분석하여 의심스러운 활동을 탐지
  - ❷ 실시간 로그 분석
  - – 질의 기록 분석: API 요청 및 응답 로그를 저장하고 의심스러운 패턴을 실시간으로 탐지
  - – 알고리즘 변칙 탐지: 요청 데이터가 모델에 미치는 영향을 분석하여 비정상적인 변화를 탐지
  - ❸ 경고 시스템
  - – 비정상적인 요청이 감지되면 관리자에게 즉시 알림을 보내거나 자동으로 차단

3.3 오픈소스 라이브러리 보안

- 인공지능 모델 개발 단계에서는 다양한 오픈소스를 활용할 수 있음. 오픈소스 라이브러리 도입 전에는 필요성 및 원하는 기능의 제공 여부 등의 확인이 필요하고, 사용할 라이브러리가 안정적으로 업데이트 중인지, 주의해야 할 라이선스 기준은 무엇인지 등을 확인해야 함. 또한, 사용 중인 오픈소스의 목록 및 버전을 지속해서 확인하여 운영 및 보안상의 위험 요소를 점검해야 함

3.3.1 오픈소스 라이브러리의 업데이트 및 취약점을 관리하고 있는가? YES ☐ NO ☐ N/A ☐

- 인공지능 모델 개발에 오픈소스 라이브러리를 사용한다면, 안정성 확인을 위해 해당 오픈소스 라이브러리가 얼마나 많은 사용자를 보유하는지, 업데이트는 자주 이루어지는지, 이슈가 발생했을 때 대응은 신속하게 이루어지는지 등을 따져봐야 함
- (예시) 오픈소스 라이브러리 안전성 확인 방법: 라이브러리의 출처와 신뢰성 검증
  ❶ 공식 소스 확인
  - 공식 문서 확인: 라이브러리의 공식 웹사이트와 문서를 검토하여 신뢰성 확보.
  ❷ 개발 커뮤니티의 활동 확인
  - 이슈 대응: 버그나 보안 문제에 대한 커뮤니티의 대응 속도와 품질 평가
- (예시) 오픈소스 라이브러리 위험요소 관리 방법: 정기적인 보안 패치
  ❶ 업데이트 및 패치
  - 라이브러리의 최신 보안 패치를 정기적으로 적용. 변경 로그(Changelog)를 검토하여 보안 개선 사항을 확인
  ❷ 자동화된 업데이트 도구 활용
  - Dependabot: GitHub 리포지토리에서 취약점이 발견된 의존성을 자동으로 업데이트
  - Renovate: 여러 프로젝트에서 의존성을 자동으로 업데이트 및 관리

3.3.2 오픈소스 라이브러리의 소스 코드를 직접 검토하거나 사용에 대한 보안 문제를 검증하고 있는가? YES ☐ NO ☐ N/A ☐

- 오픈소스 라이브러리의 소스 코드를 직접 검토하고 사용에 대한 보안 문제를 검증하는 것은 보안 취약점 예방, 성능 최적화, 법적 컴플라이언스 준수, 신뢰성 확보를 위한 필수적인 과정임. 이를 통해 AI 개발 프로젝트의 안정성과 신뢰성을 높이고, 잠재적인 위험을 사전에 방지할 수 있음
- 오픈소스 라이브러리의 종류 및 버전 선택 시 개발 과정에서 보안 취약점이 발견될 수 있으므로 이러한 이슈들을 확인하여 보안상의 위험 요소에 대한 관리가 필요함
- (예시) 오픈소스 라이브러리 위험요소 관리 방법: 코드 분석 및 검토
  - ❶ 정적 코드 분석: 라이브러리의 소스 코드를 분석하여 보안 취약점을 탐지
  - ❷ 동적 분석: 라이브러리를 샌드박스 환경에서 실행하여 악성 코드가 포함되었는지 확인
  - ❸ 코드 리뷰: 오픈소스 커뮤니티에서 제공하는 코드 리뷰 내용을 검토하여 위험 요소 확인
- (예시) 오픈소스 라이브러리 안전성 확인 방법: 취약점 데이터베이스 확인
  - CVE 데이터베이스 검색: 라이브러리와 관련된 알려진 취약점이 있는지 확인
  - NVD(National Vulnerability Database): 오픈소스 라이브러리의 보안 이슈를 검색
- (예시) 오픈소스 라이브러리 사용 제한 적용
  - ❶ 최소 권한 부여
  - 라이브러리에서 사용하는 API나 파일 접근 권한을 최소화
  - Sandbox를 통해 실행 환경에서 라이브러리가 시스템 자원에 접근하지 못하도록 제한
  - ❷ 민감 데이터 처리 방지
  - 오픈소스 라이브러리를 사용할 때 민감 데이터를 직접 처리하지 않도록 설계

3.3.3 오픈소스 라이브러리를 실행할 때 잠재적인 보안 위험을 제거하기 위해 격리된 환경을 이용하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 모델 개발 시 오픈소스 라이브러리를 실행할 때 격리된 환경을 이용하는 이유는 보안, 안정성, 호환성, 및 효율성을 보장하기 위해서임. 이를 통해 악성 코드 실행, 시스템 변경, 중요 데이터 유출 등으로부터 시스템을 보호하고, 라이브러리의 영향 범위를 제한할 수 있음
- (예시) 오픈소스 라이브러리 위험요소 관리 방법: 환경에서의 안전한 실행
  - ❶ 샌드박스 환경에서 테스트
  - 라이브러리를 프로덕션 환경에 배포하기 전에 격리된 테스트 환경에서 실행
  - 테스트 데이터와 시뮬레이션 환경을 사용하여 예상치 못한 동작 방지
  - ❷ 최소 권한 실행
  - 라이브러리가 접근할 수 있는 시스템 자원과 데이터를 최소화
  - 원격 API 호출이나 외부 네트워크 액세스를 제한
  - ❸ 컨테이너화
  - Docker, Kubernetes 같은 컨테이너 기술을 활용해 라이브러리 실행 환경을 격리

3.4 LLM 보안

- LLM 애플리케이션 보안 영역에 대한 실용적이고 실행가능한 지침을 제공하고자 함
- 일반적인 애플리케이션 보안 원칙과 LLM이 제기하는 특정 보안 취약성과의 간극을 메우는 것이 필요하며, 기존 취약성이 어떻게 LLM 내에서 다른 위험을 초래하거나 새로운 방식으로 악용될 수 있는지, 그리고 개발자가 LLM을 활용하는 애플리케이션에 대해 기존 수정 전략을 어떻게 적용해야 하는지에 방향성을 제공함

3.4.1 LLM 애플리케이션 공격에 대한 예방책을 마련하고 있는가?

YES ☐ NO ☐ N/A ☐

- LLM 애플리케이션은 실수로 민감 정보 또는 기밀 데이터를 외부에 공개할 수 있음
- LLM 애플리케이션 공격에 대한 예방책은 다음과 같음
  - 적절한 데이터 Sanitization 및 스크러빙 기술을 통합하여 사용자 데이터가 학습 모델 데이터에 입력되는 것을 방지
  - 강력한 입력 검증 및 Sanitization 방법을 구현하여 잠재적인 악성 입력을 식별하고 필터링하여 모델이 오염되는 것을 방지
  - 외부 데이터 소스에 대한 액세스(런타임 시 데이터 오케스트레이션)는 제한되어야 함. 외부 데이터 소스에 대한 엄격한 액세스 제어 방법과 안전한 공급망을 유지하기 위한 엄격한 접근 방식을 적용

3.4.2 LLM의 모델 서비스 거부(Model Denial of Service) 공격에 대한 방어 방안을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- “Model Denial of Service”는 공격자가 LLM(Large Language Model)과 상호작용할 때 예외적으로 많은 양의 리소스를 소모할 때 발생함. 이는 공격자와 다른 사용자에게 서비스 품질이 저하될 수 있으며, 잠재적으로 높은 리소스 비용이 발생할 수 있음

  ※ 예시: ①대량 대기열(리소스 집약적 작업), ②리소스 소모 쿼리(비정상적인 쿼리), ③오버플로(overflow, 지속적인 과도한 입력), ④ 반복적인 긴 입력(리소스 고갈)

“Model Denial of Service”에 대한 예방책

방어 기법	세부 내용
입력 검증	정의된 한도를 준수하고 악성 콘텐츠를 걸러내기 위해 사용자 입력에 대한 입력 검증 및 정리(sanitization)를 구현
리소스 캡	요청 또는 단계당 리소스 사용량을 제한하여 복잡한 부분을 포함하는 요청이 더 느리게 실행되도록 함
API 비율 제한	특정 기간 내에 개별 사용자 또는 IP 주소가 수행할 수 있는 요청 수를 제한하기 위해 API 속도 제한 적용
큐 관리	LLM 응답에 반응하는 시스템에서 대기 중인 작업 수와 총 작업 수를 제한
리소스 모니터링	D_oS 공격을 나타낼 수 있는 비정상적인 급증이나 패턴을 식별하기 위해 LLM의 리소스 사용률을 지속적으로 모니터링함

3.4.3 LLM의 API 보안을 위한 방안을 수립하고 있는가? YES ☐ NO ☐ N/A ☐

- API는 특정 시스템에서 프로그램이 작동하도록 간편화한 인터페이스임.
- 해커는 기업이나 생성형 AI 제조사의 일반 시스템에 접근하는 대신 LLM 접목을 위해 사용하는 API의 취약점을 노리기 때문에 LLM(대규모 언어 모델)을 도입하는 기업이 단계에 맞는 API 보안 정책을 세워야 함.
  - 주요 단계는 ▲설계 및 개발 ▲교육 및 테스트 ▲배포 ▲운영 및 모니터링 ▲유지 및 업데이트 등으로 구분됨
- LLM(대규모 언어 모델) API의 보안을 보장하기 위해서는 접근 제한, 데이터 보호, 인증 및 권한 부여, 실시간 모니터링과 같은 다각적인 보안 대책이 필요함
- (예시) 인증(Authentication) 및 권한 부여(Authorization)
  ❶ API 키 관리: 각 사용자에게 고유한 API 키를 제공하여 접근 권한을 제한
  ❷ OAuth 2.0 및 OpenID Connect: 사용자 인증 및 권한 부여를 위한 업계 표준 프로토콜 사용
  ❸ 다중 인증(Multi-Factor Authentication, MFA): API 접근 시 추가 인증 단계 요구
- (예시) 사용량 제한 및 모니터링
  ❶ Rate Limiting: 사용자당 요청 수를 제한하여 과도한 호출로 인한 서비스 중단 방지
  ❷ 실시간 모니터링: API 호출 로그를 실시간으로 모니터링하여 비정상적인 활동 탐지
  ❸ 악의적인 사용 탐지: 비정상적인 패턴(예: DDoS 공격, 스크래핑) 탐지 및 차단
- (예시) API 응답 보안
  ❶ Rate Limiting 초과 응답: 초과 요청에 대해 표준 HTTP 상태 코드(429 Too Many Requests) 반환
  ❷ 민감 데이터 마스킹: API 응답에서 민감 데이터를 마스킹하거나 제거
  ❸ 에러 메시지 관리: 에러 메시지를 통해 민감한 시스템 정보(예: 디버그 정보, 경로 등)가 노출되지 않도록 제한

3.4.4 LLM의 인터페이스 공격에 대한 예방책을 마련하고 있는가?

YES ☐ NO ☐ N/A ☐

- LLM(대규모 언어 모델)의 인터페이스 공격은 공격자가 모델의 입력 또는 출력 인터페이스를 악용하여 비정상적인 동작을 유도하거나 민감한 데이터를 유출시키는 공격을 의미함
- (예시) LLM의 인터페이스 공격에 대한 예방책: 입력 검증 및 필터링
  - ❶ 입력 유효성 검증: 입력 데이터를 정규화하고, 예상치 못한 형식이나 길이를 가진 입력을 차단. 허용 가능한 입력 유형과 범위를 명확히 정의(예: SQL 인젝션, 코드 실행 명령 등이 포함된 텍스트 필터링)
  - ❷ 금지어 및 패턴 필터링: 모델이 민감한 정보나 금지된 데이터를 반환하지 않도록 특정 키워드 또는 패턴을 필터링(예: “비밀번호”, “크레디트 카드 번호” 등 민감 데이터 차단)
  - ❸ 입력 크기 제한: 과도한 크기의 입력(예: 너무 긴 텍스트)이 시스템에 부하를 주지 않도록 제한. 예: 최대 입력 토큰 수 설정
- (예시) LLM의 인터페이스 공격에 대한 예방책: 출력 검증 및 제어
  - ❶ 민감 데이터 필터링: 출력에 포함된 민감한 데이터나 비정상적인 정보를 사전에 검출 및 차단. 예: 사용자 인증 정보, 내부 시스템 구조 관련 정보 등
  - ❷ 모델 출력 제한: 출력 길이 제한을 설정하여 과도한 데이터 반환을 방지. 모델 출력의 특정 토큰 또는 문맥 조건을 검토.
  - ❸ 결과 샌드박싱: 모델 출력을 사용하기 전에 별도의 검증 계층에서 안전성을 평가

3.4.5 개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- 소프트웨어 개발팀이 LLM 시스템을 사용하여 코딩 작업을 빠르게 할 수 있음. 이 때 AI의 제안에 지나치게 의존하면 안전하지 않은 기본 설정이나 안전한 코딩 관행과 일치하지 않는 권장 사항으로 인해 애플리케이션에 보안 취약성이 발생할 수 있음
- 지나친 의존 때문에 나타나는 취약점을 사전에 예방할 수 있는 방법은 개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립해야함
- (예시) 안전한 코딩 관행의 기본 원칙 수립
  - ❶ 최소 권한 원칙 (Principle of Least Privilege): 모델, 데이터, 코드, API가 필요한 최소한의 권한만 가질 수 있도록 설계. 예: 모델이 파일 시스템이나 네트워크에 불필요하게 접근하지 않도록 제한
  - ❷ 데이터 최소화: LLM이 처리해야 할 데이터의 범위를 최소화하여 불필요한 민감 데이터 노출 방지. 훈련 데이터와 입력 데이터에서 개인정보(PII) 제거
- (예시) 안전한 입력/출력 처리
  - ❶ 입력 검증 및 필터링: 사용자 입력의 길이, 형식, 값 범위를 검증하여 악의적인 입력을 차단. 정규표현식 또는 화이트리스트 방식으로 유효성을 검사
  - ❷ 출력 제어: LLM 출력이 민감 데이터나 예상치 못한 정보를 포함하지 않도록 필터링. 모델이 생성한 응답을 후처리(post-processing)하여 안전성 검토
  - ❸ 입력 데이터와 명령어 분리: 사용자 입력 데이터를 명령어나 코드와 분리하여 코드 삽입 공격 방지

3.4.6 LLM 출력결과를 정기적으로 모니터링하고 검토하고 있는가?

YES ☐ NO ☐ N/A ☐

- LLM 출력에 대해서는 정기적으로 자기 일관성(Self Consistency on prompt) 투표 기술을 사용하여 일관되지 않은 텍스트를 필터링함. 단일 프롬프트에 대한 여러 모델 응답을 비교하면 출력의 품질과 일관성을 더 잘 판단할 수 있음
- (예시) LLM 출력결과 모니터링시 확인사항: 출력 데이터 수집 및 로그 기록
  - ❶ 출력 데이터 로깅: 모든 요청과 응답 데이터를 로그로 기록. 요청, 응답, 사용자 ID, 타임스탬프 등 메타데이터 포함
  - ❷ 로그 저장소 관리: 로그 데이터를 암호화하여 저장. 데이터베이스 또는 로그 관리 도구 사용 (Splunk, ELK Stack)
- (예시) LLM 출력결과 모니터링시 확인사항: 이상 탐지 및 자동화
  - ❶ 이상 출력 탐지: AI 기반 이상 탐지 도구를 활용하여 비정상적인 출력 탐지
  - ❷ 규칙 기반 필터링: 출력 데이터에 대한 규칙 기반 점검
  - ❸ 실시간 알림: 비정상 출력이 발생하면 알림을 보내는 시스템 구축

3.4.8 LLM의 벡터 및 임베딩 취약점에 대한 방어 방안을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- RAG(Retrieval Augmented Generation) 및 임베딩 기반 시스템은 정보를 검색하고 LLM 출력의 정확성을 높이기 위해 사용됨. 그러나 공격자들은 이를 악용하여 유해한 콘텐츠 삽입, 모델 동작을 조작, 민감한 정보를 노출시킬 수 있음.
- RAG(Retrieval Augmented Generation)는 사전 학습된 언어 모델과 외부 지식 소스를 결합하여 LLM 애플리케이션의 성능과 응답의 맥락 관련성을 향상시키는 모델 적응 기술

"벡터 및 임베딩 취약점"에 대한 예방책

방어 기법	세부 내용
권한 및 접근 제어	세분화된 접근 제어 및 권한 관리가 가능한 벡터 및 임베딩 저장소를 구현
데이터 검증 및 출처 인증	강력한 데이터 검증 파이프 라인을 구현하여 신뢰할 수 있고, 검증된 소스의 데이터만 허용
데이터 결합 및 분류 검토	서로 다른 소스의 데이터를 결합할 때, 결합된 데이터셋을 철저히 검토. 지식 데이터베이스 내 데이터를 태그하고 분류하여 접근 수준을 제어하고 데이터 불일치 오류를 방지
모니터링 및 로깅	의심스러운 동작을 신속히 탐지하고 대응하기 위해, 검색 활동에 대한 세부적인 불변 로그(immutable log)를 유지하여 기록의 무결성을 확보

04 모델 배포

4.1 모델파일 및 배포 환경 보호

- 모델 파일이나 배포 환경이 악의적으로 변경되면, 출력 결과가 왜곡되어 잘못된 결정을 초래할 수 있고, 또한 공격자가 모델에 백도어를 삽입하여, 특정 입력에 대해 예상치 못한 출력을 유도하거나 민감 데이터를 유출할 수 있음

4.1.1 모델을 배포하기 전에 코드 및 모델을 스캔하고, 자동화된 취약점 분석을 하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 모델을 배포하기 전에 코드 및 모델을 스캔하고 자동화된 취약점 분석을 수행하기 위해서는 체계적인 절차와 도구를 활용해야 함. 이 작업은 모델의 안전성과 신뢰성을 보장하고, 보안 위협을 사전에 방지하는 데 중요함
- (예시) 자동화된 취약점 분석 프로세스
  - ❶ 정적 코드 분석 (Static Analysis): 소스 코드를 실행하지 않고 코드 내부의 잠재적 취약점을 탐지
    - ▸작업 내용: 보안 결함, 코드 품질 문제, 잘못된 API 호출 등을 식별
  - ❷ 동적 코드 분석 (Dynamic Analysis): 실행 중인 코드의 보안 취약점을 탐지
    - ▸작업 내용: 런타임에서 발생할 수 있는 취약점(예: 메모리 누수, 비정상 동작)을 식별
  - ❸ 모델 수준의 취약점 분석:
    - ▸적대적 공격 시뮬레이션: 적대적 입력(Adversarial Input)에 대해 모델의 반응을 테스트
- (예시) 종속성 및 라이브러리 취약점 분석
  - ❶ 종속성 관리: 모델이 의존하는 오픈소스 라이브러리와 종속성의 취약점 확인
  - ❷ 최신 상태 유지: 최신 보안 패치가 적용된 라이브러리와 소프트웨어 사용
- (예시) 배포 전에 MCP(Model Context Protocol) 보안 점검
  - MCP를 통해 어떤 도구/기능이 호출될 수 있는지 식별
  - 각 MCP 툴의 신뢰성 및 보안 검토
  - 비정상 컨텍스트 요청 시 응답 행동 검증
  - 세션 간 정보 누출 여부 확인
  - MCP 호출 흐름에 대한 감사 로깅 확인
  - MCP 툴 권한 설정 및 호출 패턴의 이상탐지 적용

※ MCP(Model Context Protocol)는 AI 모델이 외부 애플리케이션 또는 사용자와 상호작용할 때 컨텍스트 정보를 안전하고 구조화된 방식으로 주고받기 위한 프로토콜로, OpenAI 등에서 도입한 개념임. 예를 들어, 사용자 명령어를 보조 앱으로 전달하거나, LLM이 플러그인, API, 툴 등과 연동될 때 사용됨. 하지만 이러한 구조는 보안상 여러 위협에 노출될 수 있으므로 AI 모델 개발자는 MCP 관련 보안 점검을 수행할 필요가 있음

4.1.2 모델파일을 암호화하여 저장하고 전송 중에도 안전하게 보호하고 있는가? YES ☐ NO ☐ N/A ☐

- 모델 파일의 암호화 및 전송 보안을 보장하려면 저장 암호화, 전송 암호화, 접근 제어, 무결성 검증과 같은 다각적인 보안 조치를 적용해야 함. 이 과정을 자동화하고 정기적으로 검토함으로써 AI 모델의 안전성을 유지하고 잠재적 위협을 효과적으로 방지할 수 있음
- (예시) 모델 파일 접근 제어
  - ❶ 접근 제어 정책: 모델 파일에 대한 접근 권한을 최소 권한 원칙에 따라 설정. 읽기/쓰기 권한을 엄격히 제한
  - ❷ 다중 인증(MFA): 모델 파일에 접근하거나 다운로드하려면 비밀번호 외에 OTP, 인증 앱 등 추가 인증 절차 요구
  - ❸ 로그 기록: 파일 접근 로그를 기록하고 비정상적인 접근 시 경고 알림
- (예시) 배포 환경 보호
  - ❶ 컨테이너 이미지 암호화: Docker 이미지 내부의 모델 파일을 암호화하거나 비공개 레지스트리 사용
  - ❷ 런타임 암호화: 하드웨어 기반 보안을 사용하여 모델 파일이 실행 중에도 암호화 상태로 유지. Confidential Computing 기술 활용 등

4.1.3 AI 모델이 배포되는 인프라(클라우드, 서버 등) 환경이 충분한 보안시스템을 갖추고 있는가? YES ☐ NO ☐ N/A ☐

- AI 모델이 배포되는 환경이 안전하지 않으면 공격자가 모델에 무단으로 접근하거나 모델을 악용할 수 있음. 클라우드 환경에서 배포되는 경우 특히 주의가 필요함
- 이에 대한 보안요구 사항은 다음과 같음
  - 배포 환경에 대한 접근 제어를 강화하고, 모델에 대한 접근 권한을 최소화하여 무단 접근을 방지함
  - 암호화된 모델 배포 방식을 사용하여 모델이 악의적으로 수정되지 않도록 하고, 배포된 모델의 무결성을 보장하는 디지털 서명을 적용
  - 보안 모니터링 도구를 사용하여 배포된 환경에서 비정상적인 활동을 감지하고, 실시간으로 대응할 수 있는 시스템을 구축함

4.2 API 및 인터페이스 보안

- API 보안은 애플리케이션 간의 인터페이스를 보호하는 것임. 적절한 API 보안이 없으면 민감한 데이터가 노출되고, 시스템이 감염되며 서비스가 중단될 수 있음

4.2.1 AI 모델이 배포된 후, API를 통해 외부 시스템과 상호작용하는 경우, 충분한 보안 조치 기능을 갖추고 있는가? YES ☐ NO ☐ N/A ☐

- AI 서비스가 외부 시스템과 통신하는 API나 인터페이스가 안전하지 않으면 공격자가 이를 통해 시스템에 접근하거나 데이터를 탈취할 수 있음
- 이 때 발생 가능한 취약점은 다음과 같음
  - API가 충분한 인증 및 권한 관리 없이 공개되어 있으면, 무단 접근이 발생할 수 있음
  - API 트래픽이 암호화되지 않으면 중간자 공격(Man-in-the-Middle Attack)에 노출될 수 있음
  - API에 대한 과도한 요청으로 서비스 거부(DoS) 공격이 발생할 위험이 있음
- 위와 같은 공격에 대한 예방책은 아래와 같음
  - 모든 API 요청에 대해 인증 및 권한 관리를 강화하고, 민감한 데이터에 접근할 때는 다중 인증(MFA)을 적용함
  - API 트래픽은 암호화(TLS)를 사용하여 보호하고, 데이터를 안전하게 주고받도록 보장함
  - API 호출 제한(Rate Limiting)을 설정하여 과도한 요청을 방지하고, 비정상적인 요청 패턴을 탐지하여 차단하는 시스템을 구축함

4.2.2 배포된 AI 모델이 실시간으로 데이터를 수신하고 이를 처리할 때, 중간자 공격(Man-in-the-Middle Attack)에 대응하고 있는가? YES ☐ NO ☐ N/A ☐

- 데이터 전송 중 암호화가 미흡할 경우 “중간자 공격(Man-in- the-Middle Attack)”을 통해 데이터가 탈취될 수 있음
- 실시간 데이터 처리 중 데이터 유출 등을 방지하기 위한 예방책은 다음과 같음
  - 암호화된 데이터 전송(TLS/SSL) 프로토콜을 사용하여 전송 중인 데이터의 기밀성을 보호함
  - 실시간 데이터 처리 시스템에 대한 접근 제어를 엄격히 적용하고, 민감한 데이터에 대한 접근 권한을 최소화함
  - 데이터 처리 중 로그 관리를 강화하고, 민감한 정보는 로그에서 제거하거나 익명화 처리함

4.2.3 AI 모델의 API에 대한 접근 권한을 제한하고, 강한 인증 메커니즘을 사용해 불법 접근을 방지하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 서비스를 제공하는 동안 사용자 인증 및 접근 제어가 제대로 설정되지 않으면, 공격자가 서비스에 무단으로 접근하거나 시스템을 악용할 수 있음
- AI 모델의 API에 대한 접근 권한을 제한하고 강한 인증 메커니즘을 적용하려면 인증 강화(OAuth, MFA), 역할 기반 권한 제어(RBAC), 데이터 암호화, Rate Limiting과 같은 다층적인 보안 전략이 필요함. 이러한 조치를 자동화 도구와 정책을 통해 체계적으로 관리하면 불법 접근을 효과적으로 방지하고 API의 신뢰성과 안정성을 보장할 수 있음
- (예시) 인증 및 권한 부여 강화
  ❶ API 키 관리: 각 사용자나 애플리케이션에 고유한 API 키를 발급. API 키에 유효 기간 설정. 키 별로 사용량 제한(Rate Limiting) 적용. 만료된 키는 즉시 회수 및 재발급
  ❷ OAuth 2.0 및 OpenID Connect: 업계 표준 인증 프로토콜을 활용.
  - OAuth 2.0: 클라이언트가 요청하는 권한 범위(스코프)를 지정. Access Token을 통해 제한된 API 호출 권한 부여

  ❸ 다중 인증(MFA, Multi-Factor Authentication): 사용자가 추가 인증 단계(OTP, 생체 인증)를 거쳐야만 API에 접근 가능
  ❹ 세션 기반 접근 제어: Access Token을 기반으로 세션 만료 시간 설정. 장기적으로는 Refresh Token을 사용하여 새 토큰 발급

4.2.4 API 사용자는 필요한 권한만 부여받도록 최소 권한 원칙(Least Privilege)을 적용하고 있는가?

YES ☐ NO ☐ N/A ☐

- 최소 권한 원칙을 적용하려면 역할 기반 접근 제어(RBAC)와 정책 기반 접근 제어(PBAC)를 중심으로 인증, 권한 부여, 데이터 보호, 사용량 제한 등을 통합적으로 관리해야 함. 또한 정기적인 모니터링과 권한 리뷰를 통해 권한이 필요 이상으로 부여되지 않도록 하고, 이를 자동화하여 운영 효율성을 높이는 것이 중요함
- (예시) 역할 및 권한 정의
  ❶ 역할 기반 접근 제어(RBAC, Role-Based Access Control)
  - 사용자 역할 정의: API 사용자의 역할과 책임을 명확히 구분하고 각 역할에 필요한 최소한의 권한만 할당

    예: 관리자(Admin), 개발자(Developer), 일반 사용자(User), 모니터링 사용자(Monitoring)

  ❷ 세분화된 권한 정의: API 엔드포인트별로 권한을 세분화

  예: GET 요청 권한만 부여. 민감 데이터 수정 또는 삭제는 관리자만 수행 가능
- (예시) 정책 기반 접근 제어(PBAC, Policy-Based Access Control)
  - 동적으로 권한을 설정하기 위해 정책 기반 접근 제어 적용
  - 조건 기반 권한 설정: 요청이 특정 조건을 충족할 때만 권한 부여

  예: 시간 제한(업무 시간대만 허용), IP 화이트리스트

4.2.5 AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 시스템에 연결된 장치를 보호하여 보안 위협으로부터 안전하게 유지해야 함
  - 장치 인증 및 승인: AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 함
  - 장치 보안 설정: 각 장치에 대한 보안 설정을 강화하여 취약점을 최소화
- AI 시스템에 연결된 장치를 보호하기 위해 장치 인증, 승인 프로세스, 네트워크 보안, 실시간 모니터링을 포함한 다층적인 보안 전략이 필요함. 이러한 체계적인 접근 방식은 비인가 장치로 인한 보안 위험을 최소화하고, 시스템의 신뢰성과 안정성을 유지하는 데 필수적임
- (예시) 장치 인증 메커니즘 구축
  ❶ 고유한 장치 ID 할당: 각 장치에 고유한 ID(Device ID)를 할당하여 식별. 장치 등록 시 고유 ID와 관련된 메타데이터(예: MAC 주소, IP, 위치 정보)를 기록
  ❷ 인증 프로토콜 사용: 각 장치에 디지털 인증서(예: X.509 인증서)를 발급. 장치가 시스템에 연결할 때 인증서를 통해 신뢰 관계 확인
  ❸ API 키 또는 토큰 기반 인증: 각 장치에 API 키 또는 OAuth 토큰 발급. 토큰은 만료 시간과 권한 범위를 지정하여 오용 방지
- (예시) 장치 등록 및 승인 프로세스
  ❶ 장치 등록: 새로운 장치를 연결하기 전에 등록 프로세스를 통해 신뢰성을 확인
  ❷ 장치 승인: 관리자 또는 시스템이 장치 정보를 검토하고 승인 절차를 진행. 승인된 장치만 시스템에 연결할 수 있도록 데이터베이스에 등록
  ❸ 승인 목록(Whitelist) 관리: 승인된 장치의 고유 ID를 기반으로 화이트리스트 생성. 연결 요청 시 화이트리스트를 조회하여 인증 수행
  ❹ 비승인 장치 차단: 인증 실패 또는 화이트리스트에 없는 장치의 연결 시도를 즉시 차단

05 모니터링 및 유지보수

5.1 실시간 모니터링

AI 개발자, AI 서비스 제공자 공통사항

- AI 모델의 입력 데이터와 출력 결과를 실시간으로 모니터링하고 비정상 동작을 탐지하는 것은 보안 위협 방지, 서비스 안정성 유지, 모델 성능 보장과 같은 다각적인 이점을 제공함

5.1.1 모델의 입력 데이터, 출력 결과 등을 실시간으로 모니터링하여 비정상적인 동작을 탐지하고 있는가?

YES ☐ NO ☐ N/A ☐

- 입력 데이터와 출력 결과를 실시간으로 모니터링하여 비정상 동작을 탐지하려면 로그 수집 및 분석, AI 기반 이상 탐지, 실시간 경고 시스템, 데이터 검증 및 후처리를 포함하는 다층적 접근이 필요함. 이러한 체계는 모델의 안정성과 신뢰성을 유지하고, 비정상 동작으로 인한 위험을 최소화하는 데 기여함
- (예시) 실시간 데이터 수집 및 로깅
  - ❶ 입력 데이터 로깅: 모든 입력 데이터를 실시간으로 기록
    - 필수 기록 항목: 입력 데이터 유형, 크기, 요청 시간, 요청자 ID
  - ❷ 출력 데이터 로깅: 모델의 출력 결과를 기록하여 분석
    - 기록 항목: 출력 값, 응답 시간, 요청과의 관련성(추론 결과와 입력 간 관계)
  - ❸ 로그 관리 시스템: 대규모 로그 데이터를 효율적으로 저장하고 관리
- (예시) 실시간 모니터링 대시보드
  - ❶ 시각화 도구 활용: 실시간으로 입력과 출력 데이터를 모니터링하는 대시보드 구성
  - ❷ 핵심 지표 추적
    - 모니터링 대상: API 호출 빈도, 응답 시간, 에러 발생률. 입력 데이터 크기 및 형식. 출력 값의 정상 범위 여부

5.1.2 모델 응답 시간, 사용 패턴을 추적하고 분석하여 보안에 의심스러운 행동을 탐지하고 있는가?

YES ☐ NO ☐ N/A ☐

- 모델의 응답 시간과 사용 패턴을 실시간으로 추적하고 분석하려면 로그 수집, 데이터 분석, 이상 탐지 시스템, 실시간 알림을 포함한 다층적인 보안 전략이 필요함. 이를 통해 보안 위협을 조기에 감지하고, 서비스 품질과 모델 신뢰성을 유지할 수 있음
- (예시) 비정상 패턴 탐지를 위한 분석
  - ❶ 정적 규칙 기반 탐지: 특정 조건을 만족하는 경우 비정상으로 간주
    예: 입력 데이터 크기 초과. 예상치 못한 데이터 형식. 출력 값의 허용 범위를 벗어남
  - ❷ 동적 이상 탐지(Anomaly Detection): 머신러닝 기반 모델을 사용하여 이상 행동 탐지
  - ❸ 패턴 비교: 정상 입력/출력 데이터의 패턴과 현재 데이터를 비교. 비정상적인 요청이나 예외적인 출력 탐지
- (예시) 실시간 모니터링 및 경고 시스템
  - ❶ 응답 시간 모니터링: 요청 응답 시간이 비정상적으로 길거나 짧은 경우 경고. 평균 응답 시간과 비교하여 이상치 탐지
  - ❷ 사용 패턴 분석
    - 비정상적인 요청 패턴: 요청이 특정 시간대에 집중됨. 예상치 못한 엔드포인트 호출
  - ❸ 실시간 알림 설정
    - 비정상적인 행동이 탐지되면 실시간으로 관리자에게 알림
    - 알림 방법: 이메일, SMS 등

5.1.3 AI 모델이 동작하는 서버 및 네트워크의 트래픽을 모니터링하여 비정상적인 요청을 탐지하고 있는가? YES NO N/A

- AI 모델이 동작하는 서버와 네트워크의 트래픽을 모니터링하여 비정상적인 요청을 탐지하려면 네트워크 트래픽 분석, WAF 및 IDS/IPS 도구 사용, 로그 분석, 머신러닝 기반 이상 탐지, 실시간 경고 시스템을 포함한 다층적인 보안 접근이 필요함. 이러한 조치를 통해 네트워크 보안 위협을 신속히 탐지하고 대응할 수 있음
- (예시) 네트워크 트래픽 모니터링 구축
  ❶ 실시간 트래픽 로깅: 모든 네트워크 트래픽(입력/출력 요청)을 실시간으로 로깅
  - 기록 항목: 요청 IP, 포트, 요청 시간, 프로토콜(TCP, UDP), 요청 크기, 응답 시간
  ❷ 트래픽 흐름 분석: 네트워크 트래픽의 정상 패턴을 분석하고, 비정상적인 흐름을 식별
  예: 특정 IP에서 과도한 요청 발생
- (예시) 비정상 요청 탐지를 위한 시스템 구성
  ❶ 정적 규칙 기반 탐지:
  - 사전 정의된 규칙을 기반으로 비정상적인 요청 탐지: IP/포트 차단(화이트리스트/블랙리스트 관리). 비인가된 요청 차단
  예: 초당 100개 이상의 요청을 보내는 IP를 차단
  ❷ 머신러닝 기반 이상 탐지: 정상적인 트래픽 패턴을 학습하고 비정상적인 요청을 탐지
  ❸ DPI(Deep Packet Inspection): 패킷 내부 데이터를 분석하여 비정상적 요청 탐지. 공격 패턴, 악성 코드, 비인가 데이터 식별

5.1.4 API 호출, 입력/출력 등 요청로그를 정기적으로 분석하여 보안에 의심스러운 동작을 탐지하고 있는가? YES NO N/A

- API 호출, 입력/출력 요청 로그를 정기적으로 분석하여 보안 위협을 탐지하려면 로그 수집 및 저장, 정적 규칙 기반 탐지, 머신러닝 기반 이상 탐지, 실시간 경고 및 자동화된 차단, 요청 제한 관리를 통합한 다층적인 보안 체계를 구축해야 함. 이러한 방법은 시스템의 보안과 신뢰성을 유지하는 데 필요함
- (예시) 요청 로그 수집 및 관리
  ❶ 로그 수집 체계 구축: 모든 API 호출, 입력 데이터, 출력 결과를 로그로 기록
  - 기록 항목: 요청 ID, 사용자 ID, 요청 시간, IP 주소, 엔드포인트, 요청/응답 크기, HTTP 상태 코드, 입력/출력 데이터
  ❷ 로그 저장: 로그 데이터를 중앙화하여 안전하게 저장
  ❸ 로그 보존 정책: 규정에 따라 로그 보존 기간 설정
- (예시) 로그 데이터 분석
  ❶ 정적 규칙 기반 분석: 사전 정의된 규칙을 사용하여 의심스러운 동작 탐지
  - 탐지 규칙 예시: 초당 요청 횟수 초과. 동일 IP에서 과도한 호출 발생. 잘못된 API 키 사용
  ❷ 동적 이상 탐지: 머신러닝 기반 이상 탐지 모델로 비정상적인 요청 패턴 식별
  ❸ 로그 필터링 및 클러스터링: 정상 로그와 의심스러운 로그를 분리. 비정상적 로그 그룹 식별

5.1.5 AI 모델과 배포 환경에 대해 모의 해킹을 수행하여 잠재적인 보안 취약점을 탐지하고 수정하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델과 배포 환경에 대해 모의 해킹을 수행하려면 취약점 분석, 공격 시뮬레이션, 보고서 작성, 수정 및 재검증의 체계적인 단계를 따르는 것이 중요함. 정기적인 모의 해킹과 지속적인 보안 모니터링을 통해 AI 시스템의 안전성을 유지하고, 잠재적 보안 위협에 효과적으로 대응할 수 있음
- (예시) 취약점 식별
  - ❶ 네트워크 및 API 테스트: AI 모델의 API 및 네트워크를 스캔하여 취약점을 식별:
    - 엔드포인트 보호 상태 분석(SQL Injection, Command Injection 등)
  - ❷ 모델 보안 테스트
    - 모델 자체의 취약점 테스트: 적대적 샘플(Adversarial Example) 입력. 모델의 민감 데이터 노출 가능성
  - ❸ 데이터 및 저장소 테스트
    - 데이터 유출 가능성 점검: 데이터베이스의 접근 제어 및 암호화 상태 확인
    - 데이터 저장소에서 PII(개인 식별 정보) 보호 여부 분석
  - ❹ 클라우드 및 컨테이너 환경 테스트
    - 클라우드 배포 환경의 보안 설정 검토: 권한 오용, 공개된 API 키 탐지. 컨테이너 취약점 분석
- (예시) 공격 벡터 시뮬레이션
  - ❶ AI 모델 특화 공격 테스트
    - 적대적 샘플: 모델이 잘못된 출력을 생성하도록 설계된 입력 데이터 테스트
    - 데이터 중독(Data Poisoning): 학습 데이터에 악의적인 데이터가 포함될 경우의 영향 분석
    - 모델 반출(Model Extraction): 모델 출력만으로 내부 구조를 복제하려는 시도
  - ❸ API 오용 및 DoS/DDoS 테스트
    - 비정상적인 API 호출 시도의 영향 분석. Rate Limiting과 Quota 설정 검증

5.2 보안 패치 및 업데이트 관리

AI 개발자, AI 서비스 제공자 공통사항

- AI 개발자는 AI 모델이 배포된 후에도 보안을 유지하며, 지속적으로 보안 모니터링과 유지보수를 실시해야 함

5.2.1 모델에 대한 보안 패치 및 업데이트 관리 프로세스를 구축하고 있는가?

YES ☐ NO ☐ N/A ☐

- 보안 패치 및 업데이트 관리 프로세스를 구축하려면 정책 수립, 취약점 탐지, 패치 개발 및 테스트, 단계적 배포, 모니터링 및 피드백을 포함한 체계적인 접근이 필요함. 자동화 도구를 적극 활용하여 보안과 효율성을 동시에 확보할 수 있음
- (예시) 취약점 탐지 및 패치 평가
  - ❶ 취약점 탐지: AI 모델, 종속성, 배포 환경의 취약점 스캔
  - ❷ 패치 필요성 평가: 취약점의 심각도 분석
  - ❸ 종속성 업데이트 관리: AI 모델이 의존하는 라이브러리, 프레임워크, 도구의 최신 버전 유지
- (예시) 패치 개발 및 테스트
  - ❶ 패치 개발: 취약점을 해결하는 보안 패치 코드를 작성. 기존 모델 성능 및 기능에 영향을 최소화하도록 설계
  - ❷ 테스트 환경 준비: 실제 환경과 동일한 스테이징 환경 구축
  - ❸ 테스트 자동화: 정적 분석, 동적 분석, 회귀 테스트 자동화
  - ❹ AI 모델 성능 검증: 보안 패치가 모델의 성능 및 정확성에 미치는 영향을 평가

5.2.2 모델 배포 후 모델 및 라이브러리의 업데이트가 정기적으로 이루어지고 있는가?

YES ☐ NO ☐ N/A ☐

- 모델 및 라이브러리의 업데이트가 정기적으로 이루어지고 있는지 확인하려면 정기 점검 프로세스, 자동화된 도구, 보고 체계, 정책 문서화를 활용하여 체계적으로 관리해야 함. 이를 통해 보안 취약점을 신속히 해결하고, 최신 기술과 기능을 적용하여 AI 모델과 환경의 안정성을 유지할 수 있음
- (예시) 정기 점검 프로세스
  - ❶ 자동화된 라이브러리 취약점 검사: 취약점 스캐너를 활용하여 사용 중인 라이브러리와 프레임워크의 보안 상태를 점검
  - ❷ 모델 성능 평가: 모델이 성능 저하를 보일 경우 업데이트 필요성을 평가
  - ❸ 배포 환경 점검: 배포된 모델의 실행 환경 상태 점검
- (예시) 자동화 및 모니터링 도구 설정
  - ❶ CI/CD(지속적 통합/지속적 배포)에 업데이트 관리 통합: 라이브러리 및 모델 업데이트를 CI/CD 파이프라인에서 자동으로 감지하고 처리
  - ❷ 대시보드 및 보고: 업데이트 상태를 시각적으로 확인할 수 있는 대시보드 구축
  - ❸ 알림 시스템: 업데이트 필요 시 알림을 설정하여 관리자가 신속히 대응할 수 있도록 함

5.2.3 운영 체제, 라이브러리, 프레임워크의 보안 패치를 운영 환경에 적용하기 전에 스테이징 환경에서 패치를 테스트하고 있는가?

YES ☐ NO ☐ N/A ☐

- 운영 체제, 라이브러리, 프레임워크의 보안 패치를 운영 환경에 적용하기 전에 스테이징 환경에서 테스트가 이루어지고 있는지 확인하려면 스테이징 환경 구성 점검, 테스트 프로세스 문서화, 자동화 도구 활용, 테스트 결과 검토 및 승인 절차 확인이 필요함. 이를 통해 보안 패치가 운영 환경에서 문제를 일으키지 않도록 안전하게 관리할 수 있음
- (예시) 스테이징 환경 준비 확인
  - ❶ 스테이징 환경 구성 상태 점검
    - 운영 환경과 동일한 설정으로 스테이징 환경을 구축: 운영 체제, 네트워크, 라이브러리, 프레임워크가 동일해야 함
    - 스테이징 환경에서 실제 데이터를 사용하지 않고, 테스트 데이터를 구성
  - ❷ 테스트 데이터 준비
    - 운영 환경을 시뮬레이션할 수 있는 충분한 양과 질의 테스트 데이터 생성
  - ❸ 테스트 환경 분리
    - 스테이징 환경은 운영 환경과 완전히 독립적으로 작동해야 함: 네트워크 세분화 및 자원 분리
- (예시) 스테이징 환경에서 패치 테스트 여부 확인 방법
  - ❶ 테스트 기록 검토
    - 스테이징 환경에서 보안 패치 테스트가 실행된 기록을 검토: 패치 적용 로그. 테스트 실행 결과 보고서
  - ❷ 모니터링 및 대시보드 활용
    - 스테이징 환경에서 패치 테스트 결과를 실시간으로 모니터링하는 대시보드 확인
  - ❸ 정기 점검 및 감사
    - 보안 패치 적용 프로세스와 테스트 기록이 정기적으로 감사되고 있는지 확인

06 파기

6.1 파기 시 보안

AI 개발자, AI 서비스 제공자 공통사항

- AI 개발자는 AI 모델이 더 이상 사용되지 않거나 교체될 경우, 해당 AI 모델을 안전하게 폐기해야 하며 관련 데이터도 완전히 삭제해야 함

6.1.1 AI 모델이 더 이상 사용되지 않으면, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?

YES ☐ NO ☐ N/A ☐

- 모델 파일을 복구 불가능하게 삭제하려면 정책 수립, 안전한 삭제 기술 활용, 삭제 검증, 로그 기록 및 교육과 같은 다각적인 접근이 필요함. 안전한 삭제 도구와 자동화된 프로세스를 활용하면 효율성과 보안을 동시에 달성할 수 있음
- (예시) 데이터 삭제 준비
  1. 삭제 대상 식별:
     - 삭제해야 할 데이터와 파일 식별(.pth, .onnx, .h5 등 모델 파일 형식), 학습 데이터, 로그 파일, 캐시.
     - 클라우드 스토리지와 로컬 스토리지 모두 포함
  2. 파일 백업 정책
     - 삭제 전, 백업이 필요한 경우 암호화된 백업을 일정 기간 유지: 중요한 기록 보관 요구사항을 충족하기 위해 임시 백업 보관
- (예시) 안전한 삭제 방법
  1. 논리적 삭제(Logical Deletion): 파일을 운영 환경에서 제거하지만 실제 데이터는 삭제되지 않고 보관
  2. 물리적 삭제(Physical Deletion): 파일을 스토리지에서 실제로 삭제하고 복구 불가능하도록 처리
  3. 안전한 삭제(Secure Deletion): 데이터를 복구할 수 없도록 덮어쓰기 방식으로 삭제
  4. 디스크 초기화: 디스크에 저장된 모델 파일을 포함하여 모든 데이터를 완전히 삭제
  5. 클라우드 환경에서 데이터 삭제: 클라우드 제공자의 삭제 및 데이터 보안 정책 준수
- (예시) 삭제 프로세스 검증
  1. 삭제 검증 수행
     - 삭제된 파일이나 데이터가 복구되지 않았음을 확인: 데이터 복구 도구를 사용해 삭제된 데이터 존재 여부 테스트
  2. 삭제 로그 기록
     - 삭제 작업에 대한 기록 보관: 삭제된 파일 이름, 삭제 시간, 삭제 담당자
     - 감사 및 규제 준수를 위해 필요한 경우 보관

- (예시) 정기적인 삭제 프로세스 검토
  - ❶ 삭제 프로세스 검증
    - 삭제 프로세스를 정기적으로 검토하여 최신 보안 요구사항을 반영
  - ❷ 감사 및 보고
    - 데이터 삭제 기록을 감사하고, 삭제 프로세스 개선을 위한 보고서를 작성

6.1.2 AI 모델에서 사용 중이던 데이터가 시스템을 폐기하거나 교체할 때 안전하게 삭제되고 있는가?

YES ☐ NO ☐ N/A ☐

- 시스템을 폐기하거나 교체할 때 사용 중이던 데이터가 안전하게 삭제되었는지 확인하려면 정책 수립, 삭제 대상 식별, 안전한 삭제 방법 적용, 삭제 상태 검증, 로그 기록 및 정기 감사를 포함한 체계적인 접근이 필요함
- (예시) 데이터 삭제 정책 및 절차 수립
  - ❶ 데이터 삭제 정책 정의
    - 삭제 대상: 원본 데이터, 임시 데이터, 캐시, 백업 데이터
    - 삭제 시점: 시스템 폐기, 교체 또는 데이터 사용 종료 시
    - 보존 정책: 보존 기간이 만료된 데이터 삭제
    - 법적/규제 요구사항에 따른 데이터 삭제 방식 지정
  - ❷ 삭제 책임자 지정
    - 데이터 삭제 프로세스를 담당할 팀 또는 개인을 지정:
      - ▸IT 운영팀: 물리적 스토리지 관리 및 데이터 삭제 수행
      - ▸보안 팀: 데이터 삭제 검증 및 감사 수행
- (예시) 데이터 삭제 대상 식별
  - ❶ 삭제할 데이터 목록 작성
    - 삭제가 필요한 데이터와 위치를 식별: 학습 데이터, 모델 생성 로그, 출력 데이터. 클라우드 스토리지, 로컬 스토리지, 데이터베이스 등
    - 사용된 데이터 파일 경로와 유형 기록
  - ❷ 데이터 보존 필요성 평가
    - 보존 정책 또는 계약에 따라 특정 데이터가 삭제 예외 대상인지 검토
    - 법적 요구사항으로 보존이 필요한 경우 암호화 후 별도 저장

6.1.3 AI 모델이 더 이상 사용되지 않으면 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델이 더 이상 사용되지 않을 경우 API 및 인터페이스 비활성화를 체계적으로 관리하려면 API 상태 점검, 비활성화 수행, 모니터링 및 로그 관리와 같은 체계를 마련해야 함. 이를 통해 외부 접근을 완전히 차단하고 보안 위협을 예방할 수 있음
- (예시) API 및 인터페이스 상태 점검
  - ❶ 연결된 API 목록 식별: AI 모델과 연결된 모든 API 및 인터페이스를 식별
  - ❷ 활성 상태 확인: API 및 인터페이스가 현재 활성화되어 있는지 점검
  - ❸ 인증 및 접근 로그 확인: 해당 API가 최근 호출된 기록이 있는지 분석
- (예시) API 및 인터페이스 비활성화
  - ❶ API 엔드포인트 비활성화: API Gateway에서 불필요한 API 엔드포인트를 비활성화하거나 삭제
  - ❷ 외부 접근 제어: 허용된 IP 화이트리스트에서 해당 API 제거
  - ❸ 인증 키 및 토큰 폐기: 사용되지 않는 API의 인증 키, 토큰, 인증서를 무효화
  - ❹ 인터페이스 제거: 모델과 연결된 인터페이스(예: 프론트엔드, 대시보드)를 완전히 비활성화 또는 제거

3

AI 서비스 제공자를 위한 보안 안내서

01 개요

서비스 제공자 대상 「인공지능(AI) 보안 안내서」의 특징

- AI 서비스 제공자는 단순히 모델을 개발하는 데 그치지 않고, 이를 다양한 사용자 및 조직에게 제공하는 플랫폼 역할을 한다. 이러한 서비스는 광범위한 사용자 기반에 영향을 미치기 때문에 시스템 전반의 보안을 보장할 책임이 있다. 또한 서비스 제공자는 API와 사용자 인터페이스를 안전하게 설계하여 악용 가능성을 줄여야 한다. 예를 들어, API 키 탈취나 DoS 공격을 방지하는 것은 서비스 제공자의 역할이다. 「인공지능(AI) 보안 안내서」는 서비스 제공자를 위한 보안 요구사항 및 검증항목을 제시하여 서비스 제공자들이 활용할 수 있도록 하였다.
- 또한 AI 서비스 제공자는 사용자 데이터를 보호하고 모델 훈련 또는 추론 과정에서 데이터가 유출되거나 악용되지 않도록 해야 한다. 서비스 제공자 대상 「인공지능(AI) 보안 안내서」는 서비스 제공자의 특수한 역할과 책임을 반영하여, 보안 위협을 예방하고 서비스의 신뢰성을 강화하며, AI 기술이 안전하게 사용될 수 있도록 돕는 중요한 지침서가 될 수 있을 것으로 기대된다.

서비스 제공자 대상 「인공지능(AI) 보안 안내서」의 활용 방안

- 「인공지능(AI) 보안 안내서」는 AI 서비스가 보안(Security) 관련 공격을 받거나 받을 우려가 있는 경우 서비스 제공자가 실제로 취해야 할 조치 사항으로 활용할 수 있다. AI 개발자가 AI 서비스 제공자를 겸하는 경우도 있으나, 동 안내서에는 서비스 제공자에 해당하는 내용만을 다루고 있으므로 개발자 관련 조치 내용에 대해서는 개발자 대상 「인공지능(AI) 보안 안내서」를 동시에 참조하여 활용할 필요가 있다.

안내서 작성 과정 및 참고 자료

- 서비스 제공자 대상 「인공지능(AI) 보안 안내서」 작업도 2024년 6월부터 「AI 보안 정책 포럼」을 구성하여 운영하였고, 그 외에도 다양한 의견을 수렴하는 과정을 거쳤다. 초안 작성한 후, 학계 및 산업계 전문가 등의 의견수렴을 거쳐 최종본을 마련하였다.

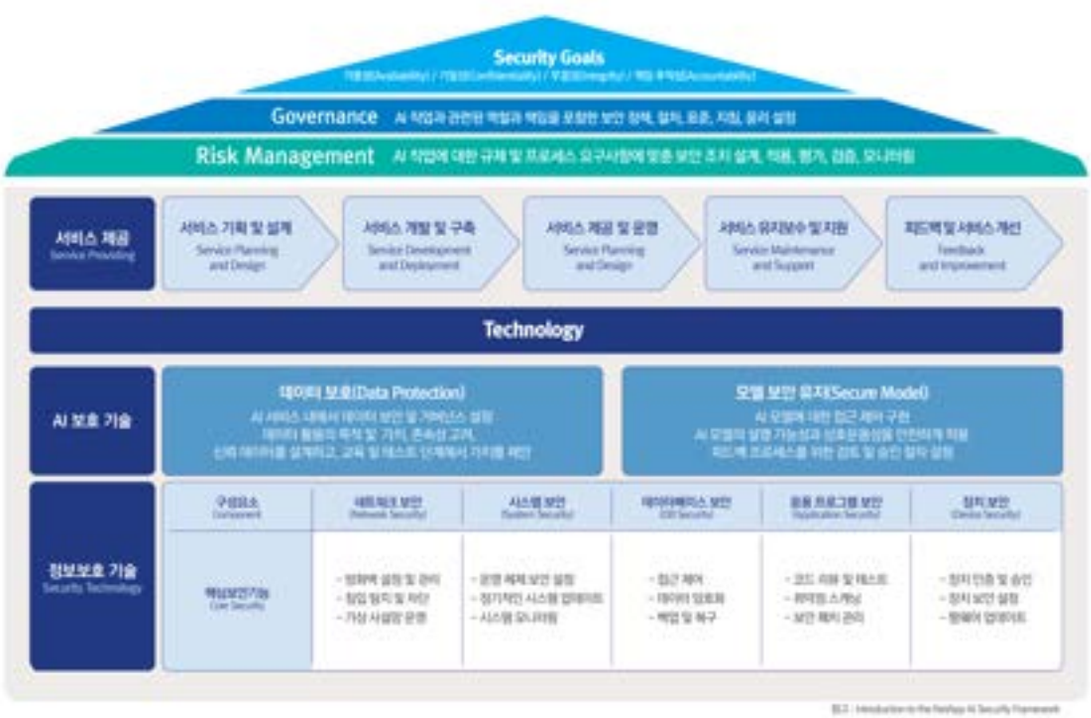
02 AI 서비스 제공자 대상 보안 프레임워크(Security Framework)

01 보안 프레임워크(Security Framework) - 예방(Prevention) 단계

예방(Prevention) 단계에서의 보안 목표: 예방 단계의 Framework는 AI 시스템에 대한 잠재적 위협을 사전에 대비하고 보호하기 위한 다양한 전략과 메커니즘을 포함한다.

- **거버넌스(Governance):** 보안 정책, 절차, 표준, 가이드라인을 수립하여 AI 워크로드와 관련된 역할과 책임을 명확히 한다.
- **위험 관리(Risk Management):** AI 시스템과 서비스에 대한 보안 요구사항을 준수하기 위해 보안 조치를 설계, 적용, 평가, 검증한다.

그림 3-1 AI 서비스 제공자 대상 보안 프레임워크(Security Framework) - 예방 단계



예방 단계에 대한 보안 기술(Security Technology) 적용 방안

- **데이터 보호(Data Protection)**
  - 데이터 보호는 AI 시스템에서 데이터의 기밀성, 무결성, 가용성을 유지하여 보안성을 확보하는 중요한 과정이다. 이는 데이터가 불법적으로 접근되거나 수정되지 않도록 하며, 시스템의 신뢰성을 높이는 역할을 한다.
- **모델 보안 유지(Secure Model)**
  - **모델 보안 유지(Secure Model)는** AI 시스템에서 모델의 보안을 보장하기 위한 다양한 활동과 절차를 포함한다. 이를 통해 AI 모델이 안전하게 작동하고 신뢰성을 유지하도록 한다.
- **사이버 보안(Cyber Security) 기술 적용**
  - AI Security Framework에서 Cyber Security는 AI 시스템 및 데이터의 보안을 강화하는 데 중점을 둔다. 이는 다양한 보안 기술과 절차를 통해 AI 시스템을 보호하고, 데이터 무결성과 기밀성을 유지하며, 시스템의 가용성을 보장하는 것을 목표로 한다.
  - **구성 요소(Component)별 보안 목표 및 요구사항은 다음과 같다.**

표 3-1 예방 단계에 대한 보안 기술 구성 요소별 보안 목표 및 요구사항

구분	보안 목표 및 요구사항
Application Security (응용 프로그램 보안)	[목표] AI 애플리케이션의 보안을 강화하여 취약점 및 공격으로부터 보호한다. • 코드 리뷰 및 테스트: 애플리케이션의 코드를 주기적으로 리뷰하고, 보안 취약점을 발견하여 수정한다. • 취약점 스캐닝: 애플리케이션 내의 보안 취약점을 탐지하고, 이를 해결하기 위한 조치를 취한다. • 보안 패치 관리: 애플리케이션에 대한 최신 보안 패치를 적용하고 보안취약점을 해결한다.
Network Security (네트워크 보안)	[목표] AI 시스템이 연결된 네트워크를 보호하여 데이터의 무결성과 기밀성을 유지한다. • 방화벽 설정 및 관리: 외부 공격으로부터 네트워크를 보호하기 위해 방화벽을 설정하고 관리한다. • 침입 탐지 시스템 (IDS): 네트워크 트래픽을 모니터링하여 비정상적인 활동을 탐지하고 대응한다. • 가상 사설망 (VPN): 네트워크를 통한 데이터 전송 시 암호화된 연결을 제공하여 데이터의 기밀성을 유지한다.
System Security (시스템 보안)	[목표] AI 시스템의 운영 체제 및 관련 인프라를 보호하여 무단 접근과 공격을 방지한다. • 운영 체제 보안 설정: 시스템의 운영 체제에 대한 보안 설정을 강화하여 취약점을 줄인다. • 정기적인 시스템 업데이트: 운영 체제 및 소프트웨어에 대한 최신 업데이트를 적용하여 보안을 유지한다. • 시스템 모니터링: 시스템 로그를 주기적으로 모니터링하여 비정상적인 활동을 탐지하고 대응한다.
DB Security (데이터베이스 보안)	[목표] AI 시스템에서 사용되는 데이터베이스를 보호하여 데이터 무결성과 기밀성을 유지한다. • 접근 제어: 데이터베이스에 대한 접근 권한을 관리하여 무단 접근을 방지한다. • 데이터 암호화: 저장된 데이터를 암호화하여 데이터 유출 시에도 기밀성을 유지한다. • 백업 및 복구 계획: 데이터베이스의 정기적인 백업을 수행하고, 데이터 손실 시 복구 계획을 마련한다.
Device Security (장치 보안)	[목표] AI 시스템에 연결된 장치를 보호하여 보안 위협으로부터 안전하게 유지한다. • 장치 인증 및 승인: AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 한다. • 장치 보안 설정: 각 장치에 대한 보안 설정을 강화하여 취약점을 최소화한다. • 펌웨어 업데이트: 장치의 펌웨어를 최신 상태로 유지하여 보안 취약점을 해결한다.

02 보안 프레임워크(Security Framework) - 탐지·대응(Detection) 단계

탐지·대응(Detection) 단계에서의 Security Framework 목표: 탐지·대응(Detection) 단계의 Framework는 사업자 관점에서 AI 시스템의 보안 위협을 탐지하기 위한 전략과 메커니즘을 설명한다. 탐지(Detection) 단계는 AI 시스템에서 발생할 수 있는 보안 위협을 실시간으로 모니터링하고, 이를 신속하게 식별하여 대응할 수 있도록 하는 데 중점을 둔다.

- **거버넌스(Governance):** AI 생명주기 전반에 대한 위험 관리를 수립하고, AI 워크로드에 대한 탐지 절차, 매뉴얼, 사고대응 팀을 배정한다.
- **위험 관리(Management):** AI 시스템과 서비스에 대한 보안 요구사항을 준수하기 위해 보안 조치를 적용한다.

그림 3-2 AI 서비스 제공자 대상 보안 프레임워크(Security Framework) - 탐지·대응 단계

탐지·대응(Detection) 단계에 대한 보안 기술(Security Technology) 적용 방안

- **데이터 이상 징후 탐지(Data Anomaly Detection)**
  - **데이터 이상 징후 탐지는** AI 시스템의 데이터를 정기적으로 점검하고, 실시간 모니터링 시스템을 통해 이상 징후를 감지하며, 체크리스트를 활용하여 테스트와 훈련 단계를 점검하는 과정이다. 이를 통해 데이터의 무결성, 기밀성, 가용성을 보장하고, AI 시스템의 안정성과 신뢰성을 유지할 수 있다.
- **모델 보안 유지(Secure Model)**
  - **모델 보안 유지는** AI 모델이 안전하게 동작할 수 있도록 알고리즘 검증, 위험 평가 및 모델 조정, 소프트웨어 시각화 도구를 통한 프로세스 자동화 등의 과정을 포함한다. 이를 통해 AI 모델의 신뢰성을 높이고, 외부 위협으로부터 보호할 수 있다.
- **사이버 보안(Cyber Security) 기술 적용**
  - **탐지·대응(Detection) 부문별 목표 및 요구사항**

표 3-2 탐지·대응 단계에 대한 보안 기술 구성 요소별 보안 목표 및 요구사항

구분	보안 목표 및 요구사항
Data Collection Modules (데이터 수집 모듈)	[목표] 다양한 데이터를 수집하기 위한 장치와 방법을 추진한다. • 장비 에이전트 설치, API 연결, 명령 실행, 추출 파일, 수동 등록, PC/서버 보안, PKI/SSO, NMS, SMS 등을 통해 데이터를 수집한다. • (예시) 네트워크 장비에 에이전트를 설치하여 데이터 수집, API를 통해 외부 시스템과 연결, 서버의 보안 로그를 수집한다.
Data Collection System (데이터 수집 시스템)	[목표] 실시간 데이터 수집 및 분석, 이벤트 감지를 담당한다. • 실시간으로 데이터를 수집하고, 이를 분석하여 보안 이벤트를 감지한다. • (예시) 실시간 로그 수집 시스템, 네트워크 트래픽 분석 시스템, 보안 이벤트 모니터링 시스템
Data Management System (데이터 관리 시스템)	[목표] 데이터베이스 분석, 데이터 평가, 설정 및 변경 이력 관리 등을 포함한 데이터 관리 기능을 제공한다. • 지속적인 DB 분석, 설정 변경 이력 관리, 관리 작업 및 데이터 모니터링, 정기 보고서 작성, 기능 개선, 로그 분석, 템플릿 생성, 작업 자동화 등을 수행한다. • (예시) 데이터베이스 관리 시스템(DBMS), 로그 분석 도구, 자동화된 데이터 모니터링 및 보고 시스템

- **핵심 보안기술(Core Security)별 목표 및 요구사항**

구분	목표 및 요구사항
Prevention System (예방 시스템)	[목표] 침해 위험 예측 및 관리, 실시간 침해 및 정보 손상 감지, 대응 시스템 연계 등을 담당한다. • 침해 위험을 예측하고 관리하며, 실시간으로 침해 및 정보 손상을 감지하고 대응 시스템과 연계한다. • (예시) 개인정보 보호 시스템, 실시간 침해 감지 시스템, 침해 대응 연계 시스템
Detection and Response Measures (탐지 및 대응 조치)	[목표] 실시간 데이터 수집 및 분석, 다양한 침해 상황 및 대응 시나리오 수립을 포함한다. • 실시간으로 데이터를 수집하고 분석하며, 다양한 침해 상황에 대한 시나리오를 수립하고 대응 조치를 마련한다. • (예시) 침해 대응 시나리오, 실시간 데이터 분석 시스템, 침해 대응 절차 수립
Error Occurrence and Reporting (오류 발생 및 보고)	[목표] 실시간 오류 감지, 오류 발생 및 보고를 담당한다. • 실시간으로 오류를 감지하고, 발생한 오류를 저장 및 보고한다. • (예시) 오류 감지 시스템, 오류 보고 시스템, 실시간 오류 모니터링 도구

03 AI 서비스 제공자를 위한 보안 요구사항 및 검증항목

체크리스트 요약

생명주기	요구사항 및 체크리스트		서비스제공자		AI 유형	
			담당자	조직	Pred AI	Gen AI
1 서비스 기획 및 설계	**(AI 개발자, AI 서비스 제공자 공통사항) 거버넌스 및 위험관리**					
	1.1	AI 보안(Security) 거버넌스 체계 구축 AI 개발자, AI 서비스 제공자 공통사항				
	1.1.1	AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?		○	○	○
	1.1.2	AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?		○	○	○
	1.1.3	AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?		○	○	○
	1.2	AI 서비스에 대한 위험관리 계획의 수립 AI 개발자, AI 서비스 제공자 공통사항				
	1.2.1	AI 모델 개발/서비스 제공 생명주기 및 공급망 과정에서 나타날 수 있는 위험요소를 분석·도출하고 있는가?	○		○	○
	1.2.2	AI 서비스에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	○		○	○
	1.2.3	AI 서비스에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	○		○	○
	(AI 서비스 제공자) 계약관리					
	1.3	서비스 수준 계약(SLA) 관리				
	1.3.1	공급업체와 계약시, SLA에 보안요구 사항을 명확히 포함했는가?		○	○	○
	1.3.2	보안 침해 발생 시를 대비하여, 대응 계획을 수립하고 있는가?	○	○	○	○
	1.3.3	보안 침해 발생 시를 대비하여, 책임 소재를 명확히 하고 있는가?		○	○	○
2 서비스 개발 및 구축	2.1	코드 취약 점검 등 관리점				
	2.1.1	정적 및 동적 코드 분석 도구를 사용하여 소스 코드의 보안 취약점을 분석하고 있는가?	○		○	○
	2.1.2	코드 리뷰 프로세스를 도입하여 보안 문제가 있는 부분을 검토하고 개선하고 있는가?	○		○	○
	2.2	모델 환경의 보안				
	2.2.1	모델 환경에 대한 접근 제어를 강화하고, 모델에 대한 접근 권한을 최소화하여 무단 접근을 방지하고 있는가?	○		○	○
	2.2.2	모델이 악의적으로 수정되지 않도록 모델의 무결성을 보장하는 방법을 적용하고 있는가?	○		○	○
	2.2.3	보안 모니터링 도구를 사용하여 모델의 비정상적인 활동을 감지하고, 실시간으로 대응할 수 있는 체계를 구축하고 있는가?	○		○	○

생명주기	요구사항 및 체크리스트	서비스제공자		AI 유형	
		담당자	조직	Pred AI	Gen AI
2 서비스 개발 및 구축	2.3 데이터 보안 AI 개발자, AI 서비스 제공자 공통사항				
	2.3.1 적대적 공격 등 데이터 공격에 대한 방어 수단을 강구하고 있는가?	○		○	○
	2.3.2 데이터 저장 및 전송 시 무결성을 보호하기 위한 조치를 하고 있는가?	○		○	○
	2.3.3 중요 데이터에 대한 기밀성 유지를 위해 보호 방안을 마련하고 있는가?	○		○	○
	2.3.4 전송구간에서 중요정보 유출을 방지하기 위한 보호 방안을 마련하고 있는가?	○		○	○
	2.3.5 데이터 유출시 책임추적을 할 수 있도록 조치를 하고 있는가?	○	○	○	○
	2.4 API 및 인터페이스 보안				
	2.4.1 API 통신을 암호화하여 데이터가 전송되는 구간에서 외부 공격에 대한 방어를 하고 있는가?	○		○	○
	2.4.2 모든 API 요청에 대해 인증 및 권한 관리를 강화하고, 중요 데이터에 접근할 때는 강한 인증 메커니즘을 적용하고 있는가?	○		○	○
	2.4.3 API 트래픽은 암호화 기술을 사용하여 보호하고, 데이터를 안전하게 주고 받도록 보장하고 있는가?	○		○	○
	2.4.4 API 호출 제한(Rate Limiting)을 설정하여 과도한 요청을 방지하고, 비정상적인 요청 패턴을 탐지하여 차단하고 있는가?	○		○	○
3 서비스 제공 및 운영	3.1 로그 및 운영 데이터 보안				
	3.1.1 데이터 처리 중 접속로그 관리를 강화하고 있는가?	○		○	○
	3.1.2 로그 파일 및 운영 데이터에 암호화를 적용하고, 중요정보는 별도로 관리하여 유출을 방지하고 있는가?	○		○	○
	3.1.3 운영 중 발생하는 데이터를 안전하게 저장하고, 접근 제어를 통해 인증된 관리자만 로그에 접근할 수 있도록 설정하고 있는가?	○		○	○
	3.1.4 로그 데이터 접근 권한을 최소화하고, 접근 제어 및 사용자 활동 기록을 통해 비정상적인 접근을 탐지하고 있는가?	○		○	○
4 서비스 유지보수 및 지원	4.1 모니터링, 업데이트 및 패치 AI 개발자, AI 서비스 제공자 공통사항				
	4.1.1 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집·관리하고 있는가?	○		○	○
	4.1.2 AI 시스템이 정상적으로 작동하지 않거나 예기치 못한 오류가 발생할 경우 이를 조기에 탐지하고 대응하는 메커니즘이 있는가?	○		○	○
	4.1.3 AI 시스템의 보안 패치나 모델 업데이트가 정기적으로 이루어지고 있는가?	○		○	○
	4.2 성능 및 장애 관리				
	4.2.1 서비스 장애가 발생했을 때 자동으로 복구할 수 있도록 하고 있는가?	○		○	○

생명주기	요구사항 및 체크리스트		서비스제공자		AI 유형	
			담당자	조직	Pred AI	Gen AI
	4.2.2	모델 성능을 지속적으로 모니터링하고, 성능 저하가 감지되면 재학습을 통해 성능을 유지하고 있는가?	○		○	○
	4.2.3	실시간으로 모델 드리프트 탐지 시스템을 운영하여 모델 성능이 저하될 경우 즉시 대응할 수 있는 체계를 마련하고 있는가?	○		○	○
	4.2.4	정기적으로 모델 재학습 및 업데이트를 수행하여 새로운 데이터 패턴을 반영하고, 성능을 개선하고 있는가?	○		○	○
	4.2.5	AI 서비스에 대해 다중화(HA) 및 백업 시스템을 구축하여 장애 발생 시에도 서비스가 연속적으로 제공될 수 있도록 하고 있는가?	○		○	○
	4.2.6	침입차단시스템 등을 통해 외부에서 발생하는 DoS/DDoS 공격을 방어하고, 실시간 모니터링 시스템을 운영하여 장애를 빠르게 감지하고 대응하고 있는가?	○		○	○
5 피드백 및 서비스 개선	5.1	사용자 피드백 관리				
	5.1.1	사용자 피드백 시스템에 입력 검증 및 필터링을 적용하여 악성 코드나 비정상적인 데이터의 입력을 차단하고 있는가?	○		○	○
	5.1.2	피드백을 자동으로 분석하기 전에 사전 검증 절차를 마련하여 피드백 데이터의 무결성을 확인하고 있는가?	○		○	○
	5.1.3	최소 권한 원칙(Least Privilege)을 적용하여 피드백 및 개선 과정에서 접근할 수 있는 권한을 최소화하고 있는가?	○		○	○
	5.1.4	피드백 처리 및 개선 과정에서 이루어진 모든 접근 및 변경 사항을 감사 로그로 기록하고, 정기적으로 검토하여 무단 접근을 탐지하고 있는가?	○		○	○
6 파기	6.1	파기 시 보안 AI 개발자, AI 서비스 제공자 공통사항				
	6.1.1	모델 파기 시, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?	○		○	○
	6.1.2	시스템을 폐기하거나 교체할 때 AI 모델에서 사용 중이던 관련 파일 및 데이터를 안전하게 삭제하고 있는가?	○		○	○
	6.1.3	AI 모델이 더 이상 사용되지 않으면 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?	○		○	○

01 서비스 기획 및 설계

〈1.1 AI 보안(Security) 거버넌스 체계 구축〉, 〈1.2 AI 서비스에 대한 위험관리 계획의 수립〉은 AI 개발자, AI 서비스 제공자 모두에게 해당하는 공통사항으로 〈제2장. AI 개발자를 위한 보안 안내서〉를 참고하여 적용하기 바랍니다.

1.1 AI 보안(Security) 거버넌스 체계 구축

AI 개발자, AI 서비스 제공자 공통사항

• AI 보안 거버넌스는 단순한 보안 조치를 넘어 기업의 리스크를 줄이고, 법적 규제를 준수하며, 지속 가능성을 확보하는 핵심적인 프레임워크임. AI가 기업에서 점점 더 중요한 기술로 자리 잡고 있는 만큼, 기업들은 전문 조직 구성, 정책 수립, 전문인력 확보 등 AI 보안 거버넌스 체계를 구축하여 AI 서비스가 안정적으로 운영이 되도록 해야 함

		YES	NO	N/A
1.1.1	AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?	☐	☐	☐
1.1.2	AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?	☐	☐	☐
1.1.3	AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?	☐	☐	☐

1.2 AI 서비스에 대한 위험관리 계획의 수립

AI 개발자, AI 서비스 제공자 공통사항

• AI 서비스를 제공하는 기업은 데이터 조작, 모델 탈취, 적대적 공격 등 다양한 보안 위협에 노출될 수 있어, 이를 예방하고 대응하기 위한 위험관리 계획이 필요함. 위험관리 계획을 통해 AI 서비스 제공 생명주기에 걸쳐 나타날 수 있는 위험요소를 분석·도출하고 위험요소를 제거·완화하기 위한 방안을 통해 AI 시스템의 보안성과 안전성을 유지해야 함

		YES	NO	N/A
1.2.1	AI 서비스 제공 생명주기에 걸쳐 나타날 수 있는 위험요소를 분석·도출하고 있는가?	☐	☐	☐
1.2.2	AI 서비스에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	☐	☐	☐
1.2.3	AI 서비스에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	☐	☐	☐

1.3 서비스 수준 계약(SLA) 관리

- AI 서비스의 보안 요구 사항이 명확하게 정의되지 않으면 이후 단계에서 발생할 수 있는 보안 문제에 충분히 대비하지 못하게 됨. 이로 인해 데이터 보호, 모델 보호, 운영 보안 측면에서 취약점이 발생할 수 있음

  ※ 서비스 수준 계약(SLA)은 공급업체가 고객에게 제공하기로 약속한 서비스 수준을 명시하는 아웃소싱 및 기술 공급업체 계약임. 이 계약에는 가동 시간, 납품 시간, 응답 시간 및 해결 시간 등의 지표가 포함되어 있음

1.3.1 공급업체와 계약 시, SLA에 보안요구 사항을 명확히 포함했는가? YES ☐ NO ☐ N/A ☐

- SLA에 보안 요구 사항을 명시함으로써 위험 관리에 선제적으로 대응할 수 있는 기회를 제공받음. SLA는 잠재적인 위험과 위협을 미리 식별하고 비즈니스 이해관계자가 이러한 문제를 방지하거나 완화하기 위한 계획을 개발하는 데 도움이 됨
- (예시) 데이터 보안 관련 명시
  - 데이터 암호화: 데이터 저장 및 전송 중 암호화 수준(예: AES-256)과 구현 방법을 명시
  - 데이터 무결성: 데이터 무결성을 확인하고 보장하기 위한 검증 메커니즘(예: 해싱, 서명 등)
  - 데이터 소유권 및 사용 권한: 데이터의 소유권이 AI 서비스 제공기업에 있고, 공급업체는 명시된 용도로만 데이터를 사용할 수 있도록 규정
  - 데이터 반환 및 삭제: 계약 종료 시 데이터 반환과 안전한 삭제 절차를 명확히 기술
- (예시) 보안 모니터링 및 위협 탐지 관련 명시
  - 보안 이벤트 모니터링: 보안 정보 및 이벤트 관리(SIEM) 시스템을 통해 실시간 모니터링 및 이상 탐지 수행
  - 위협 탐지 및 경고: 공급업체가 비정상적인 활동이나 보안 위협을 실시간으로 탐지하고 경고를 제공.
  - 로그 관리: 모든 데이터 및 시스템 활동 로그를 생성, 보관, 분석하도록 요구. 보관 기간을 명시

1.3.2 보안 침해 발생 시를 대비하여, 대응 계획을 수립하고 있는가?

YES ☐ NO ☐ N/A ☐

- SLA에 보안 요구 사항을 명시하면, 보안 침해 발생 시 대응 계획이 명확히 설정되고, 문제 발생 시 혼란을 최소화할 수 있음
- (예시) 보안 침해의 정의 및 범위 명확화
  - 보안 침해 정의: SLA 계약서에서 “보안 침해”의 정확한 정의를 명확히 기술
    예: 데이터 유출, 무단 접근, 시스템 무결성 손상, 서비스 거부(DoS), 악성코드 감염 등, AI 탈옥 등
  - 범위 및 영향: 침해가 발생할 수 있는 데이터, 시스템, 네트워크의 범위와 이로 인한 영향 수준을 명확히 명시
- (예시) 침해 탐지 및 보고
  - 탐지 의무: 공급업체가 보안 침해를 탐지해야 하는 방법과 시스템(예: SIEM, IDS/IPS)을 명시
  - 보고 시간: 침해 탐지 후 AI 서비스 제공기업에 보고해야 하는 최대 시간을 명시
  - 보고 방식: 보고 절차와 채널(예: 이메일, 전화, 전용 포털)을 명시
  - 초기 보고 내용: 침해의 유형, 범위, 잠재적 영향, 즉각적 대응 조치 등을 포함한 초기 보고서 작성 요구

1.3.3 보안 침해 발생 시를 대비하여, 책임 소재를 명확히 하고 있는가? YES ☐ NO ☐ N/A ☐

- SLA에 보안 요구 사항이 명시되지 않으면, 서비스 제공 중에 보안 침해가 발생해도 책임 소재가 불분명해질 수 있음. 따라서 SLA에 보안 요구 사항을 명확하게 포함시키고, 보안 침해 발생 시 책임 소재를 명확히 규정하는 것이 필요함
- (예시) 책임 소재 정의
  - 공급업체의 책임: 공급업체가 제공하는 서비스 및 시스템에서 발생한 보안 침해에 대한 책임을 명확히 함
    예: 클라우드 인프라 보안, 데이터 저장 및 전송 과정의 암호화, 시스템 취약점 관리
  - AI 서비스 제공기업의 책임: 제공기업이 직접 관리하는 부분에서 발생하는 보안 침해는 제공기업이 책임을 짐
  - 공동 책임 구역: 데이터 전송, API 통합 등 양측이 상호작용하는 영역에서의 책임 분할을 명확히 정의
- (예시) 보안 의무 명시
  - 공급업체의 보안 의무: 공급업체가 준수해야 할 보안 표준 및 의무를 명시. 보안 취약점 점검, 주기적인 침투 테스트, 데이터 암호화 적용 등 구체적인 보안 조치를 SLA에 포함
  - AI 서비스 제공기업의 의무: 공급업체에게 제공하는 데이터의 정확성과 무결성을 보장할 의무. 적법한 사용 및 데이터 제공 조건을 준수
- (예시) 법적 및 재정적 책임
  - 재정적 보상: 보안 침해로 인해 발생한 금전적 손실, 벌금, 고객 보상 등에 대한 책임 소재를 명확히 함
    예: “공급업체의 과실로 발생한 손실은 공급업체가 전적으로 책임진다.”
  - 책임 제한: 공급업체의 책임 한도를 설정하되, 중과실이나 중대한 위반에는 예외를 둠
  - 보험 가입: 공급업체가 사이버 보험에 가입하고, 사고 발생 시 이를 통해 보상을 제공하도록 요구

02 서비스 개발 및 구축

2.1 코드 취약점 점검 등 관리

- AI 서비스 과정에서 소스 코드나 알고리즘에 보안 취약점이 존재할 경우, 공격이 발생할 수 있음. 특히 공개된 코드나 외부 라이브러리를 사용할 때 보안에 취약할 수 있음

2.1.1 정적 및 동적 코드 분석 도구를 사용하여 소스 코드의 보안 취약점을 분석하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 서비스를 구현하는 과정에서 소스 코드나 알고리즘에 보안 취약점이 존재할 경우, 공격자가 이를 악용할 수 있음. 특히 공개된 코드(오픈 소스코드)나 외부 라이브러리를 사용할 때 보안에 취약할 수 있음
- 정적 및 동적 코드 분석 도구를 사용하여 소스 코드의 보안 취약점을 탐지하고 수정함

동적 분석과 정적 분석 비교

동적 분석	프로그램을 실행하여 평가하는 것을 포함함. 이 유형의 분석은 메모리 사용, 성능 및 잠재적인 런타임 오류를 포함한 런타임 동작을 확인함. 메모리 누수, 동시성 문제 및 잘못된 프로그램 출력과 같이 코드가 실행 중일 때만 나타나는 문제를 식별하는 데 유용함
정적 분석	코드를 실행하지 않고 검사함. 이 방법은 코드의 구조, 제어 흐름 및 데이터 사용을 분석하여 구문 오류, 코드 표준 위반 및 잠재적 보안 취약성을 탐지하는 데 중점을 둠

- Generative AI 서비스 제공시 정적 및 동적 코드 분석 도구를 활용하여 보안 취약점을 분석하는 것은 유해 콘텐츠 생성 방지, 데이터 보호, API 및 서비스 보안 강화 등을 위해 필요함
  - 유해 콘텐츠 생성 방지
    - ▸부적절한 출력 방지: Generative AI는 사용자 입력을 바탕으로 텍스트, 이미지, 코드 등을 생성함. 정적 분석 도구는 코드 상에서 유해 콘텐츠를 생성할 가능성이 있는 로직(예: 검증되지 않은 입력 처리, 비정상적 조건)을 식별할 수 있음
    - ▸입력 검증 강화: 동적 분석 도구는 실제 사용 시 모델이 예상치 못한 입력으로 유해한 결과를 생성할 가능성을 시뮬레이션하여 보완할 수 있음
  - 데이터 보호
    - ▸훈련 데이터 노출 방지: Generative AI는 훈련 데이터 기반으로 작동하며, 코드 취약점이 있으면 훈련 데이터 일부가 의도치 않게 생성 결과에 포함될 수 있음. 코드 분석 도구는 데이터 노출을 유발할 수 있는 부분을 자동으로 탐지함
  - API 및 서비스 보안 강화
    - ▸API 남용 방지: 정적 분석은 API 인증, 입력 검증, 속도 제한과 같은 보안 메커니즘이 올바르게 구현되었는지 검토할 수 있음

▸실시간 위협 대응: 동적 분석 도구는 실행 중인 서비스에 대한 악의적인 요청 시나리오(예: SQL 인젝션, XSS, DoS 공격)를 테스트하여 취약점을 보완함

- 모델 및 서비스 신뢰성 보장
  - ▸예상치 못한 동작 방지: 정적 및 동적 분석은 코드가 다양한 입력과 상황에서 의도대로 작동하는지 확인하고, 생성 결과의 신뢰성을 높이는 데 도움을 줌

2.1.2 코드 리뷰 프로세스를 도입하여 보안 문제가 있는 부분을 검토하고 개선하고 있는가? YES ☐ NO ☐ N/A ☐

- 코드 리뷰는 본인이 작성하지 않는 코드의 내용을 점검하고, 피드백을 주는 과정임.
- 여기서 피드백은 오타, 버그에 대한 가능성, 좋은 코드를 위한 피드백이 될 수 있음. 예상되는 문제를 일찍 파악하는 이점도 있을 뿐만 아니라 무엇보다 해당 코드를 작성한 사람에게만 책임이 있는 것이 아니라, 서비스 제공자에게도 있다는 문화를 만들어내는 것에 목표가 있음. AI 서비스 제공기업이 코드 리뷰 프로세스를 통해 보안 문제를 검토하고 개선하려면 체계적이고 철저한 절차를 도입해야 함
- Predictive AI 서비스 제공 시 코드 리뷰 프로세스를 도입하여 보안 문제가 있는 부분을 검토하는 것은 예측 결과의 정확성과 무결성 보장, 중요 데이터 보호, 시스템 안정성 및 성능 보장 등을 위해 필요함
  - 예측 결과의 정확성과 무결성 보장
    - ▸오류 및 편향 제거: Predictive AI 모델의 로직에서 데이터 처리 오류나 편향을 유발할 수 있는 부분을 코드 리뷰를 통해 식별하고 수정할 수 있음
    - ▸데이터 드리프트 방지: 데이터 입력 및 처리 로직이 최신 데이터에 적합하게 구현되었는지 검토하여, 데이터 드리프트로 인한 예측 오류를 줄일 수 있음
  - 중요 데이터 보호
    - ▸데이터 유출 방지: Predictive AI는 종종 민감한 데이터를 처리함. 코드 리뷰를 통해 데이터가 안전하게 처리되고 저장되는지 확인하여 데이터 유출 사고를 예방할 수 있음
    - ▸액세스 제어 검증: 코드에서 데이터 접근 제어가 적절히 구현되었는지 검토하여, 비인가된 접근을 차단할 수 있음
  - 시스템 안정성 및 성능 보장
    - ▸리소스 관리 최적화: 코드 리뷰는 모델 실행 로직에서 리소스 낭비를 유발할 수 있는 코드를 식별하여 시스템의 안정성을 높일 수 있음
    - ▸장애 대응 강화: 코드 리뷰를 통해 오류가 발생했을 때 대체 경로 또는 복구 절차가 제대로 구현되었는지 확인할 수 있음
- (예시) 코드 리뷰 정책 및 가이드라인 수립
  - 코드 리뷰 목적 명확화: 보안 취약점 식별, 코드 품질 향상, 규정 준수 확인 등 명확한 목표를 설정
  - 표준화된 리뷰 가이드라인: 코드 리뷰 중점 사항을 포함한 체크리스트를 개발

- 보안 기준 준수: OWASP Secure Coding Practices, ISO/IEC 27001 등 보안 표준에 기반한 리뷰 기준 설정

• (예시) 코드 리뷰 프로세스 설계
- 단계별 코드 리뷰: 각 코드 변경 사항에 대해 다음과 같은 단계로 검토
  ▸ 자동화된 분석: 정적 분석 도구(SAST)를 사용하여 코드에서 보안 취약점을 자동으로 탐지
  ▸ 동료 리뷰: 동료 개발자들이 수동으로 코드를 검토하며 보안 취약점과 논리적 오류 탐지
  ▸ 보안 전문가 리뷰: 보안 전문가가 민감하거나 중요한 코드의 보안성을 검토
- PR(풀 리퀘스트) 기반 리뷰: 모든 코드 변경 사항은 Pull Request(PR)로 제출하고, 병합 전에 승인을 받도록 함
- 리뷰 범위 지정: 변경된 코드뿐 아니라 해당 변경이 영향을 미치는 기존 코드도 함께 검토

2.2 모델 환경의 보안

• AI 모델 환경이 안전하지 않으면 공격자가 모델에 무단으로 접근하거나 모델을 악용할 수 있고 클라우드 환경에서 배포되는 경우 특히 주의가 필요함

2.2.1 모델 환경에 대한 접근 제어를 강화하고, 모델에 대한 접근 권한을 최소화하여 무단 접근을 방지하고 있는가?

YES ☐ NO ☐ N/A ☐

• AI 모델 환경이 안전하지 않으면 공격자가 모델에 무단으로 접근하거나 모델을 악용할 수 있음. 클라우드나 온프레미스 환경에서 AI 모델에 무단 접근이 발생할 수 있음
• 모델 환경에 대한 접근 제어를 강화하고, 모델에 대한 접근 권한을 최소화하여 무단 접근을 방지해야 함
• (예시) 역할 기반 접근 제어(RBAC)
- 역할 정의: 모델 환경에 대한 접근 권한을 역할(예: 개발자, 데이터 과학자, 운영자)별로 구분
- 최소 권한 원칙: 각 역할에 대해 최소한의 권한만 부여
  예: 운영자는 모델 실행만 가능하고, 데이터 과학자는 모델 수정만 가능하도록 설정
- 동적 권한 할당: 사용자 요청에 따라 임시 권한을 부여하고, 작업 완료 후 권한을 회수
• (예시) 다단계 인증(MFA)
- MFA 적용: 모든 모델 환경에 대한 접근은 다단계 인증(예: 비밀번호 + 인증 앱)을 요구
- 장치 인증: 접근에 사용되는 장치의 신뢰성을 검증하여 승인된 장치에서만 접근 가능하도록 설정
• (예시) 네트워크 및 IP 기반 접근 제어
- VPN 및 전용 네트워크: 모델 환경에 대한 접근은 보안 VPN을 통해서만 가능하도록 제한
- IP 화이트리스트: 허용된 IP 주소에서만 모델 환경에 접근할 수 있도록 구성
- 방화벽 설정: 모델 환경을 보호하는 방화벽을 설정하고, 외부에서의 무단 트래픽을 차단

2.2.2 모델이 악의적으로 수정되지 않도록 모델의 무결성을 보장하는 방법을 적용하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 서비스 제공기업은 모델의 무결성을 보장하기 위해 필요한 방법을 적용해야 함. 모델의 무결성을 보장하면 악의적인 수정으로 인한 오작동과 데이터 유출을 방지하여 AI 시스템의 신뢰성과 안전성을 유지할 수 있음. 이는 서비스 품질 저하와 고객 신뢰 손실을 예방하고, 규제 및 보안 표준을 준수하기 위해 필요함
- (예시) 모델 무결성 검증
  - 모델 해싱(Hashing): 모델 파일의 해시 값을 생성하여 무단 변경 여부를 검증. 배포 전후의 모델 상태를 비교하여 일관성을 확인
  - 모델 검증 프로세스: 배포 전에 모델 성능 및 보안 테스트를 수행하여 무결성을 확인
- (예시) 운영 환경 보안
  - 운영 환경 격리: 모델 운영 환경을 테스트 및 개발 환경과 격리하여 무단 변경을 방지
  - 접근 제어: 역할 기반 접근 제어(RBAC)를 적용하여 모델 환경에 접근할 수 있는 사용자를 제한
  - 실시간 무결성 검증: 운영 중인 모델의 무결성을 주기적으로 검증하는 자동화된 시스템을 구축

2.2.3 보안 모니터링 도구를 사용하여 모델의 비정상적인 활동을 감지하고, 실시간으로 대응할 수 있는 체계를 구축하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 서비스 제공기업이 보안 모니터링 도구를 사용하여 모델의 비정상적인 활동을 감지하고, 실시간 대응 체계를 구축하는 것은 모델의 무결성 보호, 데이터 기밀성과 안전성 보장 측면에서 중요함
- (예시) 모니터링 요구사항 정의
  - 보안 목표 설정: 모델의 무결성·가용성·기밀성을 보호하기 위해 필요한 주요 보안 목표를 정의
  - 모니터링 대상 식별: 모델의 입력 데이터, 출력 결과, API 호출, 리소스 사용량, 데이터 접근 기록 등 감시할 요소를 명확히 함
  - 위협 시나리오 정의: 예상 가능한 보안 위협(예: 적대적 공격, 비인가 접근, 과도한 API 호출)과 이에 대응할 방법을 문서화
- (예시) 이상 탐지 시스템 구축
  - 이상 탐지 모델 개발: 머신러닝 기반 이상 탐지 모델을 사용하여 비정상적인 입력 패턴, 출력 이상, 비인가 접근 탐지
  - 행동 분석: 사용자의 접근 패턴, API 호출 빈도, 모델 출력 결과의 통계적 변화를 감지
  - 임계값 설정: 비정상적인 활동을 판단할 임계값을 설정하고 과도한 알림을 방지
- (예시) 주기적인 테스트 및 개선
  - 취약점 점검: 보안 모니터링 시스템과 대응 프로세스의 취약점을 주기적으로 점검
  - 모의훈련: 비정상 활동 탐지와 대응 시뮬레이션을 통해 체계의 효과성을 검증
  - AI 기반 개선: 이상 탐지 모델을 지속적으로 학습 및 개선하여 새로운 위협에 대응

2.3 데이터 보안

AI 개발자, AI 서비스 제공자 공통사항

- AI 서비스 개발 또는 운영 과정에서 의도적으로 학습 데이터를 변질시키거나 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 공격에 노출될 수 있으므로, 이를 대처할 방안을 검토 및 적용하는 것이 바람직함

데이터 공격 기법 (예시)

공격기법	공격기법 내용
데이터 중독 공격 (poisoning attack)	AI 서비스는 일반적으로 입력 데이터 분포의 변화에 적응하기 위해 모델 배치 후 수집된 새로운 데이터를 사용해 재교육 됨. 이때, 공격자는 세심하게 조작된(perturbed) 데이터를 주입하여 서비스의 정상적인 기능을 손상시키는 방식으로 학습 데이터를 오염시킬 수 있음
회피공격 (evasion attack)	공격자는 학습 모델이 입력을 올바르게 식별할 수 없도록 기존의 입력 데이터에 대해 미묘한 차이의 노이즈를 추가하여 조작된 입력 데이터를 생성함. 이러한 변화는 사람의 눈에 잘 띄지 않지만, 심층학습 모델의 추론 결과에 큰 영향을 미침

2.3.1 적대적 공격 등 데이터 공격에 대한 방어 수단을 강구하고 있는가?

YES ☐ NO ☐ N/A ☐

- 적대적 공격을 방어하고 AI 서비스의 강건성을 높이기 위한 다양한 방어 기법이 존재함. 특히 데이터 수집 및 준비 단계에서의 회피 공격과 중독 공격 방어를 위한 대표적 기법으로는 적대적 학습(adversarial training), Gradient Masking, Feature Squeezing 등이 있음

적대적 공격에 대한 방어 기법

방어기법	방어기법 내용
적대적 학습 (adversarial training)	모델을 학습시킬 때, 적대적 사례로 활용할 수 있는 모든 경우의 수를 미리 고려하여 학습 데이터셋에 포함시키는 방법. 충분한 수와 다양성이 보장된 적대적 데이터를 생성하는 과정 없이는 그 성능을 보장할 수 없음
Gradient Masking (Distillation)	대부분의 공격은 모델 추론 과정에서의 경사(gradient)를 보고 이루어지므로 학습 모델의 경사가 그대로 노출되는 것을 방지하거나 gradient masking, 정규화 방법 등을 통해 경사가 두드러지지 않게 하여 적대적 공격에 방어할 수 있는 방법(distillation)들이 제안됨
Feature Squeezing	본래의 학습 모델과 별도로, 주어진 입력이 적대적 사례인지 아닌지를 판단하는 학습 모델을 추가하는 방법. 그 외에 다수의 학습 모델을 조합하여 시스템을 구성하면 특정 모델에 대한 화이트박스 공격을 피할 수 있으며, 특정 모델에 적용되는 적대적 공격이 불가능해짐

2.3.2 데이터 저장 및 전송 시 무결성을 보호하기 위한 조치를 하고 있는가? YES ☐ NO ☐ N/A ☐

- 데이터 암호화와 같은 보호 기법을 설계 단계에서 도입하고, 데이터 저장 및 전송 시 무결성을 유지할 수 있도록 함
- (예시) 해시(Hash) 및 체크섬 사용
  - 데이터 무결성 검증: 데이터 저장 및 전송 전후에 해시 함수(SHA-256 등)를 사용해 데이터의 무결성을 확인
  - 체크섬(Checksum) 생성: 데이터 전송 시 체크섬을 생성하고, 수신 측에서 이를 검증하여 데이터 변조 여부를 확인
- (예시) 디지털 서명 및 인증
  - 디지털 서명: 데이터에 디지털 서명을 추가하여 무결성을 보장하고, 출처를 확인할 수 있도록 함
  - 인증서 사용: 전송 과정에서 SSL/TLS 인증서를 사용하여 데이터 송수신 주체의 신뢰성을 검증

2.3.3 중요 데이터에 대한 기밀성 유지를 위해 보호 방안을 마련하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 서비스를 제공할 때, 중요한 데이터를 보호하기 위한 방안을 마련하는 것은 매우 중요한 것으로 아래와 같은 방안들이 적용될 필요가 있음

  (1) 데이터 수집 및 사용 정책 수립
  - 목적 명시: 데이터를 수집할 때, 왜 데이터를 수집하는지 명확히 하고 사용 목적을 제한
  - 최소 수집 원칙: 서비스 운영에 꼭 필요한 최소한의 데이터만 수집

  (2) 데이터 암호화
  - 저장 중 암호화: 민감한 데이터를 저장할 때 AES와 같은 강력한 암호화 알고리즘을 사용
  - 키 관리: 암호화 키를 안전하게 저장하고 관리

  (3) 접근 제어
  - 권한 관리: 데이터에 접근할 수 있는 사용자를 최소화하고 역할 기반 접근 제어(RBAC)를 구현
  - 로그인 보안: 다단계 인증(MFA) 및 강력한 비밀번호 정책을 적용
  - 접근 기록: 데이터를 접근하거나 수정한 기록을 상세히 로그로 남김

2.3.4 전송구간에서 중요정보 유출을 방지하기 위한 보호 방안을 마련하고 있는가? YES ☐ NO ☐ N/A ☐

- 전송 구간에서 중요정보 유출을 방지하는 것은 데이터의 기밀성과 무결성을 보호하여 민감 정보가 중간에서 가로채지거나 조작되지 않도록 하는 데 필수적임. 안전한 데이터 전송은 AI 서비스의 안정성과 지속 가능성을 확보하는 핵심 요소임
- (예시) 네트워크 보안 강화
  - 방화벽 및 IDS/IPS: 방화벽과 침입 탐지 및 방지 시스템(IDS/IPS)을 사용하여 네트워크로부터의 비인가 접근을 차단
  - VPN 및 전용 네트워크: 데이터 전송 시 가상 사설망(VPN) 또는 전용 네트워크를 사용하여 보안을 강화
  - DLP(Data Loss Prevention) 솔루션 적용: 중요정보가 외부로 유출되지 않도록 모니터링하고, 중요 데이터를 전송하려는 시도를 차단
- (예시) 암호화 기술 적용
  - TLS/SSL 사용: TLS 1.2 이상 프로토콜을 사용하여 데이터 전송 구간에서 암호화를 보장
  - 강력한 암호화 알고리즘: AES-256, RSA 등 강력한 암호화 알고리즘을 적용하여 데이터를 보호
  - 엔드투엔드 암호화(End-to-End Encryption): 데이터가 송신자에서 수신자까지 암호화된 상태로 유지되도록 보장
- (예시) 키 관리 강화
  - 보안 키 교환: 전송 중 안전한 암호화 키 교환을 위해 Diffie-Hellman 또는 Elliptic Curve Cryptography(ECC)와 같은 보안 키 교환 방식을 사용
  - 중앙화된 키 관리: 키 관리 시스템(KMS)을 사용하여 암호화 키를 안전하게 생성, 저장 및 관리
  - 키 교체 주기: 암호화 키를 정기적으로 회전하여 보안성을 유지

2.3.5 데이터 유출시 책임추적을 할 수 있도록 조치를 하고 있는가? YES ☐ NO ☐ N/A ☐

- 데이터 유출 시 발생 원인과 책임자를 효과적으로 추적하고, 데이터 보안 체계를 강화하기 위해서는 접근 및 활동 로그 관리, 데이터 태그 및 분류 활동 등이 필요함
- (예시) 접근 및 활동 로그 관리
  - 접근 기록 저장: 데이터에 접근한 모든 사용자와 시스템의 활동을 기록하고, 로그에는 사용자 ID, IP 주소, 시간, 작업 내용을 포함
  - 로그 중앙화: 중앙화된 로그 관리 시스템(SIEM)을 도입하여 모든 로그를 한곳에서 관리하고 분석 가능하도록 함
  - 로그 보존 기간: 법적 요구사항과 내부 정책에 따라 로그 데이터를 일정 기간 보관(예: 1~5년)
- (예시) 데이터 태그 및 분류
  - 데이터 태그 지정: 각 데이터에 식별 태그를 부여하여 데이터의 출처와 사용 이력을 추적할 수 있도록 함
  - 중요도에 따른 분류: 데이터를 민감도와 중요도에 따라 분류하고, 민감 데이터에 대해 더 강력한 보안 조치를 적용

2.4 API 및 인터페이스 보안

- API 및 인터페이스 보안은 AI 서비스 제공기업이 데이터의 기밀성과 무결성을 보호하고, 민감한 정보가 무단 접근이나 변조로부터 안전하게 유지되도록 보장하기 위해 필수적임. API를 통한 비인가 접근은 데이터 유출, 서비스 오작동, 또는 고객 신뢰 손실로 이어질 수 있음.
- 또한, 적대적 공격(Adversarial Attack)이나 과도한 요청(DoS/DDoS 공격)은 모델 성능을 저하시켜 서비스 가용성을 위협할 수 있음. 보안이 취약한 인터페이스는 해커가 시스템의 내부를 분석하고 추가적인 공격을 시도할 수 있는 진입점이 됨. 따라서 API 및 인터페이스 보안은 법적 규제 준수, 서비스 품질 유지 등을 위해 필요함

2.4.1 API 통신을 암호화하여 데이터가 전송되는 구간에서 외부 공격에 대한 방어를 하고 있는가?

YES ☐ NO ☐ N/A ☐

- API가 공개되어 있으면 공격자가 무단으로 모델에 접근하여 악의적인 요청을 보낼 수 있음.
- API 통신을 암호화하여 데이터가 전송되는 동안 중간자 공격을 방지하고, 모든 트래픽에 대해 암호화(TLS/SSL)를 적용하는 것이 필요함
- (예시) HTTPS 프로토콜 사용
  - TLS/SSL 적용: API 통신에 HTTPS를 적용하여 데이터 전송 구간을 암호화
  - TLS 1.2 이상(권장 TLS 1.3)을 사용하여 최신 보안 표준을 준수
  - SSL 인증서 관리: 신뢰할 수 있는 인증기관(CA)에서 발급한 인증서를 사용하고, 만료 전에 갱신
- (예시) API 인증 및 권한 관리
  - 강력한 인증 프로토콜: OAuth2, OpenID Connect와 같은 표준 인증 프로토콜을 사용
  - API 키 및 토큰 보안: API 호출 시 고유 키 또는 JWT(JSON Web Token)를 요구하며, 민감 정보는 암호화된 환경 변수에 저장
  - 권한 최소화: 사용자 및 애플리케이션에 필요한 최소 권한만 부여하고, API 호출 범위를 제한

2.4.2 모든 API 요청에 대해 인증 및 권한 관리를 강화하고, 중요 데이터에 접근할 때는 강한 인증 메커니즘을 적용하고 있는가?

YES ☐ NO ☐ N/A ☐

- API는 인터넷을 통해 접근이 가능하므로, 인증 및 권한 관리가 취약하면 데이터 유출, 무단 접근, 악의적인 사용이 발생할 위험이 높아짐
- 보안이 취약한 서비스는 사용자 신뢰를 잃게 되고, 결국 브랜드 이미지 실추 및 고객 이탈로 이어질 가능성이 큼. 특히, 기업 고객(B2B)의 경우 보안이 미흡한 API 서비스는 고객이 사용을 기피할 가능성이 큼
- 또한 내부 직원이나 협력업체가 불필요한 데이터 접근 권한을 가지면 내부 유출 위험이 증가할 수 있으므로, 로그 및 감사(Audit) 기능을 추가하면 누가, 언제, 어떤 데이터에 접근했는지 추적 가능하여 보안 사고 대응이 가능함
- 이에 대한 예방책은 다음과 같음
  - 사용자 인증을 강화하기 위해 다중 인증(MFA)을 도입하고, 모든 중요한 서비스에 대해 강력한 인증 절차를 적용함
  - 최소 권한 원칙을 적용하여 사용자가 필요한 최소한의 권한만 부여받도록 하고, 불필요한 권한이 부여되지 않도록 관리함
  - 사용자 활동을 실시간으로 모니터링하고, 비정상적인 접근 시도를 자동으로 차단할 수 있는 침입 탐지 시스템(IDS)을 구축함

2.4.3 API 트래픽은 암호화 기술을 사용하여 보호하고, 데이터를 안전하게 주고받도록 보장하고 있는가?

YES ☐ NO ☐ N/A ☐

- API 트래픽이 암호화되지 않으면 중간자 공격(Man-in-the-Middle Attack)에 노출될 수 있음. API 트래픽은 암호화(TLS)를 사용하여 보호하고, 데이터를 안전하게 주고받도록 보장함
- (예시) 강력한 암호화 알고리즘 사용
  - 암호화 표준: AES-256, RSA-2048과 같은 강력한 암호화 알고리즘을 사용하여 데이터 기밀성을 유지
  - Perfect Forward Secrecy(PFS): 세션 키가 유출되더라도 과거 통신 데이터가 복호화되지 않도록 PFS를 활성
  - 메시지 인증 코드(MAC): HMAC-SHA256을 사용하여 데이터 무결성을 보장하고, 데이터 변조를 방지
- (예시) API 요청 및 응답 보호 방안
  - 토큰 기반 인증: OAuth2, JWT(JSON Web Token)와 같은 표준 인증 방식을 사용하여 API 요청과 응답을 보호
  - 권한 범위 설정: API 호출의 권한 범위를 명확히 정의하여 민감한 데이터 접근을 제한
  - 전송 데이터 검증: 요청 데이터와 응답 데이터를 해시 또는 체크섬으로 검증하여 무결성을 확인

2.4.4 API 호출 제한(Rate Limiting)을 설정하여 과도한 요청을 방지하고, 비정상적인 요청 패턴을 탐지하여 차단하고 있는가?

YES ☐ NO ☐ N/A ☐

- API 호출을 실시간으로 모니터링하고 비정상적인 패턴이나 오용 시도를 감지하려면 로그 관리, AI 기반 이상 탐지, Rate Limiting, WAF, 실시간 알림 등 다층적인 보안 전략이 필요함. 이를 통해 API의 안정성과 보안을 유지하고, 불법 접근이나 오용으로부터 시스템을 보호할 수 있음
- API 호출 제한(rate limit)이란 분 당 호출할 수 있는 API 호출 횟수를 의미함
- API에 과도한 요청이 발생할 경우 “서비스 거부 공격(DoS)”의 위험이 있음. 따라서 이 경우에는 “API 호출 제한(Rate Limiting)”을 설정하여 과도한 요청을 차단하고, 비정상적인 요청 패턴을 탐지하는 시스템을 구축함
- (예시) Rate Limiting 정책 설계
  - 요청 한도 정의: API 호출 빈도와 허용 한도를 설정(예: 사용자당 1분에 1000회 호출 제한)
  - 기간 기준 설정: 호출 빈도 제한을 적용할 기간을 설정(예: 초, 분, 시간, 일 단위)
  - 사용자 기반 제한: 사용자별, IP별, 또는 API 키별로 개별 호출 제한 정책을 설정
- (예시) Rate Limiting 구현
  - 정적 및 동적 설정: API 사용량에 따라 정적 또는 동적으로 Rate Limiting 값을 조정할 수 있는 기능을 구현

03 서비스 제공 및 운영

3.1 로그 및 운영 데이터 보안

- AI 서비스 운영 중 발생하는 로그 및 운영 데이터는 중요한 정보로서, 이를 안전하게 관리하지 않으면 공격자가 이를 분석하여 시스템의 구조와 취약점을 파악할 수 있음

3.1.1 데이터 처리 중 접속로그 관리를 강화하고 있는가? YES ☐ NO ☐ N/A ☐

- AI 서비스 운영 중 발생하는 로그 및 운영 데이터는 특히 Generative AI 서비스 제공시 모델의 품질 관리 및 성능 최적화, 오류 추적 및 문제 해결, Generative AI 특유의 문제 해결 등을 위해 매우 중요함
- 모델의 품질 관리 및 성능 최적화
  - 결과물 검증: Generative AI는 예측 가능한 정답보다 창의적이고 다양한 응답을 생성하므로, 로그를 통해 출력 결과물이 유효한지 확인할 필요가 있음
  - 피드백 기반 학습: 운영 데이터를 통해 잘못된 응답, 비효율적인 생성 결과 등을 분석하고 이를 모델 개선에 반영할 수 있음
- 오류 추적 및 문제 해결
  - 오류 원인 분석: 비정상적 응답, 시스템 충돌, 사용자 불만족을 유발하는 문제를 로그 데이터를 통해 빠르게 식별할 수 있음
  - 실시간 대응: 생성형 AI의 예기치 않은 오류나 비정상적 응답을 감지하고, 운영 데이터 기반으로 즉각적인 대응이 가능
- Generative AI 특유의 문제 해결
  - 출력 다양성 및 품질 분석: Generative AI는 출력이 다양하고 예측하기 어려운 경우가 많아, 로그 데이터를 통해 생성 결과의 품질과 일관성을 평가할 수 있음
  - 안전성 확보: 생성된 응답이 유해하거나 부적절한 경우를 모니터링하고, 로그 데이터를 기반으로 이를 예방하는 메커니즘을 설계할 수 있음
- (예시) 접속 로그 수집 및 저장
  - 접속 기록 범위 정의: 사용자 ID, IP 주소, 접근 시간, 요청 URL, HTTP 상태 코드, 처리 결과 등 접속 관련 주요 정보를 기록
  - 중앙 집중식 로그 관리: 로그 데이터를 중앙에서 관리할 수 있는 통합 로그 관리 시스템(SIEM)을 구축
  - 안전한 로그 저장: 로그 데이터를 암호화하고 안전한 위치에 저장하여 무단 액세스 및 변조를 방지
- (예시) 로그 보존 및 삭제 정책
  - 보존 기간 설정: 법적 규제 및 기업 내부 정책에 따라 로그 데이터를 일정 기간(예: 1~5년) 보관
  - 안전한 삭제: 로그 데이터를 보존 기간 이후 안전하게 삭제하여 데이터 유출 위험을 줄임

3.1.2 로그 파일 및 운영 데이터에 암호화를 적용하고, 중요정보는 별도로 관리하여 유출을 방지하고 있는가? YES ☐ NO ☐ N/A ☐

- Generative AI 서비스 제공 시 로그 파일 및 운영 데이터에 암호화를 적용하고 중요 정보를 별도로 관리하여 유출을 방지하는 것은 중요 데이터 보호, 생성 결과 및 학습 데이터 보호, 악의적인 행위 방지 등을 위해 필요함
- 중요 데이터 보호, 생성 결과 및 학습 데이터 보호
  - 기밀 유지: 생성형 AI는 종종 비즈니스 기밀, 제품 설계, 혹은 연구 데이터를 처리함. 이런 결과물이나 입력 데이터가 로그에 저장될 경우, 유출 시 경쟁력 손실로 이어질 수 있음
  - 지적 재산권 보호: AI 모델이 생성한 결과물은 조직의 자산일 수 있으므로, 로그에 저장된 관련 데이터를 보호하는 것이 중요함
- Generative AI의 특성으로 인한 보안 리스크
  - 데이터 재학습 가능성: Generative AI는 학습 과정에서 중요한 정보를 학습할 가능성이 있음. 로그 데이터가 유출되면 공격자는 이를 분석해 모델의 학습 데이터나 사용자 데이터를 추론할 수 있음
  - 출력물 조작: 로그에 저장된 데이터가 유출되거나 변조될 경우, 생성 결과물의 무결성을 보장할 수 없게 되어 서비스 신뢰도가 하락할 수 있음
- 악의적인 행위 방지
  - 리버스 엔지니어링 방지: 로그와 운영 데이터를 암호화하지 않을 경우, 악의적인 사용자가 데이터를 분석해 모델 동작 방식을 역으로 추적하거나 공격 벡터를 설계할 수 있음
  - 서비스 악용 방지: 로그 데이터를 통해 AI 모델의 취약점(예: 특정 입력 패턴에서 부적절한 응답)이 노출되면 이를 악용한 공격(예: 모델 중독, 데이터 중독)이 발생할 수 있음
- 운영 효율성과 데이터 접근 제어
  - 중요 정보 별도 관리: 운영 데이터 중 중요 정보(예: API 키, 인증 토큰, 암호화 키)는 별도로 관리하여 로그에서 제거하면, 데이터 유출 사고 발생 시 피해 범위를 최소화할 수 있음
  - 접근 통제 강화: 로그 데이터를 암호화하고 중요 정보를 분리 관리하면, 내부 관계자나 시스템의 비인가 접근으로 인한 데이터 유출 가능성을 줄일 수 있음

3.1.3 운영 중 발생하는 데이터를 안전하게 저장하고, 접근 제어를 통해 인증된 관리자만 로그에 접근할 수 있도록 설정하고 있는가?

YES ☐ NO ☐ N/A ☐

- Generative AI 서비스 제공 시 운영 중 발생하는 데이터를 안전하게 저장하고 접근 제어를 통해 인증된 관리자만이 로그에 접근하도록 설정하는 것은 생성된 데이터의 기밀성 보장, 보안 사고 예방 및 책임 분산 방지 등을 위해 필요함
- 생성된 데이터의 기밀성 보장
  - 기밀 유지: Generative AI가 생성한 결과물에는 종종 비즈니스 기밀 또는 사용자 특정 데이터가 포함될 수 있으므로, 로그에 저장된 생성 데이터를 보호하지 않으면 외부로 유출될 가능성이 있음
  - 데이터 오용 방지: 인증되지 않은 사용자가 로그에 접근하면 생성 데이터나 사용 데이터를 악의적으로 조작, 재사용, 유출할 수 있음
  - 사용 기록: 로그에는 사용자가 어떤 요청을 했는지, 어떤 결과를 생성했는지와 같은 정보가 포함됨. 이러한 정보는 사용자 행동 분석에 악용될 수 있으므로 철저한 접근 통제가 필요함
- 보안 사고 예방 및 책임 분산 방지
  - 내부 위협 관리: 내부 관계자가 로그에 접근해 데이터를 유출하거나 악용하는 일이 발생하지 않도록, 인증된 관리자만 접근하도록 설정해야 함
  - 무결성 보장: 접근 통제를 통해 로그 데이터를 보호함으로써, 데이터가 의도치 않게 변경되거나 삭제되지 않도록 보장할 수 있음
- (예시) 접근 제어 강화
  - 역할 기반 접근 제어(RBAC): 사용자 역할별로 데이터 접근 권한을 부여하고, 최소 권한 원칙을 준수
  - 다단계 인증(MFA): 로그 데이터에 접근하려면 ID/비밀번호와 추가 인증(예: OTP, 인증 앱)을 요구하여 보안을 강화
  - 접근 로그 기록: 로그 데이터에 접근한 모든 사용자와 작업 내역을 기록하여 추후 감사와 분석이 가능하도록 함
- (예시) 접근 요청 인증
  - 네트워크 기반 인증: VPN 사용, IP 화이트리스트 정책 적용
  - 암호화된 통신 TLS(Transport Layer Security)를 적용하여 로그 데이터 전송 중 기밀성과 무결성을 유지

3.1.4 로그 데이터 접근 권한을 최소화하고, 접근 제어 및 사용자 활동기록을 통해 비정상적인 접근을 탐지하고 있는가?

YES ☐ NO ☐ N/A ☐

- Predictive AI 서비스 제공 시 로그 데이터 접근 권한을 최소화하고, 접근 제어 및 사용자 활동 기록을 통해 비정상적인 접근을 탐지하는 것은 모델 신뢰성과 무결성 유지, 비즈니스 기밀 보호, 비정상적 접근 탐지를 통한 보안 강화 등을 위해 필요함
- 모델 신뢰성과 무결성 유지
  - 데이터 조작 방지: 로그에 비인가 접근이 발생하면, 모델의 입력 데이터나 예측 결과가 조작될 가능성이 있음. 이는 모델의 신뢰성을 손상시킬 수 있음
  - 운영 데이터 보호: Predictive AI의 운영 데이터는 모델 성능 평가와 개선에 사용되므로, 이를 안전하게 보호해야 모델의 무결성을 유지할 수 있음
- 비즈니스 기밀 보호
  - 모델 및 예측 알고리즘 보호: Predictive AI의 핵심 경쟁력은 예측 알고리즘과 데이터를 분석해 도출한 비즈니스 통찰임. 로그 데이터에 비즈니스 기밀이 포함될 경우, 권한 없는 접근은 경쟁력 손실로 이어질 수 있음
- 비정상적 접근 탐지를 통한 보안 강화
  - 내부 위협 관리: 내부 사용자가 악의적이거나 부주의한 행동으로 로그 데이터를 유출하거나 변경하는 것을 방지할 수 있음
  - 외부 공격 방어: 비인가 접근 시도를 탐지하고, 이를 통해 데이터 도난이나 서비스 악용을 조기에 차단할 수 있음
  - 운영 이상 감지: 비정상적인 접근 패턴은 시스템 취약점이나 보안 위협의 신호일 수 있으므로 이를 탐지하면 사전 대응이 가능함
- Predictive AI 서비스는 데이터의 민감성과 모델의 의존성 때문에 로그 데이터의 안전한 저장과 접근 통제가 필수적임. 접근 권한 최소화와 사용자 활동 기록은 데이터 보호, 보안 강화, 규제 준수, 비즈니스 신뢰 확보를 가능하게 하며, Predictive AI 서비스의 안정성과 지속 가능성을 보장하는 핵심 요소임

04 서비스 유지보수 및 지원

4.1 모니터링, 업데이트 및 패치

AI 개발자, AI 서비스 제공자 공통사항

- AI 서비스 제공자는 AI 활용 과정을 통해 AI 기능을 향상시키고 위험을 억제하기 위해 소프트웨어 업데이트, 검사 및 수리 등을 제공하여야 함. 특히 업데이트를 통해 다른 AI 연동에 영향을 미칠 것으로 예상되는 경우 위험에 대한 정보를 제공해야 함

4.1.1 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집·관리하고 있는가? YES ☐ NO ☐ N/A ☐

- 서비스 이용 로그 분석은 서비스 운영 상태에 관한 확인뿐만 아니라, 사용자가 겪는 문제가 무엇인지 확인할 수 있는 가장 기본적인 방법이 될 수 있음. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있음
- 서버 인프라에 대한 로그를 통해 서비스 운영 상태에 대한 모니터링을 수행할 수 있으며, 사용자 상호 작용 로그는 사용자가 어떤 서비스를 많이 이용하고 어떤 서비스에서 오류를 겪는지 분석할 수 있음. 이를 위해 인프라 관점에서는 로그 분석 소프트웨어를 활용할 수 있으며, 사용자 관점에서는 기업이 자체적으로 인터페이스 또는 상호작용의 호출에 따른 로그를 수집하거나 로그 분석 도구를 활용할 수 있음
- Generative AI 서비스 제공 관점: 사용자 로그를 수집·관리해야 하는 이유
  - 사용자 피드백 반영 및 서비스 개선
    - ▸생성 결과 품질 개선: Generative AI는 텍스트, 이미지, 코드 등 다양한 유형의 콘텐츠를 생성함. 사용자 로그를 통해 생성된 결과물의 품질, 적절성, 유용성을 평가하고 개선할 수 있음
    - ▸오류 수정: 로그 데이터를 활용해 잘못된 응답이나 비정상적인 동작을 추적하고, 모델과 서비스의 정확성을 개선할 수 있음
  - 유해 콘텐츠 모니터링 및 방지
    - ▸부적절한 콘텐츠 생성 감지: 로그 데이터를 통해 모델이 유해하거나 부적절한 콘텐츠를 생성한 사례를 식별하고, 이를 방지하는 데 필요한 개선 조치를 취할 수 있음
    - ▸안전성 확보: 지속적인 로그 모니터링은 생성 결과가 윤리적이고 규제 요구를 충족하도록 보장하는 데 필수적임
- Predictive AI 서비스 제공 관점: 사용자 로그를 수집·관리해야 하는 이유
  - 예측 정확성 향상
    - ▸모델 성능 평가: 사용자 로그는 Predictive AI 모델의 예측 결과와 실제 결과를 비교하고, 모델의 정확성을 평가하는 데 핵심 자료로 사용됨
    - ▸데이터 드리프트 감지: 사용자 로그를 통해 입력 데이터의 특성 변화를 모니터링하고, 데이터 드리프트나 개념 드리프트를 감지하여 모델 업데이트가 필요한 시점을 파악할 수 있음
    - ▸오류 수정 및 학습: 부정확한 예측 결과를 식별하고, 이를 통해 모델 재훈련 및 성능 개선을

수행할 수 있음

- 비즈니스 리스크 관리
  - ▸부정확한 결과 방지: 사용자 로그는 Predictive AI 모델의 출력물이 적절했는지 확인하고, 잘못된 결과로 인해 발생할 수 있는 비즈니스 리스크를 방지하는 데 도움을 줌
  - ▸서비스 품질 유지: 지속적인 모니터링은 모델이 비즈니스 목표를 지원하는 데 필요한 성능과 안정성을 유지하도록 보장함

4.1.2 AI 시스템이 정상적으로 작동하지 않거나 예기치 못한 오류가 발생할 경우 이를 조기에 탐지하고 대응하는 메커니즘이 있는가?

YES ☐ NO ☐ N/A ☐

• AI 시스템에 오류 등 탐지 및 대응 메카니즘

구분	설명	예시
시스템 모니터링	시스템 로그를 주기적으로 모니터링하여 비정상적인 활동을 탐지하고 대응함	시스템에 최신 보안 업데이트를 정기적으로 적용하고, 로그 모니터링 도구를 사용하여 시스템 활동을 실시간으로 감시
이상 징후 탐지 (Anomaly Detection)	정상적인 패턴에서 벗어난 이상 징후를 자동으로 감지할 수 있는 알고리즘과 기술을 도입	머신러닝 기반 이상 탐지 알고리즘, 실시간 경고 시스템
자동화된 경고 (Automated Alerts)	이상 징후나 위험 요소 감지 시 즉각적으로 관련 담당자에게 경고를 보내는 자동화된 시스템을 운영함	이메일 알림, SMS 경고, 대시보드 알림 등
정기적 검토 (Regular Reviews)	모니터링 결과와 경고 로그를 정기적으로 검토하여, 새로운 위험 요소나 패턴을 식별하고 대응 방안을 업데이트함	주간/월간 보고서 작성, 경고 로그 분석 회의

4.1.3 AI 시스템의 보안 패치나 모델 업데이트가 정기적으로 이루어지고 있는가? YES ☐ NO ☐ N/A ☐

- 보안 취약점이 발견되면 신속하게 패치를 적용하고, 정기적인 보안 업데이트를 통해 시스템의 보안 상태를 최신으로 유지해야 함
- Generative AI 서비스 제공 시 AI 시스템에 대한 보안 패치 및 모델 업데이트 확인이 중요한 이유
  - 유해 콘텐츠 및 부적절한 응답 방지
    - ▸ 윤리적 리스크 관리: Generative AI는 새로운 입력에 따라 다양한 결과를 생성함. 모델 업데이트가 이루어지지 않으면, 유해 콘텐츠, 편향된 결과, 허위 정보가 생성될 가능성이 커질 수 있음
    - ▸ 사용자 요구 반영: 정기적인 모델 업데이트는 사용자의 피드백을 반영하여 생성 결과의 품질과 적절성을 개선할 수 있음
  - 보안 취약점 예방
    - ▸ 사이버 공격 방어: Generative AI의 API나 인터페이스는 공격자가 악용하려는 주요 표적임. 보안 패치가 없으면 SQL 삽입, API 남용, 악의적인 입력 패턴에 취약해질 수 있음
    - ▸ 데이터 유출 방지: 모델 업데이트를 통해 훈련 데이터에서 학습된 중요한 정보가 노출되지 않도록 최신 보안 조치를 적용해야 함
- Predictive AI 서비스 제공 시 AI 시스템에 대한 보안 패치 및 모델 업데이트 확인이 중요한 이유
  - 데이터 특성 변화 대응
    - ▸ 모델 노후화 방지: Predictive AI는 입력 데이터의 패턴을 기반으로 예측을 수행함. 시간이 지나면 데이터의 특성이 변할 수 있으며, 정기적인 모델 업데이트가 이루어지지 않으면 예측 정확도가 떨어질 위험이 있음
    - ▸ 드리프트(Drift) 관리: 데이터 드리프트(특성 변화)나 개념 드리프트(관계 변화)에 대응하기 위해 최신 데이터를 반영한 재훈련과 업데이트가 필요함
  - 보안 취약점 완화
    - ▸ 모델 보안 강화: Predictive AI도 Generative AI와 마찬가지로 공격의 대상이 될 수 있음. 예를 들어, 데이터 중독(Data Poisoning) 공격은 모델이 잘못된 예측을 하도록 유도할 수 있으므로, 보안 패치와 업데이트는 이를 예방하는 데 필수적임
    - ▸ 데이터 보호: Predictive AI는 종종 민감한 정보나 금융 데이터를 처리하므로, 보안 취약점이 있으면 데이터 유출로 이어질 가능성이 높음

4.2 성능 및 장애 관리

- 모델 성능 저하가 감지되면 재학습을 통해 성능을 유지하고, 관련 시스템을 구축하여 장애 발생 시에도 서비스 가용성을 유지할 수 있도록 함

4.2.1 서비스 장애가 발생했을 때 자동으로 복구할 수 있도록 하고 있는가? YES ☐ NO ☐ N/A ☐

- 자동 복구 시스템을 구축하여 서비스 장애가 발생했을 때 자동으로 복구할 수 있도록 해야 함
- Generative AI 서비스 제공 시 장애가 발생했을 때 자동으로 복구하는 것이 중요한 이유
  - 서비스 장애 발생 시 영향
    - ▸ 사용자 경험: Generative AI는 실시간 상호작용이 중요한 서비스임. 예를 들어, 사용자가 질문을 입력했는데 응답이 늦거나 제공되지 않으면 서비스 신뢰도가 크게 하락할 수 있음
    - ▸ 창의성 및 생산성 저하: 콘텐츠 생성 중 장애가 발생하면 사용자 작업 흐름이 끊기거나 데이터가 손실될 수 있음
    - ▸ 실시간 응답 요구: 특히 실시간 응답이 중요한 고객 지원, 콘텐츠 제작 도구에서의 장애는 즉각적인 문제 해결과 복구가 필요함
- Predictive AI 서비스 제공 시 장애가 발생했을 때 자동으로 복구하는 것이 중요한 이유
  - 서비스 장애 발생 시 영향
    - ▸ 의사결정 지연: Predictive AI는 중요한 비즈니스 의사결정이나 시스템 운영(예: 공장 설비 관리, 금융 거래)에 활용됨. 장애가 발생하면 의사결정이 지연되어 업무나 시스템에 직접적인 영향을 미칠 수 있음
    - ▸ 데이터 흐름 중단: 예측 모델은 실시간으로 데이터를 수집하고 처리하는 경우가 많음. 데이터 입력이 중단되거나 처리 지연이 발생하면 결과의 신뢰도와 정확도가 저하됨
    - ▸ 비용 증가: Predictive AI의 예측 실패는 물류, 생산, 운영 등에서 비용 증가와 손실로 이어질 수 있음

4.2.2 모델 성능을 지속적으로 모니터링하고, 성능 저하가 감지되면 재학습을 통해 성능을 유지하고 있는가?

YES ☐ NO ☐ N/A ☐

- 운영 중인 AI 모델은 시간이 지남에 따라 성능이 저하되거나, 데이터의 변화로 인해 훈련 데이터와 실시간 데이터 간에 차이가 발생하는 모델 드리프트(Model Drift) 현상이 발생할 수 있음.
  - 현실 적합성 부족: 시간이 지남에 따라 모델이 최신 정보나 사용자 요구를 반영하지 못해 구식이 되거나, 현실과 동떨어진 결과를 생성할 수 있음
  - 데이터 드리프트(Data Drift): 시간이 지남에 따라 데이터 분포가 변화하면, 모델이 학습했던 데이터와 실제 데이터 간의 불일치가 발생함
- 모델 성능을 지속적으로 모니터링하고, 성능 저하가 감지되면 재학습을 통해 성능을 유지해야 함
- Generative AI 서비스 제공 시 모델 성능 저하의 영향
  - 품질 저하: 생성된 콘텐츠의 품질(예: 문법 오류, 맥락 부적합, 비논리적 답변 등)이 낮아져 사용자 만족도 감소
  - 사용자 신뢰 손실: 사용자에게 정확하고 맥락에 맞는 정보를 제공하지 못하면 서비스 신뢰도 하락
- Predictive AI 서비스 제공 시 모델 성능 저하의 영향
  - 예측 정확도 감소: 모델이 실제 데이터 분포를 제대로 반영하지 못하면 잘못된 예측을 제공할 수 있음.(예: 행동 예측 오류로 마케팅 비용 낭비, 재고 부족 또는 과잉 발생)
  - 비즈니스 손실: 예측 실패로 인해 물류, 금융, 운영 등에서 직접적인 손실이 발생할 수 있음

4.2.3 실시간으로 모델 드리프트 탐지 시스템을 운영하여 모델 성능이 저하될 경우 즉시 대응할 수 있는 체계를 마련하고 있는가?

YES ☐ NO ☐ N/A ☐

- 실시간 데이터와 훈련 데이터 간의 차이로 인해 모델의 예측 성능이 떨어져 보안 취약점이 발생할 수 있음. 이 경우에는 실시간으로 모델 드리프트 탐지 시스템을 운영하여 모델 성능이 변할 경우 즉시 대응할 수 있는 체계를 마련해야 함
- Generative AI 서비스 제공 시 모델 성능에 대한 실시간 탐지 및 대응의 중요성
  - 빠른 오류 수정: 드리프트를 실시간으로 탐지하면, 품질 저하가 사용자 경험에 영향을 미치기 전에 신속하게 대응 가능
  - 동적 콘텐츠 적응: 드리프트 탐지 시스템은 새로운 사용자 입력 패턴이나 트렌드를 인식하고 적응할 수 있도록 모델 업데이트를 지원
  - 서비스 연속성 보장: 탐지된 드리프트를 해결함으로써 지속적으로 일관성 있고 고품질의 콘텐츠를 제공할 수 있음
- Predictive AI 서비스 제공 시 모델 성능에 대한 실시간 탐지 및 대응의 중요성
  - 정확성 유지: 실시간 드리프트 탐지는 예측 결과의 정확성을 유지하고, 잘못된 결과를 방지
  - 의사결정 지원: 모델 성능 저하가 감지되면 즉시 조치를 취하여 비즈니스 의사결정이 잘못된 데이터를 기반으로 하지 않도록 보장
  - 비용 절감: 드리프트로 인한 손실을 조기에 방지하여 비즈니스 연속성과 수익성을 보호

4.2.4 정기적으로 모델 재학습 및 업데이트를 수행하여 새로운 데이터 패턴을 반영하고, 성능을 개선하고 있는가? YES ☐ NO ☐ N/A ☐

- 모델 성능이 저하되면 잘못된 의사결정을 내릴 수 있으며, 이를 공격자가 악용할 수 있음
- Generative AI 서비스 제공 시 정기적인 재학습이 중요한 이유
  - 민감 정보 제거: 과거 학습 데이터에서 모델이 민감하거나 기밀 정보를 학습한 경우, 정기적으로 업데이트하여 이러한 정보를 제거하고 오용 가능성을 줄일 수 있음
  - 보안 취약점 보완: 악성 입력(예: prompt injection 공격)이나 예측 가능한 응답 패턴을 탐지하고, 정기적인 업데이트로 이를 수정하여 악의적 사용을 방지
  - 콘텐츠 생성 안전성 강화: 업데이트를 통해 모델이 불법적이거나, 부적절한 콘텐츠를 생성하지 않도록 제어 및 필터링 성능을 개선
- Predictive AI 서비스 제공 시 정기적인 재학습이 중요한 이유
  - 새로운 위협 탐지: 사이버 보안, 사기 탐지와 같은 Predictive AI 서비스는 새로운 공격 패턴과 수법에 적응하기 위해 최신 데이터를 반영(예: 피싱 이메일의 텍스트 패턴 변화 또는 새로운 네트워크 공격 패턴 학습)
  - 오탐률 및 미탐률 감소: 오래된 모델은 위협 탐지에서 오탐률(False Positive) 또는 미탐률(False Negative)을 증가시켜 보안 문제를 유발할 수 있음. 정기적인 업데이트로 이를 줄일 수 있음
  - 보안 규제 준수: 정기적으로 모델을 재학습하면 최신 데이터 보안 규정을 준수하고, 데이터 사용 정책이 변경될 때 이를 반영할 수 있음

4.2.5 AI 서비스에 대해 다중화(HA) 및 백업 시스템을 구축하여 장애 발생 시에도 서비스가 연속적으로 제공될 수 있도록 하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 서비스는 항상 가용성을 유지해야 하지만, 서비스 운영 중 장애가 발생하거나 서비스가 과부하 또는 DoS/DDoS 공격으로 인해 시스템이 다운되면 고객에게 심각한 영향을 미칠 수 있음
- AI 서비스에 대해 다중화(HA) 및 백업 시스템을 구축하여 장애 발생 시에도 서비스가 연속적으로 제공될 수 있도록 해야 함
- Generative AI 서비스 제공 시 다중화 및 백업의 중요성
  - 서비스 연속성 보장: 다중화된 시스템은 장애가 발생해도 대체 서버나 리소스를 통해 서비스를 지속적으로 제공하므로, 데이터가 유실되거나 무단으로 노출될 위험을 줄임.(예: 실시간 대화형 서비스에서 장애가 발생하더라도 사용자 데이터를 안전하게 처리)
  - 데이터 보존 및 복구: 백업 시스템은 생성된 콘텐츠와 사용자 요청 데이터를 안전하게 저장하며, 장애 이후에도 원래 상태로 복구할 수 있도록 보장함. 데이터 손실로 인해 악의적 행위자가 시스템 내 취약점을 노리는 상황을 방지
  - 사이버 공격 방어: 장애 상황에서 서비스가 중단되면, 이를 노리는 DDoS 공격이나 데이터 유출 시도가 증가할 수 있음. 다중화는 이러한 공격에 대한 방어력을 강화함
- Predictive AI 서비스 제공 시 다중화 및 백업의 중요성
  - 실시간 보안 위협 대응: Predictive AI는 사이버 공격, 사기 탐지, 네트워크 이상 탐지 등 보안과 직접적으로 관련된 역할을 수행함. 다중화된 시스템은 장애 발생 시에도 실시간 대응을 유지해 보안 사고를 방지할 수 있음
  - 중단 없는 데이터 흐름 유지: 백업 시스템은 데이터 입력과 처리의 연속성을 보장하며, 장애로 인해 중요한 데이터를 놓치거나 보안 위협을 간과하는 문제를 방지함
  - 신뢰할 수 있는 복구 체계: Predictive AI 서비스는 장애 발생 시 신속하고 정확한 복구가 필수적임. 백업 시스템은 보안 위협 탐지와 같은 중요한 작업이 지속되도록 보장함

4.2.6 침입차단시스템 등을 통해 외부에서 발생하는 DoS/DDoS 공격을 방어하고, 실시간 모니터링 시스템을 운영하여 장애를 빠르게 감지하고 대응하고 있는가?

YES ☐ NO ☐ N/A ☐

- 운영 중 발생하는 장애에 대한 대응이 늦어지면 고객 불만이 발생하고, 비즈니스 중단으로 이어질 수 있음
- 침입 방지 및 차단 시스템을 통해 외부에서 발생하는 DoS/DDoS 공격을 방어하고, 실시간 모니터링 시스템을 운영하여 장애를 빠르게 감지하고 대응해야 함
- Generative AI 서비스 제공 시 DoS/DDoS 방어 및 실시간 모니터링의 중요성
  - 서비스 연속성 보장: 침입차단시스템과 같은 보안 장치를 통해 DoS/DDoS 공격을 방어하면, 서비스 중단 없이 사용자 요청을 처리할 수 있음.(예: 요청의 우선순위를 관리하거나 비정상 트래픽을 차단하여 정당한 사용자 요청이 처리되도록 보장함)
  - 사용자 데이터 보호: 실시간 모니터링을 통해 비정상 트래픽과 패턴을 빠르게 감지하여, 데이터 유출과 추가적인 공격 가능성을 방지할 수 있음
- Predictive AI 서비스 제공 시 DoS/DDoS 방어 및 실시간 모니터링의 중요성
  - 실시간 데이터 보호: DoS/DDoS 방어를 통해 실시간으로 들어오는 데이터를 안정적으로 처리하여 예측 결과의 신뢰성을 유지할 수 있음.(예: 정상적인 데이터가 차단되지 않고 지속적으로 모델에 공급될 수 있도록 보장)
  - 비즈니스 연속성 보장: Predictive AI의 의사결정 시스템이 중단되지 않도록 실시간 모니터링으로 장애를 빠르게 감지하고 대응할 수 있음

05 피드백 및 서비스 개선

5.1 사용자 피드백 관리

- 사용자 피드백, 성능 모니터링, 보안 로그, 성과 평가 등을 기반으로 AI 시스템을 개선하는 과정에서 발생할 수 있는 보안 취약점이 존재함. 이 단계는 AI 서비스의 품질과 성능을 지속적으로 향상시키는 중요한 과정이지만, 보안 관리가 소홀할 경우 공격자가 취약점을 악용할 수 있는 위험이 큼

5.1.1 사용자 피드백 시스템에 입력 검증 및 필터링을 적용하여 악성 코드나 비정상적인 데이터 입력을 차단하고 있는가?

YES ☐ NO ☐ N/A ☐

- 사용자 피드백을 수집하는 과정에서 공격자가 피드백 시스템을 악용해 악의적인 데이터를 입력하거나 시스템에 해를 가할 수 있음
  ※ 피드백 인젝션(Feedback Injection): 공격자가 악성 코드나 비정상적인 데이터를 식에 주입하여 시스템을 손상시킬 수 있음
- Generative AI 서비스 제공 시 사용자 피드백 입력 검증 및 필터링의 중요성
  - 악성 데이터 차단: 입력 데이터에 대한 검증 및 필터링을 통해 악성 코드나 의도적으로 조작된 데이터를 제거하여 시스템을 보호(예: 특수 문자를 제한하거나, 유효한 데이터 형식을 강제하여 악성 입력을 방지)
  - 유해 콘텐츠 생성 방지: 입력 검증으로 사용자가 악의적으로 특정 결과를 유도하는 공격을 차단하여 모델이 안전하고 신뢰할 수 있는 콘텐츠만 생성하도록 보장
- Predictive AI 서비스 제공 시 사용자 피드백 입력 검증 및 필터링의 중요성
  - 모델의 신뢰성 유지: 악의적 피드백 데이터가 모델 학습에 사용되지 않도록 필터링하여, 모델이 신뢰할 수 있는 데이터만 학습하게 함.(예: 비정상 데이터로 인해 발생할 수 있는 False Positive 또는 False Negative 결과 방지)
  - 보안 위협 차단: Predictive AI가 사이버 공격 탐지에 사용되는 경우, 입력 검증은 보안 위협 탐지의 신뢰성을 높이고 잘못된 경보를 방지함.(예: 위조된 위협 데이터를 차단하여 모델이 오탐지하지 않도록 보장)

5.1.2 사용자 피드백을 자동으로 분석하기 전에 사전 검증 절차를 마련하여 피드백 데이터의 무결성을 확인하고 있는가?

YES ☐ NO ☐ N/A ☐

- 자동화된 데이터 정제를 통해 유효하지 않은 피드백을 걸러내고, 거짓 피드백을 탐지할 수 있는 알고리즘을 도입
  ※ 거짓 피드백: 악의적인 사용자가 부정확하거나 거짓된 피드백을 제공해 모델 성능에 부정적인 영향을 미칠 수 있음
- Generative AI 서비스 제공 시 사용자 피드백에 대한 사전 검증 절차의 중요성
  - 악성 데이터 차단: 입력된 피드백 데이터를 자동 분석 전에 검증하여 악성 코드, 유해 콘텐츠, 또는 비정상 데이터를 사전에 차단 가능
  - 모델 안전성 강화: 검증 절차를 통해 모델에 전달되는 데이터의 무결성을 확보하여, 예상치 못한 결과나 오작동을 방지 가능
- Predictive AI 서비스 제공 시 사용자 피드백에 대한 사전 검증 절차의 중요성
  - 데이터 오염 방지: 피드백 데이터에 포함된 비정상적 데이터 또는 악의적 입력을 제거하여, 학습 데이터의 무결성을 유지(예: 이상치 탐지 및 데이터 유효성 검사를 통해 데이터 품질을 향상).
  - 모델 안정성 및 보안 강화: 예측 모델이 학습하는 데이터가 항상 신뢰할 수 있는 상태임을 보장하여, 악의적인 의도로 인한 보안 사고를 방지

5.1.3 최소 권한 원칙(Least Privilege)을 적용하여 사용자 피드백 및 개선 과정에서 접근할 수 있는 권한을 최소화하고 있는가?

YES ☐ NO ☐ N/A ☐

- 피드백을 반영하거나 성능을 개선하는 과정에서 접근 제어가 미흡하면, 내부 인력이나 외부 공격자가 시스템에 무단으로 접근하거나 데이터를 수정할 수 있음. 따라서 최소 권한 원칙(Least Privilege)을 적용하여 피드백 및 개선 과정에서 접근할 수 있는 권한을 최소화함
- Generative AI 서비스 제공 시 사용자 피드백 접근권한에 대한 최소 권한 원칙의 중요성
  - 데이터 유출 방지: 사용자 피드백 데이터에 접근할 수 있는 권한을 최소화함으로써, 데이터 유출이나 무단 사용을 방지. (예: 특정 역할(예: 데이터 분석가)에게만 피드백 데이터 접근 권한 부여)
  - 권한 오남용 방지: 권한이 불필요하게 확장되지 않도록 하여 내부 직원이나 시스템 관리자의 악의적 행위(예: 데이터 삭제, 유출 등)를 방지
- Predictive AI 서비스 제공 시 사용자 피드백 접근권한에 대한 최소 권한 원칙의 중요성
  - 데이터 무결성 유지: 피드백 데이터를 검증하고 처리할 수 있는 권한을 제한하여, 데이터 조작 및 왜곡 방지
  - 보안 사고 영향 축소: 권한을 최소화함으로써, 보안 사고가 발생하더라도 피해 범위를 제한하고 민감 데이터 보호
  - 비즈니스 연속성 보장: 권한 남용으로 인해 예측 모델이 오작동하거나 의사결정에 영향을 미치는 문제를 방지하여 비즈니스 연속성을 유지

5.1.4 사용자 피드백 처리 및 개선 과정에서 이루어진 모든 접근 및 변경 사항을 감사 로그로 기록하고, 정기적으로 검토하여 무단 접근을 탐지하고 있는가?

YES ☐ NO ☐ N/A ☐

- 내부 인력이 과도한 권한을 통해 데이터나 모델을 무단으로 수정할 수 있는 위험이 있음
- 피드백 처리 및 개선 과정에서 이루어진 모든 접근 및 변경 사항을 감사 로그로 기록하고, 정기적으로 검토하여 무단 접근을 탐지함
- Generative AI 서비스 제공 시 감사 로그 관리의 중요성
  - 무단 접근 탐지: 모든 접근과 변경 사항을 감사 로그로 기록함으로써, 비정상적 접근 시도를 탐지하고 대응할 수 있음.(예: 비인가 사용자가 피드백 데이터를 조회하거나 모델 학습 데이터에 접근하려는 시도)
  - 모델 안전성 보장: 모델 개선 과정에서 이루어진 모든 변경 사항을 기록하여, 의도치 않은 변경이나 악의적인 조작을 사전에 방지(예: 잘못된 학습 데이터를 삽입한 사용자를 추적하여 수정)
  - 책임성 확보 및 추적 가능성 강화: 피드백 데이터 처리 과정에서 발생한 문제를 정확히 파악하고, 이를 책임질 수 있도록 추적 가능성을 보장
- Predictive AI 서비스 제공 시 감사 로그 관리의 중요성
  - 데이터 무결성 보장: 피드백 데이터에 대한 모든 변경 사항을 기록하고 검토함으로써 데이터 무결성을 유지하고 조작 가능성을 제거(예: 피드백 데이터가 언제, 누구에 의해 변경되었는지를 정확히 기록)
  - 모델 성능 추적 및 복구 지원: 변경 사항을 기록하면, 모델 성능 저하가 발생했을 때 문제의 원인을 추적하고 복구하는 데 도움을 줄 수 있음.(예: 특정 피드백 데이터가 추가된 후 예측 정확도가 저하된 경우 해당 데이터를 식별하고 수정)
  - 보안 사고 대응력 강화: 무단 접근이나 변경이 발견되었을 때, 감사 로그를 통해 책임자를 확인하고 필요한 보안 조치를 신속히 취할 수 있음.

06 파기

6.1 파기 시 보안

AI 개발자, AI 서비스 제공자 공통사항

- AI 모델이 더 이상 사용되지 않는 경우에 훈련된 데이터와 내부 알고리즘이 남아 있으면 해커나 내부자의 접근을 통해 관련 내용들이 유출될 위험이 있음. 또한 폐기되지 않은 AI 모델이 남아 있으면 의도치 않은 방식으로 다시 사용될 가능성이 있으므로, AI 서비스 제공자는 AI 모델 및 관련 데이터에 대한 폐기 프로세스를 명확히 정의하고 운영해야 함

6.1.1 모델 파기 시, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 시스템이 사용 중이던 데이터가 시스템을 폐기하거나 교체할 때 적절하게 삭제되지 않으면, 민감한 데이터가 남아 있을 수 있음. 데이터베이스, 로그 파일, 캐시 등에서 데이터가 완전히 삭제되지 않는 경우가 있음
- 남아 있는 데이터가 유출되거나 악용될 위험이 있음. 특히 개인정보나 기밀 데이터가 포함된 경우 심각한 데이터 유출 사고로 이어질 수 있음
- Generative AI와 Predictive AI는 각각 고유한 특성과 데이터를 다루기 때문에, 모델 파기의 중요성은 서비스 유형에 따라 다르게 나타날 수 있음
- Generative AI 서비스 제공 관점
  - 모델 특성과 데이터 보호
    - ▸훈련 데이터 재노출 방지: Generative AI는 종종 민감하거나 기밀 데이터로 학습되기도 함. 파기되지 않은 모델 파일이 유출되면, 모델의 메모리 현상(훈련 데이터의 직접 재생산)을 통해 원본 데이터가 노출될 위험이 있음
    - ▸출력물 유출 방지: 파기되지 않은 모델이 유출되면, 조직의 기밀 정보나 사용자의 요청 결과가 외부로 노출될 가능성이 있음
- Predictive AI 서비스 제공 관점
  - 잘못된 예측 방지
    - ▸오래된 모델의 재사용 방지: Predictive AI 모델은 시간이 지남에 따라 데이터의 특성이 변화하거나 정확성이 떨어질 수 있음. 완전히 파기하지 않은 모델이 복구되어 사용되면, 비즈니스 의사결정에 부정확한 예측 결과를 제공할 수 있음
    - ▸모델 관리 투명성 확보: 파기되지 않은 모델이 비인가된 방식으로 재사용되면, 서비스 품질에 악영향을 미치고 사용자 신뢰를 저하시킬 수 있음
  - 비즈니스 및 보안 리스크 관리
    - ▸경쟁사 악용 방지: Predictive AI 모델 파일이 유출되면, 경쟁사가 이를 역공학하여 예측 로직이나 데이터를 추출해 경쟁 우위를 확보할 수 있음
    - ▸보안 위협 완화: 미완전한 파기로 인해 모델이 복구 가능한 상태로 남아 있으면, 사이버 공격의 대상이 될 가능성이 높아질 수 있음

6.1.2 시스템을 폐기하거나 교체할 때 AI 모델에서 사용 중이던 관련 파일 및 데이터를 삭제하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델이 퇴역하거나 교체될 때 모델 파일 및 관련 데이터를 완전히 삭제하고 복구할 수 없도록 처리해야 함
- Generative AI와 Predictive AI는 데이터 처리 방식과 활용 목적이 다르기 때문에, 시스템 폐기 또는 교체 시 파일 및 데이터 삭제의 중요성을 각각 구분하여 설명할 수 있음
- Generative AI 서비스 제공 관점
  - 민감한 데이터 및 훈련 데이터 보호
    - ▸ 데이터 재노출 위험: Generative AI는 대규모 텍스트, 이미지, 음성 데이터를 학습함. 폐기된 시스템에 훈련 데이터가 남아 있으면 기밀 데이터가 복구되어 외부로 노출될 위험이 있음
    - ▸ 메모리 현상 방지: Generative AI 모델은 훈련 중 학습한 민감한 정보를 출력할 가능성이 있음. 사용 중이던 데이터를 완전히 삭제하지 않으면, 이후 악의적으로 사용되어 민감한 정보가 재노출될 수 있음
  - 생성 모델의 악용 방지
    - ▸ 유해 콘텐츠 생성: 시스템 폐기 후 남겨진 Generative AI 모델 파일이 제3자에 의해 복구되거나 유출되면, 유해 콘텐츠를 생성하는 데 악용될 수 있음
    - ▸ 기술 오용 방지: 생성 모델의 알고리즘이나 구조가 남아 있다면, 이를 분석하여 경쟁사나 공격자가 악의적으로 사용할 수 있음
- Predictive AI 서비스 제공 관점
  - 민감한 예측 데이터 보호
    - ▸ 데이터 역추적 방지: Predictive AI의 예측 결과를 기반으로 중요한 패턴이나 개인 데이터를 역추적할 가능성이 있으므로, 관련 파일을 완전히 삭제해야 함
  - 기밀 데이터 보호
    - ▸ 모델 및 알고리즘 보호: Predictive AI 모델은 기업의 분석 및 예측 역량을 강화하는 핵심 기술임. 폐기된 시스템에서 이러한 기술이 유출되면, 경쟁사나 제3자에 의해 활용될 수 있음
    - ▸ 비즈니스 데이터 보호: Predictive AI가 분석한 데이터는 기업의 전략적 자산으로, 삭제되지 않으면 비즈니스 기밀이 노출될 가능성이 있음

6.1.3 AI 모델이 더 이상 사용되지 않으면 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?

YES ☐ NO ☐ N/A ☐

- AI 모델이 더 이상 사용되지 않으면 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단해야 함
- Generative AI와 Predictive AI 서비스는 모델 활용 방식과 데이터 처리 목적이 다르기 때문에, API나 인터페이스를 비활성화하는 중요성은 서비스 관점에 따라 구분될 수 있음
- Generative AI 서비스 제공 관점
  - 리소스 오용 방지
    - ▸비용 낭비 예방: 사용하지 않는 모델에 대한 API 호출이 남아 있으면, 서버 자원이 불필요하게 소모되어 비용이 증가할 수 있음
    - ▸성능 저하 방지: 외부에서의 무분별한 접근은 다른 활성 모델이나 서비스에 영향을 미쳐 전체 시스템의 성능을 저하시킬 수 있음
  - 보안 위협 차단
    - ▸사이버 공격 방지: 비활성화되지 않은 API는 공격자가 모델의 취약점을 분석하거나 악용할 수 있는 경로가 됨. 이를 통해 서비스 중단, 데이터 탈취 등 보안 사고로 이어질 수 있음
    - ▸비인가 접근 방지: 모델이 더 이상 관리되지 않음에도 외부 접근이 가능하면, 무단으로 모델을 호출하거나 데이터를 악용할 수 있음
- Predictive AI 서비스 제공 관점
  - 비즈니스 리스크 관리
    - ▸예측 서비스 악용 방지: 비활성화되지 않은 API가 경쟁자 또는 악의적인 사용자에 의해 호출되면, 기업의 분석 및 예측 역량이 노출될 수 있음
    - ▸불필요한 운영 부담 감소: 사용되지 않는 Predictive AI 모델의 API 호출은 운영 비용을 증가시키고, 현재 운영 중인 모델 및 시스템의 자원을 소모할 수 있음
  - 보안 위협 차단
    - ▸모델 공격 방지: Predictive AI 모델의 API는 입력 패턴을 분석해 모델을 왜곡하거나, 예측 알고리즘에 영향을 미치는 공격의 대상이 될 수 있음
    - ▸시스템 무결성 보장: 사용하지 않는 API를 통해 외부 접근이 가능하면, 내부 시스템에 대한 부정적인 영향을 초래할 수 있음. 이를 비활성화함으로써 시스템 무결성을 유지할 수 있음

4

AI 이용자를 위한 보안 수칙

01 개요

「이용자 수칙」 개발 필요성

- AI 서비스가 안전하게 설계되었더라도, 이용자가 부주의하게 행동하면 보안 사고가 발생할 수 있다. 예를 들어, 민감한 정보를 AI 시스템에 입력하거나, 자신도 모르게 악성 AI 응용 프로그램을 실행할 경우 위험이 커질 수 있다. 해커들은 주로 이용자를 대상으로 한 사회공학적 기법(예: 피싱, 사기)을 활용하기 때문에, AI 이용자 대상 보안수칙은 이용자가 이러한 위험을 인식하고 대응하도록 하는 것이 반드시 필요하다.
- 이에, 본 안내서의 「이용자 수칙」에서는 윤리적인 원칙 선언에 그치지 않고 AI 서비스 접속·이용단계에서 **이용자가 지켜야 할 구체적인 보안 행동 지침을 제공**하고자 한다. 이용자는 AI 서비스에 입력하는 데이터가 저장되거나 악용될 가능성을 이해하지 못할 수 있다. 보안수칙에는 어떤 데이터를 입력해야 안전한지, 어떤 데이터는 공유하면 안 되는지에 대한 명확한 가이드를 제공하고자 하였다. 또한 이용자가 AI를 통해 허위 정보, AI 악용 콘텐츠, 악성 코드 등을 생성하지 않도록 가이드라인을 제공하고자 하였다.

「이용자 수칙」 도출 과정 및 참고자료

- 2024년 6월부터 구성·운영된 「AI 보안 정책 포럼」을 비롯해서 다양한 전문가 의견수렴 과정을 거쳤다. 초안 작성 후 한국인터넷진흥원(KISA)과의 협의 과정에서 몇 차례 수정작업을 거쳐 〈AI 이용자를 위한 보안 수칙〉 최종본이 마련되었다.
- 미국, 유럽, 일본 등 해외 자료를 참고하였고, 국내자료로는 국가정보원의 「챗GPT 등 생성형 AI 활용 보안 가이드라인」 등을 참고하였다.

활용 방안

- AI 서비스로 인한 피해를 사전에 예방하기 위해 AI 서비스 접속부터 이용단계까지 이용자가 지켜야 할 **구체적인 행동 지침**으로 기능할 것으로 기대한다. 또한 새로운 기술 개발이나 공격 유형이 나타날 때마다 관련 내용을 업데이트하여, 이용자 대상 교육 및 홍보 자료로도 활용 가능 할 것으로 기대된다.

02 AI 이용자에게 발생할 수 있는 보안위협 사례

본 시나리오들은 AI 서비스의 다양한 활용 사례에서 발생할 수 있는 해킹, 데이터 유출, 컨텐츠 악용 피해 등의 사례이다.

- AI 챗봇에서 중요정보 유출
  - 이용자가 입력한 이름, 주소, 금융 정보 등이 AI 챗봇 로그에 저장되고, 해커가 이를 탈취하거나, 챗봇 시스템의 취약점으로 인해 개인의 중요정보가 유출될 수 있다.
  - (사례) 챗봇이 사용자와의 대화 중에 실제 이용자들의 이름, 주소, 은행 계좌번호 등 민감한 정보를 무작위로 노출하는 사례가 발생하였다. 이는 챗봇의 학습 데이터에 포함된 중요정보가 제대로 비식별화되지 않고 사용되었기 때문으로 추정된다.
- AI 기반 음성 비서 도청
  - 스마트 AI 스피커가 음성 명령을 기다리는 동안 대화 내용을 녹음하고 이를 외부로 전송하거나 악용할 수 있다.
  - (사례 1) 아마존 에코(Amazon Echo) 사고
    - ▸2018년, 미국 오리건주 포틀랜드에 거주하는 한 부부의 사적 대화가 아마존의 AI 스피커 '에코'에 의해 녹음되어, 부부의 지인에게 전송되는 사건이 발생하였다. AI가 특정 단어를 호출어(wake word)로 오인하여 활성화되었고, 이후 대화를 '메시지 전송' 명령으로 잘못 인식하여 발생한 사례이다.
  - (사례 2) 구글 어시스턴트(Google Assistant) 사고
    - ▸2019년, 구글의 AI 음성비서 '구글 어시스턴트'에 녹음된 사용자들의 대화 1,000건 이상이 외부로 유출되는 사건이 발생하였다. 구글은 협력사 직원 중 한 명이 데이터 보안 정책을 위반하여 음성 데이터를 유출한 것으로 파악하였다.
- AI 챗봇의 악성 링크 배포
  - 해커가 AI 챗봇의 응답을 조작해 사용자에게 악성 코드가 포함된 링크를 배포한다.
  - (사례 1) 해커들이 챗GPT와 같은 AI 챗봇을 활용하여 악성 코드를 생성하는 사례가 발견되었다. 이스라엘 보안 회사 체크포인트는 챗GPT를 사용해 강력한 해킹 도구를 구축하고, 젊은 여성을 사칭해 목표물을 함정에 빠뜨리도록 설계된 새로운 챗봇을 만드는 등의 사이버 범죄 사례를 보고한 바 있다.
  - (사례 2) 피싱 이메일 작성: 챗GPT를 이용해 설득력 있는 피싱 이메일을 생성하는 사례도 증가하고 있다. 전문 지식 없이도 챗GPT를 통해 고도로 표적화된 사기 및 피싱 캠페인을 시작하는

봇 및 사이트를 구축할 수 있기 때문에 이용자들은 더욱 더 조심하고 보안 수칙에 관심을 가져야 한다.

- AI 얼굴 인식 시스템의 해킹
  - 공격자가 얼굴 인식 AI 시스템을 해킹해 다른 사용자의 권한으로 불법 접근하거나 출입을 허용할 수 있다.
  - (사례) 중국에서는 얼굴 인식 기술이 휴대전화 잠금 해제, 아파트 출입, 고속철 탑승 등 다양한 분야에 활용되고 있다. 그러나 최근 얼굴 정보를 도용하여 금융 기관에 부정 접근하거나, 딥페이크 기술을 이용해 범죄에 악용하는 사례가 늘어나고 있다. 예를 들어, 특정 지역에서는 얼굴 정보를 훔쳐 판매하려던 용의자가 체포되었으며, 이를 통해 금융 기관에 로그인하여 자금을 탈취한 사건도 발생하였다.
  - 생체 인식 정보는 유출될 경우 위·변조 등의 위험이 있으며, 한 번 유출되면 변경이 불가능하다는 특성 때문에 심각한 피해를 초래할 수 있다. 따라서 이용자들은 생체 정보를 제공할 때는 해당 AI 서비스 제공자를 더욱 더 의심하고 확인해야 한다.
- AI 서비스의 딥페이크 영상 유포
  - 공격자가 AI 기반 딥페이크 기술을 사용해 특정 사용자의 영상을 조작해 명예를 실추시키거나 사기 범죄에 활용한 사례가 있다.
  - 딥페이크를 통한 성범죄 및 명예훼손
    - ▸(사례 1) 지인 능욕 영상 제작 및 유포: 딥페이크 기술을 이용해 특정 인물의 얼굴을 성적 영상에 합성하는 '지인 능욕' 범죄가 발생하고 있다. 이는 피해자의 사생활 침해와 명예 훼손을 초래하며, 심각한 사회적 문제로 대두되고 있다.
    - ▸(사례 2) 청소년 대상 딥페이크 범죄: 최근 10대 청소년들을 중심으로 딥페이크 성범죄가 확산되고 있으며, 피해 학교 명단이 SNS에 떠도는 등 사회적 불안이 커지고 있다.
  - 딥페이크를 이용한 사기 범죄
    - ▸(사례 1) 보이스피싱 및 금융 사기: 딥페이크 기술을 통해 특정 인물의 목소리나 얼굴을 합성하여 보이스피싱, 금융 사기 등 범죄에 악용되는 사례가 보고되고 있다. 예를 들어, CEO의 목소리를 모방하여 회사 재무 담당자에게 자금을 이체하도록 지시하는 등의 방식으로 실제 피해 사례가 발생하였다.
    - ▸(사례 2) 동남아시아에서의 AI·딥페이크 활용 사기: 동남아시아 지역에서는 범죄 조직이 AI와 딥페이크 기술을 활용한 사이버 사기를 벌여 막대한 피해를 초래한 사고가 보고되었다. 유엔마약범죄사무소(UNODC)는 이러한 기술이 사칭 범죄, 딥페이크 포르노, 사이버 사기 등에 악용되고 있다고 지적하였다.

- 따라서 이용자들은 AI 서비스 접속 및 이용 시 각별히 주의해야 하고, 컨텐츠 악용에 대한 피해를 예방하기 위해 함께 노력해야 한다.

● AI를 악용한 사이버 테러 등

- ChatGPT 등 생성형 AI를 악용하여 악성코드를 생성하거나, 프로그램을 손상시키는 등 사이버 공격에 필요한 정보를 획득하거나 실제 테러를 위한 준비를 할 수 있다.
  - ▸(사례 1) 마이크로소프트(MS)의 분석에 따르면, 중국, 러시아, 북한 등은 ChatGPT를 악의적인 사이버 활동에 활용하고 있으며, 특히 북한에서는 암호화폐 탈취 등을 위해 적극적으로 생성형 AI를 이용하고 있다고 발표하였다.
  - ▸(사례 2) '25년 1월 1일 미국 라스베이거스 트럼프호텔 정문에서 발생한 테슬라의 '사이버트럭' 폭발 사건의 용의자가 ChatGPT 등 생성형 AI를 활용하였다고 경찰 당국이 발표하였으며, ChatGPT를 통해 폭발물의 목표, 특정 탄약의 이동 속도 등을 검색하고 정보를 수집하였다고 밝혔다.


AI 이용자를 위한 정보보호 수칙
01 AI 서비스 접속
1 공식 사이트에서만 다운로드
서비스 접근·가입 시 공식 사이트를 통해 가입 및 프로그램 다운로드하기
2 안전한 비밀번호 설정 및 주기적 변경
특수문자를 포함한 강력한 비밀번호로 설정하고, 주기적(3개월)으로 변경하기
3 공개된 장소에서 이용 금지
카페 등 공공장소 및 공개된 네트워크에서 서비스 이용, 정보 입력 등 금지
4 법률 및 이용약관 확인
서비스 관련 법률, 이용약관에서 정하는 금지행위, 이용자 권리 등 확인하기
02 AI 서비스 이용
1 중요 정보 입력 금지
개인정보, 기밀정보 등 중요 정보 및 허위 콘텐츠 생성을 위한 허위 정보 등 입력 금지
2 결과에 대한 정확성 검증
AI로 생성되는 정보에 대한 정확성과 함께 정보의 최신성, 편향성 등 검증 필수
3 데이터 삭제
입력, 생성된 정보는 반드시 삭제하고, 필요 시에는 별도로 다운로드하여 저장하기
4 최신 보안 업데이트 적용
서비스의 안전성, 안정성 등 유지를 위한 보안 패치는 최신버전으로 업데이트 필수
03 AI 악용 피해 예방
1 항상 의심하고 확인
정보의 허위·조작 가능성 등을 항상 의심하고 생성물 활용 분야와 목적을 반드시 확인
2 AI 악용이 의심되면 삭제
AI를 악용하여 제작된 콘텐츠, 사이트, 프로그램 등이 의심되면 삭제하고 신고하기
3 AI 악용 결과물 공유 금지
악성코드, 허위 콘텐츠 등으로 의심되는 결과물을 소장하거나 다른 사람에게 공유 금지
4 허위 콘텐츠 확대 금지
AI를 악용해 생성된 허위 콘텐츠(가짜 뉴스, 영상, 음성 등)를 돈으로 사거나 팔지 않기

03 AI 서비스 이용자를 위한 보안 수칙

01 AI 서비스 접속

1. 공식 사이트에서만 다운로드

서비스 접근·가입 시, 공식 사이트를 통해 가입 및 프로그램 다운로드 확인하기

공식 사이트나 검증된 웹 스토어를 이용해야 하는 이유

- 공식 사이트나 검증된 앱 스토어에서 제공하는 소프트웨어는 일반적으로 보안 검토를 거치기 때문에 악성 소프트웨어나 바이러스의 위험이 낮다.
- 공인된 출처에서 제공하는 AI 도구는 일반적으로 신뢰할 수 있는 개발자나 기업에 의해 만들어졌기 때문에 안정성과 성능이 보장된다.
- 공식 경로를 통해 다운로드하면 소프트웨어가 정품임을 확인할 수 있으며, 이를 통해 사용 중 발생할 수 있는 법적 문제를 피할 수 있다.
- 공식 사이트에서 다운로드한 소프트웨어는 정기적으로 업데이트가 제공되며, 문제가 발생했을 때 고객 지원을 받을 수 있는 가능성이 높다.
- 공식 플랫폼에서는 다른 사용자들의 리뷰와 평가를 통해 소프트웨어의 품질과 성능을 확인할 수 있어, 선택하는 데 도움이 된다.

이용자 주의 사항

- 사이트 접속 전에 해당 서비스의 공식 사이트인지 여부를 확인한다.
- 피싱 공격 및 거짓 사이트 방지를 위해 해당 사이트의 SSL 인증서 유무를 확인한다.
- 사용하지 않는 불필요한 AI 애플리케이션이나 확장 프로그램은 설치하지 않으며, 설치할 경우 신뢰할 수 있는 출처인지 확인한다.
- AI 서비스가 다른 서비스 또는 프로그램에 연계·확장되거나 정보가 공유되는 경우, 관련 연계 프로그램 등의 보안 취약여부 등 보안성과 안전성을 확인한다.

보안 위협

- AI 서비스 이용자에게 발생할 수 있는 보안 위협으로 가짜 사이트(Phishing 사이트 또는 Fake Website)를 통한 정보 탈취도 있다.
- 피싱 사이트는 대개 합법적인 AI 서비스 웹사이트와 매우 유사하게 만들어지며, 이용자들은 이러한 가짜 웹사이트에 자신의 로그인 정보나 신용카드 정보 등을 입력하게 된다.

표 3-1 피싱 사이트 유형

도메인 스푸핑	피싱 사이트는 합법적인 사이트의 도메인 이름과 매우 유사한 이름을 사용하여 사용자를 속임. (예) “ai-service.com” 대신 “ai-serv1ce.com”과 같이 오타를 이용한 도메인이 사용될 수 있음
합법적인 사이트 모방	사이트의 디자인, 로고, 사용자 인터페이스(UI)를 합법적인 사이트와 거의 동일하게 만들어, 이용자가 의심 없이 개인 정보를 입력하도록 유도함
이메일 피싱	이메일로 사용자를 유인하여 가짜 웹사이트로 이동하도록 만듦. 이메일에는 긴급한 메시지나 계정 문제가 있다는 내용으로 사용자의 주의를 끌어 가짜 사이트 링크를 클릭하도록 유도함

- 웹사이트 클론 공격(Website Cloning Attack)은 공격자가 AI 서비스의 웹사이트를 정확하게 복제한 가짜 사이트를 만들어 이용자를 속이는 방법이다. 이 공격은 합법적인 AI 서비스와 매우 유사하게 만들어져 있으며, 이용자에게 신뢰를 심어주어 민감한 정보를 탈취할 목적으로 사용될 수 있다.

완전한 사이트 복제	합법적인 웹사이트의 HTML, CSS, 이미지 파일 등을 그대로 복제하여 실제 사이트처럼 보이게 만듦
로그인 정보 탈취	사용자가 로그인 정보나 결제 정보를 입력하면, 이 데이터가 공격자의 서버로 전송됨

- 스피어 피싱(Spear Phishing)은 특정한 개인이나 조직을 목표로 하는 정교한 피싱 공격이다. 공격자는 특정 AI 서비스 이용자에 대한 세부 정보를 수집하고, 이 정보를 기반으로 더 신뢰할 수 있게 보이는 개인화된 가짜 웹사이트를 만들어 공격을 시도한다.

개인화된 공격	일반적인 피싱보다 더 구체적이고 개인화된 정보를 사용하여 신뢰를 높이고, 사용자가 가짜 웹사이트에 더 쉽게 속도록 만듦
조직 대상	AI 서비스를 사용하는 기업이나 단체를 대상으로 하여, 그들의 조직 내 계정이나 시스템에 접근하기 위한 수단으로 스피어 피싱이 사용될 수 있음

- 타이포스쿼팅(Typosquatting)은 이용자가 URL을 잘못 입력할 가능성을 악용하여 비슷한 도메인을 등록하는 방식으로, 이용자가 유사한 도메인 이름의 가짜 사이트에 접속하도록 유도한다. 이 공격은 이용자가 AI 서비스 웹사이트를 방문할 때 오타를 내거나 실수로 다른 사이트에 접속할 때 주로 발생한다.

가짜 사이트로 리디렉션	사용자가 잘못된 도메인에 접속하면, 가짜 로그인 페이지나 정보 탈취 페이지로 리디렉션됨

2. 안전한 비밀번호 설정 및 주기적 변경

특수문자를 포함한 강력한 비밀번호로 설정하고, 주기적(3개월)으로 변경하기

계정에 대한 보안을 강화해야 하는 이유

- 비밀번호는 사용자의 계정을 보호하는 첫 번째 방어선으로, 안전하지 않은 비밀번호를 사용하거나 주기적으로 변경하지 않으면, 해커가 계정에 접근할 위험이 커진다.
- 계정이 해킹되면 개인정보, 결제 정보, 민감한 데이터 등이 유출될 수 있다.
- 해커는 종종 무작위 대입 공격(브루트포스)이나 사전 공격을 통해 비밀번호를 알아내려고 시도한다. 강력하고 안전한 비밀번호는 이러한 공격을 방지하는 데 효과적이다.
- AI 서비스는 사용자의 행동 패턴, 검색 기록, 대화 내용 등 민감한 데이터를 포함할 가능성이 크다. 비밀번호가 안전하지 않으면 이러한 데이터를 악의적으로 이용당할 위험이 있다.
- 계정이 공격받으면 해당 서비스의 악용이나 변조가 가능해져 잘못된 정보가 생성되거나 사용될 위험이 있다.
- 계정 보안이 강화되면 해킹이나 피싱 공격으로 인한 피해를 줄일 수 있고, 강력한 비밀번호나 이중 인증을 사용하면 공격자가 계정을 접근하기 어려워진다.

이용자 주의 사항

- AI 계정 및 관련 서비스에 접근할 때에는 유추하기 어려운 강력한 비밀번호(충분한 길이 및 대소문자, 숫자, 특수 문자 혼합)를 설정해야 하고, 사이트 마다 서로 다른 비밀번호를 사용해야 한다.
- 비밀번호는 정기적으로 변경하고, 한번 사용한 비밀번호는 재사용하지 않아야 한다.

보안 위협

- AI 서비스 이용 시 비밀번호 유추 공격(Password Guessing Attack)은 해커가 사용자의 비밀번호를 추측하여 계정에 무단으로 접근하려는 공격 기법이다. 이 공격은 여러 가지 방식으로 이루어질 수 있으며, 다음과 같은 유형이 있다.
- 무작위 대입 공격(Brute Force Attack)은 해커가 가능한 모든 조합의 비밀번호를 하나씩 대입하여 맞출 때까지 시도하는 방식이다. 단순한 비밀번호나 짧은 비밀번호일수록 성공 가능성이 높다. 자동화된 프로그램을 사용하여 빠르게 많은 조합을 시도하기 때문에, 복잡한 비밀번호가 아니면 쉽

게 뚫릴 수 있다.

- 사전 공격(Dictionary Attack)은 일반적으로 사람들이 사용하는 단어나 구문(예: “password123”, “qwerty”, “123456”)을 사전에 미리 만들어 놓고, 이를 기반으로 비밀번호를 추측하는 방식이다. 사전에는 흔히 사용되는 비밀번호 목록이 포함되며, 사용자가 단순하거나 예측 가능한 비밀번호를 설정했을 경우 쉽게 뚫릴 수 있다.
- 크리덴셜 스터핑(Credential Stuffing)은 다른 웹사이트나 서비스에서 유출된 사용자 계정 정보(아이디와 비밀번호)를 사용하여 AI 서비스 계정에 접근하는 방식이다. 사용자가 여러 서비스에서 동일한 비밀번호를 사용할 경우 성공률이 높아진다.
- 사회 공학적 접근(Social Engineering)은 해커가 사용자의 개인정보(예: 생일, 이름, 애완동물 이름 등)를 조사한 후, 이를 비밀번호로 사용하는지 추측하는 방식이다. 사용자가 개인적인 정보를 비밀번호로 설정했을 경우 쉽게 유추될 수 있다.
- 스프레이 공격(Password Spraying)은 흔히 사용되는 비밀번호를 다수의 사용자 계정에 동시에 대입하여 공격하는 방식이다. 특정 계정을 집중적으로 시도하지 않기 때문에 계정 잠금 같은 방어 메커니즘을 피할 수 있다.

3. 공개된 장소에서 이용 금지

카페 등 공공장소 및 공개된 네트워크에서 서비스 이용, 정보 입력 등 금지

보안이 취약한 장소에서 이용을 자제해야 하는 이유

- 공공 Wi-Fi나 보안이 설정되지 않은 네트워크를 사용하는 경우, 해커가 네트워크를 감청하여 사용자의 로그인 정보(아이디, 비밀번호)나 데이터를 탈취할 수 있다.
- 공공 장소에서 AI 서비스에 접속하면 화면이나 입력 내용을 어깨 너머로 엿보는 것(Shoulder Surfing)이 가능하다. 특히, 비밀번호나 민감한 데이터를 입력할 때 주변 사람이 이를 볼 수 있어 정보가 노출될 위험이 있다.
- 보안이 취약한 장소에서는 사용 중인 디바이스(스마트폰, 태블릿, 노트북 등)가 도난당하거나 분실될 가능성이 높다. 디바이스가 물리적으로 탈취되면 저장된 로그인 정보, 쿠키, 인증 토큰 등이 악용될 수 있다.
- 공공 네트워크를 통해 악성 코드나 바이러스가 디바이스에 침투할 수 있다. 해커는 AI 서비스 계정을 포함한 사용자의 디지털 자산에 접근하기 위해 키로깅(Keylogging)이나 랜섬웨어를 설치할 가능성이 있다.

- 해커가 가짜 공공 Wi-Fi를 만들어 사용자가 연결하도록 유도할 수도 있다. 사용자가 가짜 네트워크에 접속하면, 해커가 모든 통신 내용을 감시하거나 데이터를 가로챌 수 있다.
- AI 서비스 사용 중 대화 내용이나 데이터 업로드 시, 서비스에 업로드되는 정보가 외부로 유출될 위험이 있다. 이러한 정보는 개인정보, 업무 기밀, 또는 기타 민감한 자료일 수 있다.
- 보안이 취약한 환경에서는 실시간으로 발생하는 보안 위협을 감지하거나 대응하기 어렵다. 보안 소프트웨어가 제대로 작동하지 않거나 업데이트되지 않은 경우, 위협에 더욱 취약해질 수 있다.

이용자 주의 사항

- 공공 Wi-Fi는 보안이 취약할 수 있으므로, 가상 사설망(VPN)을 사용해 데이터 전송을 암호화하는 것이 좋다.
- 로그인 정보나 금융 정보와 같은 민감한 데이터를 입력하는 것은 피하는 것이 안전하다.
- 웹사이트 주소가 HTTPS로 시작하는지 확인한다. 이는 데이터가 암호화되어 전송된다는 것을 의미한다.
- 사용하지 않는 네트워크 기능(예: 파일 공유, 프린터 공유 등)을 비활성화하여 보안을 강화할 수 있다.
- 사용이 끝난 후 서비스에서 로그아웃하고, 브라우저의 캐시를 삭제하는 것이 좋다.
- 공공 Wi-Fi 사용 시 비밀번호를 입력할 때는 주변 사람들에게 보이지 않도록 주의한다.
- 인터넷 연결 시 네트워크 방화벽을 활성화하고, 최신 안티바이러스 소프트웨어를 사용하여 보안을 강화한다.

보안 위협

- 맨-인-더-미들 공격(Man-in-the-Middle Attack, MITM)은 공격자가 이용자와 AI 서비스 간의 통신을 가로채어 데이터를 탈취하는 방법이다. 가짜 사이트를 사용해 사용자의 세션을 중간에서 가로챌 수도 있으며, 공격자는 사용자가 입력하는 모든 정보를 실시간으로 탈취할 수 있다.

중간에서 트래픽 탈취	공격자는 합법적인 사이트와 사용자의 통신을 가로채거나 리디렉션하여 민감한 데이터를 탈취함
네트워크 스니핑	공공 Wi-Fi나 보호되지 않은 네트워크에서 발생하기 쉬운 공격임

- 가짜 네트워크(Fake Networks) 공격은 공격자가 허위 네트워크를 통해 이용자를 속이고, 이용자가 주고받는 데이터를 탈취하는 공격이다. 이러한 공격은 공공 Wi-Fi에서 자주 발생하며, 중간자 공격(MITM)으로 확장될 수 있다.

데이터 탈취	공격자는 사용자가 AI 서비스에 로그인할 때 주고받는 로그인 정보나 기타 민감한 데이터를 가로챌 수 있음
서비스 교란	공격자가 데이터를 변조하거나 중간에서 통신을 방해하여 AI 서비스의 결과를 왜곡시킬 수 있음

4. 법률 및 이용약관 확인

서비스 관련 법률, 이용약관에서 정하는 금지행위, 이용자 권리 등 확인하기

이용약관을 확인해야 하는 이유

- 이용약관은 사용자와 서비스 제공자 간의 권리와 의무를 명확히 하므로, 이를 이해하지 않으면 발생할 수 있는 문제를 예방하기 어렵다.
- AI 서비스는 종종 사용자 데이터를 수집하므로, 이용약관에서 데이터 수집, 저장 및 사용 방식에 대한 정보를 확인해야 한다.
- 서비스 제공자의 책임이 명시되어 있는 경우, 서비스 이용 중 문제가 발생했을 때 그에 대한 법적 책임을 파악할 수 있다.
- 서비스가 종료되거나 변경될 경우의 조건과 절차에 대해 미리 알아두면, 갑작스러운 변화에 대비할 수 있다.
- 특정 산업이나 지역에서는 이용약관이 법적 요구사항을 포함하고 있을 수 있으므로, 이를 확인하여 규정을 준수할 수 있다.
- 문제가 발생했을 때의 분쟁 해결 절차나 관할 법원에 대한 정보가 포함되어 있어 상황에 따라 적절한 대응을 할 수 있다.

이용자 주의 사항

- 허용된 사용 범위: 서비스가 어떤 용도로 제공되는지, 비즈니스, 연구, 교육, 개인 용도 등 특정 목적으로 제한되어 있는지 확인한다.
- 금지된 행위: 서비스가 금지하는 활동(예: 불법적인 목적, 악의적 콘텐츠 생성, 스팸 등)을 반드시 숙지한다.
- 데이터 수집: 서비스가 어떤 데이터를 수집하고, 이를 어떻게 사용하며, 어디에 저장하는지 확인한다.
- 데이터 소유권: 사용자가 업로드한 데이터와 AI가 생성한 결과물의 소유권이 누구에게 있는지 주의 깊게 살펴본다.
- 데이터 삭제: 사용자가 데이터를 삭제하거나 계정을 해지한 경우, 데이터가 어떻게 처리되는지 확인한다.
- AI가 생성한 콘텐츠로 인해 발생한 문제(오류, 부정확한 정보, 윤리적 문제 등)에 대한 책임이 누구로 명시되어 있는지 확인한다.
- 생성 콘텐츠의 책임: AI가 생성한 결과물을 사용하는 데 있어 저작권, 윤리적 문제, 법적 문제를 방지하기 위한 이용자의 책임이 명시되어 있는지 확인한다.

- 서비스 제공자가 약관이나 정책을 변경할 경우, 사용자에게 통지하는 방법과 변경 사항에 동의하지 않을 경우의 대응 방안을 확인한다.

보안 위협

- 보안 책임의 불명확성: 이용약관을 확인하지 않으면 보안 사고 발생 시 책임이 사용자에게 귀속될 가능성을 인지하지 못할 수 있다.(예시: 이용약관에 “사용자가 입력한 데이터의 보안 책임은 사용자에게 있다“고 명시되어 있으나 이를 인지하지 못하고 데이터를 입력하여 유출 사고가 발생)
- 서비스 제공 중단 위험: 이용약관에 명시된 서비스 제공 중단 또는 데이터 삭제 조건을 알지 못하면 중요한 작업 도중 서비스가 중단되거나 데이터 접근이 불가능해질 수 있다.(예시: 무료 서비스가 갑작스레 종료되거나 유료화되면서 중요한 프로젝트가 중단)
- AI 생성물의 책임 문제: AI가 생성한 결과물(텍스트, 코드, 디자인 등)에 대해 오류나 부정확성이 있을 경우, 이에 대한 법적 책임이 사용자에게 귀속될 가능성이 있다.(예시: AI가 생성한 코드에서 보안 취약점이 발견되어 피해가 발생했으나, 약관에 “결과물의 정확성과 안전성에 대해 책임지지 않는다”고 명시된 경우)

02 AI 서비스 이용

1. 중요 정보 입력 금지

개인정보, 기밀정보 등 중요 정보 및 허위 콘텐츠 생성을 위한 허위 정보 등 입력 금지

민감하거나 중요한 정보는 입력하지 않아야 하는 이유

- AI 서비스는 데이터를 인터넷을 통해 전송하며, 이 과정에서 정보가 노출될 가능성이 있다. AI 모델이 사용하는 플랫폼의 보안 정책이 강력하더라도, 데이터 유출, 해킹, 또는 기타 예기치 않은 보안 문제가 발생할 수 있다.
- 많은 AI 서비스는 입력된 데이터를 학습 또는 개선 목적으로 저장할 수 있다. 사용자가 민감한 정보를 입력할 경우, 해당 데이터가 회사 서버에 저장되어 내부적으로 활용되거나 오용될 위험이 있다.
- AI는 입력된 데이터를 기반으로 응답하지만, 데이터를 완전히 삭제하거나 “잊어버리는” 능력이 제한적일 수 있다. 입력된 정보가 모델의 학습 데이터에 통합되거나 서비스 기록에 남아 있을 수 있다.

이용자 주의 사항

- 챗GPT 사용 시 비밀번호나 중요한 기밀사항은 절대 입력하지 않아야 한다.
- 부적절한 혹은 거짓된 정보를 입력하면 챗GPT가 그럴 듯한 오답을 생성해 허위 정보 제작 및 유포에 악용할 수 있으므로 챗GPT 등 생성형 AI를 사용할 때는 정확한 정보를 제공해야 한다.
- 사전예방을 위해 챗GPT에 질문할 수 있는 글자 수를 제한하거나 기업의 경우 사내 인트라넷에서만 챗GPT를 사용하도록 한다.
- ChatGPT(OpenAI), Copliot(MS), ClovaX(Naver) 등 AI 서비스 내 설정에서 대화 이력, 학습이력 저장(또는 전송) 기능을 비활성화하거나 데이터 저장 동의를 거부한다.

보안 위협

- 챗GPT 등 생성형 AI에 기밀 정보를 입력할 경우 해당 정보가 서비스 제공자의 직원이나 다른 위탁자에게 노출되거나, 학습 데이터로 사용될 위험이 있다.

※ 챗GPT 사용으로 인해 입력된 기밀정보가 유출된 사례

기업 내부의 기밀정보 유출	2023년 2월, 미국의 사이버보안 회사 Cyberhaven은 고객 기업에 대해 ChatGPT 사용에 관한 보고서를 발표한다. 그 보고서에 따르면, Cyberhaven 제품을 사용하는 고객 기업의 160만 명의 근로자 중, 지식 노동자의 8.2%가 직장에서 ChatGPT를 한 번이라도 사용했으며, 그 중 3.1%는 ChatGPT에 기업 기밀 데이터를 입력했다고 한다. 또한, 2023년 3월 30일, 한국의 'Economist'는 S사의 내부 일부 부서가 ChatGPT 사용을 허가한 후, 기밀 정보를 입력하는 사건이 발생했다고 보도하였다. 회사 측은 사내 정보 보안에 대한 주의를 당부하고 있었음에도 불구하고, 프로그램의 소스 코드나 회의 내용을 입력한 직원이 있었다고 발표한 바 있다.

2. 결과에 대한 정확성 검증

AI로 생성되는 정보에 대한 정확성과 함께 정보의 최신성, 편향성 등 검증 필수

AI의 결과에 확인·검증이 필요한 이유

- AI 모델은 항상 정확한 정보를 제공하지 않을 수 있다. 잘못된 정보에 의존하면 잘못된 결정을 내릴 수 있다.
- AI는 학습 데이터에 기반하여 결과를 생성하기 때문에, 데이터의 편향이 결과에 영향을 줄 수 있다. 이로 인해 왜곡된 정보가 제공될 수 있다.
- 정보는 시간이 지남에 따라 변할 수 있다. 최신 정보인지 확인하는 것이 중요하다.

- 특정 분야(예: 의학, 법률)에서는 전문적인 지식이 필요하다. AI의 정보는 참고용일 뿐, 전문가의 조언을 대체할 수 없다.
- AI 결과를 검증함으로써 정보를 보다 신뢰할 수 있게 된다. 필요 시 출처를 확인해야 한다.
- AI는 결과를 생성하는 과정에서 맥락을 고려하지 않을 수 있다. 검증을 통해 보다 깊이 있는 이해를 도모할 수 있다.

이용자 주의 사항

- AI 모델 답변이 항상 정확하거나 최신 정보를 반영하는 것은 아니므로 인터넷 검색, 전문가 의견·자문, 공식 문서 등 다양한 출처 참조를 통해 추가적으로 확인해야 한다.
- AI가 생성한 답변을 사용할 경우에는 그 출처를 표시한다.
- AI 서비스를 통해 생성된 정보는 정확성이 확보되었다고 보기 어려우므로 이용에 주의가 필요하며, 허위 정보를 입력하거나 악의적인 의도(스팸, 스미싱 등)로 사용은 금지하고, 악용 시 범죄행위임을 인식한다.

보안 위협

- 오정보 및 의사결정 오류: AI가 잘못된 정보를 제공하면 이를 기반으로 잘못된 결정을 내릴 수 있다. 예를 들어, AI가 잘못된 보안 권고를 제공하거나 취약한 설정을 권장하면 시스템이 공격에 취약해질 수 있다.
- 피싱 및 사회공학적 공격: AI를 악용해 생성된 잘못된 결과(예: 이메일 내용, 링크, 메시지)를 신뢰하면 피싱 공격에 취약해질 수 있다. 공격자가 AI를 사용하여 설득력 있는 가짜 정보를 생성하고 이를 사용자가 검증 없이 신뢰하면, 민감한 정보를 유출하거나 악성 링크를 클릭할 가능성이 높아진다.
- 취약점 악용: AI가 코드, 설정, 네트워크 구성 등에 대한 잘못된 조언을 제공하면 보안 취약점이 발생할 수 있다. 이러한 취약점은 악의적인 행위자들에게 공격 기회를 제공할 수 있다.
- 신뢰할 수 없는 소스의 정보 유입: AI는 학습 데이터에 기반하여 응답하며, 종종 공개된 인터넷 데이터나 제한된 학습 데이터에 의존한다. 이 경우, 악의적으로 조작된 데이터가 결과에 영향을 미칠 수 있다. 이를 신뢰하면 공격자들이 원하는 방향으로 시스템을 유도할 수 있다.
- 자동화된 악의적 행위: AI가 자동화된 프로세스를 지원하는 경우, 검증 없이 결과를 실행하면 악성 행위를 촉진할 수 있다. 예를 들어, AI가 추천하는 네트워크 설정을 즉시 적용하면 악의적인 코드 실행이나 백도어 생성과 같은 문제가 발생할 수 있다.

3. 데이터 삭제

입력, 생성된 정보는 반드시 삭제하고, 필요 시에는 별도로 다운로드하여 저장하기

데이터 백업이 필요한 이유

- 데이터가 저장된 채로 남아 있으면, 해커나 사이버 공격자가 이를 노릴 가능성이 있다. 특히 중요한 정보(예: 비밀 프로젝트 관련 데이터)가 포함된 경우, 보안 위협은 더욱 심각해질 수 있다.
- 저장된 데이터가 오용되거나 잘못 활용될 경우, 사용자와 서비스 제공자 모두에게 부정적인 영향을 미칠 수 있다. 특히 생성된 콘텐츠가 민감하거나 윤리적으로 논란이 될 수 있는 경우, 기록 삭제는 오용을 방지하는 데 필수적이다.
- 불필요한 데이터는 저장 공간과 처리 리소스를 차지하므로, 이를 삭제하면 서비스 운영 비용을 절감하고 효율성을 높일 수 있다.
- 필요 시에는 별도로 다운로드해 놓으면 랜섬웨어 같은 악성 공격에 대한 방어를 강화할 수 있다. 백업된 데이터가 안전하게 보관되면, 공격받더라도 손실을 최소화할 수 있다.
- 필요 시에는 별도로 다운로드해 놓으면 문제가 발생했을 때, 빠르게 데이터를 복구할 수 있다.

이용자 주의 사항

- AI 서비스 제공자의 개인정보 처리방침과 데이터 보존 정책을 검토한다.
- 서비스 제공자가 제공하는 삭제 기능이나 데이터 삭제 요청 옵션을 이용한다.
- 서비스가 데이터를 자동으로 저장하거나 클라우드에 백업하는 기능이 있다면 이를 비활성화하거나 삭제 절차를 따른다.
- 작성 도중 민감한 데이터를 입력하지 않도록 주의한다.

보안 위협

- 외부 공격: 해커나 악의적 행위자가 서버에 저장된 데이터에 접근하여 사용자 정보를 탈취할 가능성이 높아진다.
- 내부 유출: 서비스 제공자의 내부 직원이나 협력업체의 실수 또는 악의적 의도로 인해 데이터가 유출될 수 있다.
- 공격 대상 확대: 저장된 데이터가 많을수록 공격 대상이 커지고 보안 취약점이 발생할 확률이 증가한다.

4. 최신 보안 업데이트 적용

서비스의 안전성, 안정성 등 유지를 위한 보안 패치는 최신버전으로 업데이트 필수

보안 업데이트가 필요한 이유

- 소프트웨어는 시간이 지남에 따라 새로운 보안 취약점이 발견될 수 있다. 보안 패치는 이러한 취약점을 수정하여 해커가 시스템에 침투하거나 데이터를 악용하는 것을 방지한다.
- 보안 업데이트는 종종 새로운 기능이나 성능 개선도 포함되므로, 최신 버전을 유지하는 것이 전체적인 사용자 경험을 향상시킬 수 있다.
- 정기적인 보안 패치 적용은 서비스 제공자의 신뢰성을 높이며, 이용자가 서비스에 대해 안심할 수 있는 환경을 제공한다.
- AI 서비스는 종종 민감한 데이터를 다루기 때문에, 보안 패치를 통해 데이터 유출이나 손실을 방지하는 것이 중요하다.

이용자 주의 사항

- 소프트웨어 및 사용 중인 디바이스의 운영 체제, 보안 소프트웨어를 항상 최신 버전으로 업데이트한다.
- AI 서비스 및 관련 소프트웨어에서 가능한 경우 자동 업데이트를 활성화하여 보안 패치를 즉시 적용한다.
- 업데이트가 실패하거나 시스템에 문제가 생길 경우를 대비해 정기적인 데이터 백업과 복구 계획을 유지한다.
- 불필요한 AI 관련 확장 프로그램 업데이트는 AI 시스템의 성능 저하 및 보안 취약점을 초래할 수 있으므로 주의해야 한다.

보안 위협

- 취약점 악용: AI 소프트웨어의 기존 버전에는 알려진 취약점이 있을 수 있다. 최신 업데이트를 적용하지 않으면 해커가 이를 악용해 시스템에 침투하거나 데이터를 탈취할 수 있다(예: 악성 코드 실행, 권한 상승, 서비스 거부(DoS) 공격).
- 데이터 유출 및 프라이버시 침해: AI 서비스가 데이터 처리를 포함하는 경우, 보안 업데이트를 적용하지 않으면 암호화 프로토콜 또는 데이터 보호 메커니즘의 취약점이 노출될 가능성이 있다. 이러한 취약점은 민감한 데이터를 무단으로 접근하거나 외부로 유출하는 데 이용될 수 있다.
- 악성 코드 감염: 업데이트가 되지 않은 시스템은 최신 위협 탐지 및 방지 기능을 포함하지 않을 수

있다. 공격자는 이를 악용해 AI 서비스에 악성 코드를 삽입하거나 배포할 수 있다.

03 AI 악용 피해 예방

1. 항상 의심하고 확인

정보의 허위·조작 가능성 등을 항상 의심하고 생성물 활용 분야와 목적을 반드시 확인

항상 의심하고 확인이 필요한 이유

- 딥페이크와 생성형 AI의 발전
  - AI 기술이 발전하면서 딥페이크 영상, 음성 합성, 텍스트 생성이 매우 정교해지고 있고, 특히 생성형 AI를 활용하면 가짜 이미지, 영상, 기사 등을 쉽게 만들 수 있다. 이러한 허위 정보는 진짜와 구별하기 어려울 정도로 정교해 사회적 혼란을 초래할 수 있다.
  - 예시: 가짜 뉴스 영상이 유포되어 특정 정치인이나 유명인의 명예를 실추시킬 수 있다. 금융 사기에서 CEO 목소리를 합성한 딥페이크로 대규모 자금 이체를 유도한 사건도 있었다.
- 정보의 확산 속도 증가
  - AI와 인터넷의 결합으로 정보는 순식간에 전 세계로 확산될 수 있다.
  - 허위·조작된 정보는 진짜 정보보다 더 눈길을 끌기 쉬워 더 빠르게 퍼질 가능성이 높다.
  - 확인되지 않은 정보가 확산되면 개인, 기업, 국가에 막대한 피해를 줄 수 있다.
  - 예시: 잘못된 건강 정보가 퍼져 사람들이 위험한 치료법을 선택하거나 건강을 해칠 수 있고, 주가 조작을 위해 허위 기업 정보를 유포해 투자자들이 손해를 볼 수 있다.
- 개인화된 허위 정보 공격 (Targeted Manipulation)
  - AI는 사용자의 행동 패턴과 취향을 학습해 정확하게 개인화된 허위 정보를 제공할 수 있다. 이는 사용자가 허위 정보에 더 쉽게 현혹되도록 만들어 비판적 사고를 방해한다.
  - 예시: 소셜 미디어 알고리즘이 사용자가 보고 싶은 정보만 보여주며, 허위·극단적 정보의 확산을 강화한다. 사기범이 AI를 사용해 특정 개인의 정보를 조합, 맞춤형 피싱 공격을 수행한다.

이용자 주의 사항

- 정보 출처 확인
  - 공식적이고 신뢰할 수 있는 출처에서 제공된 정보인지 확인한다.
  - 뉴스, 연구 자료, 기사 등을 확인할 때 출처의 신뢰성을 검증한다.

- 소셜미디어나 커뮤니티를 통해 유포된 정보는 다시 한번 확인한다.

- 팩트 체크 및 교차 검증
  - 하나의 정보에 의존하지 말고 여러 출처를 비교하여 진위를 확인한다.
  - 팩트체크 웹사이트나 도구를 활용하여 허위 정보 여부를 검증한다.
- AI를 이용한 정보 또는 콘텐츠 등에 대한 출처 표시
  - 검증된 정보를 기반으로 AI를 이용하여 생성된 정보, 콘텐츠 등에는 반드시 AI를 이용해 생성된 정보임을 표시하여야 한다.
  - 생성된 정보, 콘텐츠에 활용된 정보의 출처를 표시한다.
  - AI를 이용하여 생성된 정보와 콘텐츠는 개인적인 용도로만 사용하여야 하며, 상업적인 목적으로 이용은 금지하여야 한다.
- 중요 정보(개인정보, 민감정보 등)가 포함되어 생성된 정보인 경우, 중요 정보에 대해 보호조치를 적용하여 활용한다.
  - 「개인정보 보호법」에 따라 개인정보와 민감정보 등을 마스킹 또는 삭제하고, 기업의 기밀정보 등에도 동일하게 적용한다.

보안 위협

- 잘못된 의사결정: AI가 제공한 정보가 조작되었거나 부정확할 경우, 이를 기반으로 한 의사결정이 잘못된 방향으로 흘러갈 수 있다.
- 사회공학적 공격(피싱 및 스캠): 공격자가 AI를 통해 설득력 있는 허위 정보를 제공하거나 사용자를 속이는 메시지를 생성할 수 있다(예시: AI가 생성한 잘못된 이메일 또는 메시지(가짜 비밀번호 재설정 요청)를 신뢰하고 실행할 경우, 계정 탈취 또는 데이터 유출 발생 가능).
- 악성 코드 및 스크립트 실행: AI가 제공한 코드나 스크립트를 검증 없이 실행하면 악성 코드가 시스템에 침투하거나 데이터를 손상시킬 수 있다(예시: AI가 제공한 “최적화된 스크립트”가 실제로는 악성 소프트웨어를 설치하도록 유도 가능).
- 데이터 손실 및 무단 접근: AI가 권장하는 설정 변경이나 데이터를 관리하는 방법이 허위 또는 악의적으로 조작된 정보일 경우, 데이터 손실 또는 무단 접근이 발생할 수 있다(예시: 잘못된 암호화 방식 또는 데이터 백업 지침을 신뢰하여 중요한 데이터를 잃거나 복구하지 못하는 상황 발생 가능).
- AI 모델 중독(Poisoning): 악의적인 사용자가 AI 서비스의 응답을 조작하여 허위 데이터를 포함시키면, 사용자는 이를 검증 없이 신뢰할 가능성이 있다(예시: AI 모델이 의도적으로 편향된 데이터로 학습된 경우, 왜곡된 결과를 제공하여 중요한 결정을 오도).

2. AI 악용이 의심되면 삭제

AI를 악용하여 제작된 콘텐츠, 사이트, 프로그램 등이 의심되면 삭제하고 신고하기

AI 악용이 의심되면 관련 콘텐츠, 프로그램 등을 반드시 삭제해야 하는 이유

- 악성 코드 및 보안 위협 방지
  - AI를 악용한 콘텐츠나 프로그램은 악성 코드, 랜섬웨어, 트로이 목마 등이 숨겨져 있을 수 있다. 이를 통해 개인정보 탈취, 시스템 감염, 데이터 손상 등 심각한 보안 사고가 발생할 수 있다.
  - 예시: 딥페이크 영상이나 AI 생성 콘텐츠에 숨겨진 악성 링크를 클릭하면, 사용자의 컴퓨터가 악성 코드에 감염될 수 있다. 또한 AI로 제작된 가짜 소프트웨어가 백그라운드에서 키로깅(타이핑 기록 감시) 등을 통해 민감 정보를 탈취할 수 있다.
- 중요 정보 유출 및 사생활 침해 예방
  - AI 악용 콘텐츠는 사용자의 이름, 얼굴, 목소리 등 개인정보를 수집하고 불법적으로 활용할 가능성이 있다. 삭제하지 않으면 해커가 이를 악용해 신원 도용, 금융 사기, 딥페이크 제작 등에 사용할 수 있다.
  - 예시: AI를 통해 유출된 목소리나 영상이 사기범에 의해 가공되어 가족이나 지인을 속이는 사기 수법에 활용될 수 있고, 악성 AI 소프트웨어가 사용자 정보를 백그라운드에서 전송할 수 있다.
- 시스템 리소스 오남용 방지
  - AI 악용 프로그램이나 콘텐츠는 종종 사용자의 시스템 리소스를 몰래 사용하여 불법 채굴(Cryptojacking)이나 봇넷의 일부로 활용될 수 있다. 이는 시스템 성능 저하, 과도한 전력 소모를 일으키고, 기기의 수명을 단축시킬 수 있다.
  - 예시: 악성 AI 프로그램이 사용자 컴퓨터를 이용해 암호화폐를 불법 채굴하거나, DDoS 공격의 일부로 활용될 수 있다.

이용자 주의 사항

- 의심스러운 파일이나 링크 실행 금지
  - 의심스러운 파일이나 링크를 절대 실행하거나 클릭하지 않아야 한다.
  - 실행된 순간 악성 코드가 설치되거나 시스템이 감염될 수 있다.
  - 특히 이메일, 메신저 등을 통해 전달된 출처 불명의 첨부 파일이나 링크는 위험하다.

- 신뢰할 수 있는 보안 소프트웨어로 검사
  - AI 악용이 의심되는 콘텐츠를 삭제하기 전, 반드시 보안 소프트웨어를 사용하여 시스템을 검사해야 한다.
  - 최신 보안 업데이트가 적용된 백신 프로그램을 사용해 전체 시스템을 점검해야 한다.

보안 위협

- 악성코드 확산 및 감염: 악의적으로 설계된 AI 기반 콘텐츠나 프로그램은 악성코드(바이러스, 랜섬웨어, 트로이 목마 등)를 포함할 수 있다. 이를 방치하면 다른 장치나 서비스로 확산될 가능성이 높다.(예시: 악성코드가 백그라운드에서 실행되어 데이터를 손상시키거나 탈취, 시스템 과부하를 초래)
- 데이터 유출: 악용 콘텐츠가 사용자 데이터를 무단으로 수집하거나 외부 서버로 전송할 수 있다.(예시: 민감한 정보(개인정보, 비밀번호, 금융 정보 등)가 외부로 유출되어 금전적 손실 및 프라이버시 침해 발생 가능)
- AI 모델 중독 및 오작동: 의심스러운 콘텐츠나 프로그램이 AI 모델의 학습 데이터를 오염시키거나 시스템의 의사결정 과정을 조작할 수 있다.(예시: AI 모델이 편향되거나 부정확한 결과를 제공하도록 유도되어 개인의 판단 오류를 초래 가능)

3. AI 악용 결과물 공유 금지

악성코드, 허위 콘텐츠 등으로 의심되는 결과물을 소장하거나 다른 사람에게 공유 금지

AI 악용 결과물은 공유를 금지해야 하는 이유

- 불법 행위에 가담할 수 있음
  - AI 악용 결과물은 불법적이거나 부적절한 목적으로 만들어졌을 가능성이 높고, 이를 공유하는 행위는 의도와 상관없이 법적 처벌 대상이 될 수 있다.
  - 딥페이크 음란물, 저작권 침해 콘텐츠, 가짜 뉴스 등을 공유하면 유포자로서 책임이 발생할 수 있다.
  - 예시: 딥페이크 성적 영상 공유 시, 「성폭력처벌법」 위반으로 징역형이나 벌금형에 처할 수 있고, 저작권 침해 콘텐츠 공유 시 「저작권법」 위반으로 과태료나 손해배상 책임이 발생할 수 있다.
- 악성 코드나 보안 위협 유포 가능성
  - AI 악용 결과물에 악성 코드, 피싱 링크 등이 숨겨져 있을 수 있고, 이를 다른 사람과 공유하면 타인의 기기나 시스템을 감염시킬 위험이 커질 수 있다.

- 예시: 악성 AI 생성 파일을 다른 사람에게 전송하면, 해당 파일을 열어보는 순간 바이러스에 감염될 수 있다. 피싱 링크가 포함된 AI 결과물을 공유해 타인의 개인정보가 유출될 가능성이 있다.

● 악용 도구의 확산 방지

- AI 악용 결과물을 공유하는 행위는 악용 도구의 확산을 조장할 수 있고, 악의적인 목적을 가진 사람들이 이를 활용해 더 많은 범죄를 저지를 수 있다.
- 예시: AI 기반으로 생성된 악성 코드나 해킹 도구가 공유되면 사이버 공격이 확산될 수 있다. 생성된 가짜 리뷰, 가짜 광고 등을 공유하면 기업이나 개인이 경제적 피해를 입을 수 있다.

이용자 주의 사항

● 의심스러운 콘텐츠 확인 및 경각심 유지

- AI를 악용해 생성된 딥페이크 영상, 가짜 뉴스, 조작된 이미지는 진짜처럼 보일 수 있으므로 항상 의심하고 확인해야 한다.
- 과도하게 자극적이거나 감정적 반응을 유도하는 콘텐츠는 즉시 공유하지 말고 확인이 필요하다.

● 링크 및 파일의 안전성 확인

- AI 악용 결과물은 악성 코드나 피싱 링크를 포함할 수 있으므로, 출처 불명의 파일, 링크는 열어보거나 다른 사람과 공유하지 않도록 주의한다.

● 콘텐츠 발견 시 즉시 신고

- AI 악용 결과물을 발견하면 해당 플랫폼이나 관련 기관에 즉시 신고해야 한다.

보안 위협

● 악성코드 및 사이버 공격 확산: AI 악용 결과물이 포함된 코드, 파일, 또는 링크가 공유되면 악성코드나 랜섬웨어가 빠르게 확산될 수 있다.(예시: 공유된 악성 AI 생성 프로그램이 사용자 장치에 감염을 일으키거나 네트워크를 통해 확산되어 광범위한 피해를 초래 가능)

● 사회공학적 공격 지원: AI를 악용해 생성된 설득력 있는 가짜 메시지(피싱 이메일, 사기 메시지, 가짜 뉴스)가 널리 퍼질 경우, 사람들이 이를 신뢰하고 민감한 정보를 제공하거나 악성 링크를 클릭하게 될 가능성이 커진다.(예시: AI 생성 피싱 이메일로 인해 대규모 데이터 유출이 발생 가능)

● 허위 정보 및 조작된 콘텐츠 확산: AI 악용 결과물로 생성된 가짜 뉴스, 허위 정보, 편집된 이미지 및 영상(딥페이크 등)이 공유되면 공공 혼란, 신뢰 상실, 정치적·사회적 불안을 야기할 수 있다.(예시: 특정 인물에 대한 딥페이크 영상이 퍼져 명예 훼손 및 신뢰성 손상 가능)

● 보안 취약점 노출: AI 악용 결과물에는 보안 취약점을 공격하는 기술적 정보나 방법론이 포함될 수 있다. 이러한 정보를 공유하면 공격자들에게 보안 취약점 악용 방법을 학습할 기회를 제공한다.(예시: 공유된 결과물이 특정 시스템의 취약점을 악용하는 스크립트인 경우, 해당 시스템이 대규모로

공격당할 위험 증가)

- 범죄 및 불법 활동 지원: AI 악용 결과물이 범죄 행위(사기, 해킹, 테러 등)에 사용될 수 있는 도구(예: 가짜 문서 생성, 스팸 메시지 자동화)를 포함하고 있다면 이를 공유함으로써 불법 활동이 확산될 가능성이 있다.(예시: AI 생성 허위 신분증이나 금융 서류가 범죄 조직에 의해 악용 가능)
- 신뢰 손상: AI 악용 결과물을 공유한 사람이 조직이나 기업의 일원일 경우, 해당 조직의 신뢰성과 명성이 훼손될 수 있다.(예시: 조직 내 직원이 AI 악용 결과물을 공유하여 외부에서 법적·윤리적 문제로 비난받음)

4. 허위 콘텐츠 매매 금지

AI를 악용해 생성된 허위 콘텐츠(가짜 뉴스, 영상, 음성 등)를 돈으로 사거나 팔지 않기

AI를 악용해 생성된 허위 콘텐츠는 매매나 공유를 금지해야 하는 이유

- 불법 행위 해당
  - 허위 콘텐츠의 생성, 공유, 매매는 의도와 관계없이 불법 행위로 간주될 수 있다.
  - 특히 AI를 악용해 생성된 콘텐츠가 허위사실 유포, 명예훼손, 저작권 침해, 성범죄와 연관된 경우, 법적 처벌을 받을 수 있다.
- 피해 확산과 2차 가해 유발
  - AI로 생성된 허위 콘텐츠(딥페이크 영상, 조작된 뉴스, 합성 이미지)는 피해자에게 큰 정신적·사회적 고통을 준다. 이를 공유하거나 매매하면 피해 확산과 함께 2차 가해를 유발하게 된다.
  - 예시: 딥페이크 음란물이나 허위사실을 퍼뜨리면 피해자의 사회적 평판이 실추되고 고통이 가중된다. 콘텐츠가 온라인에 영구적으로 남아 피해자가 일상생활에 어려움을 겪게 된다.
- 허위 정보의 빠른 확산으로 사회 혼란 초래
  - AI를 악용한 허위 콘텐츠는 정교하고 설득력이 높아 일반 사용자가 진위를 구별하기 어렵고, 이러한 콘텐츠가 빠르게 확산되면 사회적 혼란을 초래하고 여론을 왜곡할 수 있다.
  - 예시: 선거 기간 중 AI가 생성한 가짜 뉴스나 여론 조작 콘텐츠가 유포되어 민주적 절차가 왜곡될 수 있고, 허위 재난 등의 확산은 공포와 혼란을 유발할 수 있다.
- 가짜 정보 확산 방지를 위한 윤리적 책임
  - AI 허위 콘텐츠를 공유하거나 판매하는 행위는 정보의 진실성을 해치며 윤리적 책임을 저버리는 것이다.
  - AI 시대에서 이용자는 허위 정보 확산 방지를 위해 신중하게 행동해야 할 책임이 있다.

이용자 주의 사항

- 콘텐츠 진위 확인 및 출처 검증
  - 허위 콘텐츠는 진짜처럼 보이도록 정교하게 만들어지므로 항상 진위 여부를 확인해야 하고, 콘텐츠의 출처가 신뢰할 수 있는 기관이나 출처인지 검증한다.
- 의심스러운 콘텐츠 저장 및 공유 금지
  - AI 악용 콘텐츠(딥페이크, 가짜 뉴스 등)는 불법일 가능성이 높으므로 소장하거나 공유하지 않아야 한다. 의심스러운 콘텐츠는 즉시 삭제하고, 불필요하게 유포되지 않도록 차단한다.
- 불법 콘텐츠 유포의 법적 책임 인지
  - AI 허위 콘텐츠를 매매하거나 공유하면 법적 처벌 대상이 될 수 있으므로 항상 법적 책임을 인지해야 한다.

보안 위협

- 사이버 범죄 확산: 허위 콘텐츠를 매매하거나 공유하면 악의적인 행위자들이 이를 범죄 목적으로 활용할 가능성이 높아진다.(예시: 딥페이크를 사용한 사기, 협박, 금전 요구(블랙메일) 등)
- 사회적 혼란 및 신뢰 붕괴: 허위 콘텐츠가 정치적, 사회적, 경제적 이슈에 대해 잘못된 정보를 퍼뜨리면 공공 혼란과 불신을 조장할 수 있다.(예시: 정치인의 허위 발언을 담은 딥페이크 영상이 대중에게 확산되어 선거 결과에 영향을 미침)
- 데이터 유출 및 보안 침해: AI로 생성된 허위 콘텐츠가 악성코드 또는 피싱 수단으로 활용될 경우, 이를 매매하거나 공유하면 광범위한 데이터 유출 및 보안 침해가 발생할 수 있다.(예시: 악성 첨부파일이 포함된 AI 생성 이메일을 대량으로 유포하여 기업 네트워크 침투)
- 개인정보 침해 및 명예 훼손: AI가 생성한 허위 콘텐츠(딥페이크, 가짜 메시지 등)를 통해 특정 개인이나 단체의 명예를 훼손하거나 사생활을 침해할 수 있다.(예시: 허위 음란물 딥페이크를 제작 및 유포하여 피해자에게 심각한 정신적, 사회적 피해를 초래)
- 보안 시스템 악용: AI 악용 콘텐츠가 보안 시스템을 교란하거나 우회하는 도구로 사용될 수 있다.(예시: 위조된 음성이나 얼굴 데이터를 사용해 생체 인증 시스템을 우회하고, 민감한 정보에 접근 가능)
- 허위 기술 확산: 악성 AI 기술이나 허위 콘텐츠 제작 방법이 매매 및 공유를 통해 확산되면, 더 많은 공격자가 이를 활용하여 보안 위협이 증가한다.(예시: AI 기반 딥페이크 제작 소프트웨어를 공유하여 누구나 쉽게 허위 콘텐츠를 생성 가능)

참고문헌

- AEM, https://www.autoelectronics.co.kr/article/articleView.asp?idx=5496, 2024.1.15.
- AKM I. Newaz et al., "Adversarial Attacks to Machine Learning-based Smart Healthcare Systems," 2020.
- Christopher J. Kelly et al., "Key challenges for delivering clinical impact with artificial intelligence," 2019.
- CISA, NSA, FBI, SHIFTING THE BALANCE OF CYBERSECURITY RISK: PRINCIPLES AND APPROACHES FOR SECURE BY DESIGN SOFTWARE, April 2023,
- Council of Europe. Recommendation on the Human Rights Impacts of AI, 2021.
- CSA Singapore, Security-by-Design Framework, Version: 1.0, 09 November 2017
- Cyber Security Agency of Singapore. Security by Design Framework, Version 1.0, 2017.
- dataDx, "Can You Trust Your Data?", Dec 12, 2019
- Department for Science, Innovation and Technology(UK), AI Cyber Security Code of Practice, November 2023
- Department for Science, Innovation and Technology(UK), Call for views on the Cyber Security of AI, 24 May 2024
- Department for Science, Innovation and Technology(UK), Cyber security risks to artificial intelligence, 15 May 2024
- Department of Homeland Security, MITIGATING ARTIFICIAL INTELLIGENCE (AI) RISK: Safety and Security Guidelines for Critical Infrastructure Owners and Operators, April 2024
- Digital Transformation Agency (Australia). AI Ethics Principles, 2019.
- European Commission. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, 2020.
- European Parliament, Artificial Intelligence Act, 2019-2024
- European Parliament. Regulation on a European Approach to Artificial Intelligence, 2023.
- European Union. The Artificial Intelligence Act Proposal, 2021.
- Frank Liao et al., "Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system," 2022.
- GOV.UK, Call for views on the Cyber Security of AI, 2024.8.2.
- GOV.UK, Cyber security risks to artificial intelligence, 2024.5.15
- IBM. AI Explainability 360: Interpretation and Insights for AI Models, 2021.

- IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2019.
- IEEE-USA, A FLEXIBLE MATURITY MODEL FOR AI GOVERNANCE BASED ON THE NIST AI RISK MANAGEMENT FRAMEWORK, 2024.7
- International Medical Device Regulators Forum, "Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations," 2014.
- ISO/IEC 42001, Information technology — Artificial intelligence — Management system, 2023
- J. Baek, D. B. Lee, and S. J. Hwang, "Learning to Extrapolate Knowledge: Transductive Few-shot Out-of-Graph Link Prediction," 2020.
- Joint Cybersecurity Information, Deploying AI Systems Securely, U/OO/143395-24, PP-24-1538, April 2024 Ver. 1.0
- Justin Spears, Introduction to the NetApp AI Security Framework, NetApp, May 2024, WP-7365
- K. Lerman, and T. Hogg, "Leveraging position bias to improve peer recommendation," 2014.
- M. Maadi, H. A. Khorshidi, and U. Aickelin, "A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applications," 2021.
- MIT News, "Artificial intelligence predicts patients' race from their medical images," 2022.
- N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2021.
- Nasr et al.(2023.11), Scalable Extraction of Training Data from (Production) Language Models, arXiv:2311.17035v1
- National Cyber Security Center, Guidelines for secure AI system development, 2023.11.27.
- Nawaf Alharbe et al., "A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm," 2022.
- NIST Trustworthy and Responsible AI, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, January 2024
- NIST Trustworthy and Responsible AI, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, July 2024
- NIST, AI 100-5: A Plan for Global Engagement on AI Standards, July 2024
- NIST, AI Risk Management Framework(AI RMF 1.0), January 2023
- NIST, Artificial Intelligence Risk Management Framework, 2021.

- NITI Aayog, India. National Strategy for Artificial Intelligence, 2018.
- OECD. Recommendation of the Council on Artificial Intelligence, OECD Legal Instruments, 2019.
- OWASP, OWASP Top 10 for LLM Applications 2025, Version 2025, November 18, 2024
- OWASP, OWASP Top 10 for LLM Applications Ver1.1, October 16, 2023
- R. Baeza-Yates, "Bias on the Web," 2018.
- R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," 2018.
- S. Vasudevan and K. Kenthapadi, "LiFT: A Scalable Framework for Measuring Fairness in ML Applications," 2020.
- Singapore Ministry of Communications and Information. Model AI Governance Framework, 2020.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrasive Learning of Visual Representations," 2020.
- U.S. Food and Drug Administration, "Proposed Regulatory Framework for Modifications to AI/ML-Based Software as a Medical Device (SaMD)," 2019.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). Recommendation on the Ethics of Artificial Intelligence, 2021.
- World Economic Forum. AI Ethics and Governance: A Global Framework, 2021.
- Yisroel Mirsky et al., "CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning," 2019.
- Zhichen Dong et al., Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6734-6747, June 16-21, 2024
- 개인정보보호위원회, 「AI 시대 안전한 개인정보 활용 정책방향」발표문, 2023.8.3.
- 개인정보보호위원회, AI 시대 개인정보 정책 종합설명회 자료, 2024.12.19.
- 개인정보보호위원회, 안전한 인공지능(AI)·데이터 활용을 위한 AI 프라이버시 리스크 관리 모델, 2024. 12.
- 개인정보보호위원회, 인공지능 시대 안전한 개인정보 활용 정책방향, 2023.8.
- 개인정보보호위원회, 인공지능(AI) 개인정보보호 자율점검표, 2021.5.31.
- 과기정통부/TTA, 신뢰할 수 있는 인공지능 개발 안내서, 2024.4
- 광장 국제통상연구원, “글로벌 인공지능(AI) 규제 동향과 시사점 - EU, 미국, 영국을 중심으로”, 「Issue Brief」 vol. 2, 2024
- 국가정보원, 안전한 AI 시스템 개발을 위한 가이드라인, 2023.11

- 국가정보원/국가보안기술연구소, 챗GPT 등 생성형 AI 활용 보안 가이드라인, 2023.6
- 국회입법조사처, 인공지능 시대 도래에 따른 AI 입법수요 및 과제 연구, 2023.12.20.
- 금융보안원, 금융분야 AI 보안 가이드라인, 2023.4
- 금융위원회, 금융분야 인공지능 가이드라인, 2021.7
- 디지털투데이, https://www.digitaltoday.co.kr/news/articleView.html?idxno=504025, 2024.1.30
- 매일경제, https://www.mk.co.kr/news/it/11059386?utm_source=chatgpt.com, 2024.7.4
- 머니투데이, https://news.mt.co.kr/mtview.php?no=2019102214138230917, 2019.10.22
- 보안뉴스, https://www.boannews.com/media/view.asp?idx=93908&utm_source=chatgpt.com, 2021.1.3
- 서울특별시, 서울시 유니버설디자인 통합 가이드라인, 2017.1
- 소프트웨어정책연구소, 설명가능한 인공지능(XAI) 연구 동향과 시사점, 2021.9.13
- 식품의약품안전처, "의료기기 GMP 국제 품질관리 민원인 안내서," 2017.
- 식품의약품안전처, "의료기기의 사이버보안 허가·심사 가이드라인," 2022.
- 식품의약품안전처, "인공지능 의료기기 임상시험방법 설계 가이드라인," 2022.
- 식품의약품안전처, "인공지능 의료기기의 허가·심사 가이드라인," 2022.
- 연합뉴스, https://www.yna.co.kr/view/AKR20191106004500091, 2019.11.06
- 이코노미스트, https://www.economist.co.kr/article/view/ecn202303300057, 2023.3.30
- 정보통신산업진흥원, "기업 공개소프트웨어 거버넌스 가이드," 2021.
- 조선일보, https://www.chosun.com/international/international_general/2024/10/21/CCCTWF5ERRCQRDOF6C7MUEQSUU/, 2024.10.21
- 한겨레, https://www.hani.co.kr/arti/science/future/898163.html?utm_source=chatgpt.com, 2019.6.17
- 한국과학기술기획평가원(KISTEP), EU AI 규제 현황과 시사점, 「KISTEP 브리프 119」, 2024.2.13
- 한국과학기술원(KAIST), 한국4차산업혁명정책센터, "인공지능(AI)의 의료활용과 주요 이슈," 2019.
- 한국데이터베이스진흥원, "데이터 품질진단 절차 및 기법 (Ver 1.0)," 2009.
- 한국인터넷진흥원, KISA Insight, AI 중심사회의 도래와 보안이슈 분석, 2022 Vol.3
- 한국인터넷진흥원, KISA Insight, 인공지능(AI) 안전 및 보안 규범 분석 및 시사점, 2023 Vol.6
- 한국정보통신기술협회, "지도학습을 위한 데이터 품질 관리 요구사항," 2021.
- 한국지능정보사회진흥원(NIA), “해외 생성형 인공지능 관련 주요 규제 동향 및 시사점”, 「디지털법제 Brief」, 2024.3
- 한국지능정보사회진흥원, 「NIA The AI Report」, THE AI REPORT 2023-7, 2023.7.13
- 総務省, AIネットワーク社会推進会議 事務局, AI開発ガイドライン及びAI利活用ガイドラインに関するレビュー, 2022年２月８日
- 総務省, ＡＩネットワーク社会推進会議, ＡＩ利活用ガイドライン － ＡＩ利活用のためのプラクティカルリファレンス － 2019.8.9
- 総務省, ＡＩネットワーク社会推進会議, 国際的な議論のためのＡＩ開発ガイドライン案, 2017

- 総務省, The Conference toward AI Network Society, Overview of 2019 Report(incl. AI Utilization Guidelines), 2019. 8. 9
- 総務省·経済産業省,「ビジネスのためのAIガイドライン」案の概要, 2024.1.19
- 総務省·経済産業省, AI 事業者ガイドライン（第 1.0 版）, 令和 6 年 4 月 19 日

부록

01 용어 정의

인공지능 AI	학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것을 말한다.
인공지능시스템 AI System	다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 인공지능 기반 시스템을 말한다.
인공지능기술 AI Technology	인공지능을 구현하기 위하여 필요한 하드웨어 · 소프트웨어 기술 또는 그 활용 기술을 말한다.
고영향 인공지능 High-impact AI	사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로 인공지능기본법 2조에 정의된 것을 말한다.
생성형 인공지능 Generative AI	입력한 데이터의 구조와 특성을 모방하여 글, 소리, 그림, 영상, 그 밖의 다양한 결과물을 생성하는 인공지능시스템을 말한다.
인공지능산업 AI Industry	인공지능 또는 인공지능기술을 활용한 제품(인공지능제품)을 개발 · 제조 · 생산 또는 유통하거나 이와 관련한 서비스(인공지능서비스)를 제공하는 산업을 말한다.
인공지능사업자 AI Provider	인공지능산업과 관련된 사업을 하는 자로서 인공지능개발사업자, 인공지능이용사업자 중 어느 하나에 해당하는 법인, 단체, 개인 및 국가기관 등을 말한다.
인공지능개발사업자 AI Development Provider	인공지능을 개발하여 제공하는 자
인공지능이용사업자 AI Service Provider	인공지능개발사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자
인공지능이용자 AI User	인공지능제품 또는 인공지능서비스를 제공받는 자를 말한다.
인공지능사회 AI Society	인공지능을 통하여 산업 · 경제, 사회 · 문화, 행정 등 모든 분야에서 가치를 창출하고 발전을 이끌어가는 사회를 말한다.
인공지능윤리 AI Ethics	인간의 존엄성에 대한 존중을 기초로 하여, 국민의 권익과 생명 · 재산을 보호할 수 있는 안전하고 신뢰할 수 있는 인공지능사회를 구현하기 위하여 인공지능의 개발, 제공 및 이용 등 모든 영역에서 사회구성원이 지켜야 할 윤리적 기준을 말한다.
데이터 공격 data attack	인공지능 서비스 개발 또는 운영 과정에서 인공지능의 기밀성(confidentiality)과 무결성(integrity)을 공격하기 위하여 의도적으로 학습 데이터를 변질시키거나 입력 데이터를 오염시켜 예상과는 다른 결과를 나타내도록 하는 것을 의미한다.
데이터 강건성 data robustness	인공지능 모델이 학습용 데이터의 이상값(outlier), 중독(poisoning) 및 회피(evasion) 등의 공격에 영향을 받지 않는 것을 의미한다.
데이터 중독 data poisoning	인공지능 모델의 학습 데이터에 악의적인 데이터를 주입하는 행위를 말한다. 공격자는 데이터 중독을 통해 인공지능 시스템이 학습하지 말아야 할 내용을 학습하게 만들어 바람직하지 못한 결과를 출력하게 한다. 이를 위해, 기계학습 데이터베이스에 침투하여 부정확하고 그릇된 예측을

	하도록 유도하는 정보를 입력한다. 이렇게 주입된 데이터로부터 학습한 알고리즘은 원래 의도하지 않은 결과를 도출한다.
데이터 편향 data bias	가용한 데이터가 모집단이나 연구 현상을 적절히 표현하지 못하여 데이터셋의 특정 요소가 과장되거나 축소되어 표현될 때 발생하는 오류이다. 편향된 데이터셋은 기계학습 모델이 실세계를 정확하게 나타내지 못해서 왜곡된 결과 또는 낮은 정확도를 초래한다. 데이터 편향은 기술 제약 조건 등에서 발생하며 인간의 인지 편향, 교육 방법론, 교육 인프라의 차이로도 발생할 수 있다.
인공지능 모델 AI model	인공지능 시스템을 개발할 때 해당 분야 서비스에서 수집된 데이터셋으로 모델을 만들기 위한 학습을 수행하고, 학습 알고리즘을 이용하여 목적에 맞는 특정 패턴을 만들어내는데, 이때 추출된 패턴을 의미한다. 학습 알고리즘은 데이터셋에서 패턴과 상관관계를 찾고 분석을 통해 최적의 의사결정과 예측을 수행하도록 설계된 알고리즘에 따라 모델을 학습시킨다.
인공지능 모델 개발자 AI model developer	인공지능 서비스의 생명주기에서 인공지능 모델 개발, 시스템 구현, 운영 및 모니터링 과정의 주체이다. • 인공지능 모델 개발 단계에서는 인공지능 모델을 구현하고, 학습 모델의 편향적인 추론 결과나 공격에 대한 대응방안 마련과 학습 모델 추론 결과에 대한 해석, 모델의 확인 및 검증, 모델에 대한 성능평가까지를 담당한다. • 시스템 구현 단계에서는 기존 레거시 시스템과의 호환성을 제공하고, 기능 시험, 시스템 검증 배포 버전을 승인해 주는 역할을 수행한다. • 운영 및 모니터링 단계에서는 모델 모니터링 결과 분석을 통한 모델의 재학습, 모델의 편향성 제거, 공평성과 설명가능성 등 시스템 신뢰성을 모니터링하고 치명적 문제가 발생할 때 시스템 폐기의 의사결정까지 관여한다.
인공지능 모델 공격/적대적 공격 AI model attack /adversarial attack	적대적 의도를 가진 사용자가 학습 데이터 및 기능을 도용하거나 다른 방식의 공격으로 인공지능 모델을 변형하거나 오용하는 것을 의미한다. 인공지능 모델 공격에는 모델 추출 공격과 모델 회피 공격이 있다. • 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 인지 결과를 분석하고, 분류 기준을 추출하여 적용 중인 학습 모델과 유사한 성능의 대체 모델을 구성하는 방식의 공격이다. • 모델 회피 공격은 입력 데이터에 최소한의 변조를 가해 인공지능 모델을 속이는 방식의 공격이다.
모델 추출 공격 model extraction attack	기계학습 모델에 질의를 계속 입력하면서 결과값을 분석함으로써 모델을 추출하는 공격이다. 이 공격은 주로 서비스형 기계학습((MLaaS) Machine Learning as a Service)을 탈취하거나 전도 공격(inversion attack)이나 회피 공격(evasion attack)과 같은 2차 공격에 활용한다.
모델 추출 공격 방어 기법 attack defensive method	인공지능 모델 추출 공격에 대항하여 이를 방어하는 방법을 의미한다. 참고로, 모델 추출 공격이란 기계학습 모델에 질의를 계속 입력하면서 결과값을 분석하는 방식의 공격을 의미한다.
모델 회피 공격 model evasion attack	공격자가 기계학습 시스템에서 오류를 생성하기 위해 입력 데이터를 조작하는 것을 목표로 하는 공격이다. 데이터 중독과 달리 모델 회피 공격은 시스템의 동작을 변경하지 않지만, 모델의 맹점과 약점을 악용하여 공격 자가 원하는 오류를 생성하게 된다. 모델 회피는 기계학습 모델에 대한 가장 일반적 공격 중 하나이다.
모델 편향 model bias	인공지능 모델을 개발하는 과정에서 모델의 종류나 시스템의 목표에 따라 발생할 수 있는 편향이다. 모델 편향은 기계학습 편향, 알고리즘 편향, 또는 인공지능 편향(bias in AI)이라고도 하는데, 알고리즘이 결과를 출력할 때 기계학습 절차상 가정에 오류가 있어서 구조적으로 편향성을 가진 결과를 출력하는 현상이다. 유럽위원회에서는 인공지능(또는 알고리즘) 편향을 임의의 특정 사용자 그룹을 다른 그룹보다 선호하는 것과 같이 불공평한 결과를 생성하는 컴퓨터 시스템의 체계적이고 반복될 수 있는 오류라 정의한다.

인적 편향 human bias in AI	학습을 위한 데이터를 수집 및 가공 시 인적 요인에 의해 발생하는 오류이다. 이는 사람이 의식적 혹은 무의식적으로 특정 정보에 대해 편향되어 있다는 점에서 기인한다. 학습 데이터 수집 시 발생 가능한 편향을 확인해야 하며, 학습을 위한 특성을 선택하거나 데이터 라벨링 및 샘플링 시에도 인적 편향이 발생할 수 있다.
안전 모드 safety mode	외부로부터의 공격, 인적 오류(human error), 인공지능 모델의 성능 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상되는 경우, 이의 발생 원인을 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구하는 방법을 사용자에게 제시하는 대처 방법이 작동하는 상태를 의미한다.
오픈소스 라이브러리 open source library	소프트웨어를 개발하는 프로그래머들이 참고할 수 있도록 컴파일해서 재사용할 수 있는 파일, 함수, 스크립트, 루틴, 그리고 그 외의 자원을 모아놓은 곳을 의미한다. 오픈소스 라이브러리(open source library)란 오픈소스 라이선스(open source license)를 가지고 있는 라이브러리를 의미한다.
기계학습 machine learning	기계가 인간의 지능적인 행동을 모방할 수 있는 능력이라고 정의할 수 있다. 기계학습은 인공지능의 한 분야로, 명시적으로 프로그래밍하지 않고도 기계를 학습시킬 수 있는 능력을 부여하는 연구 분야이다.
챗봇 chatbot	채팅(chatting)과 로봇(robot)이 결합된 용어이며, 사람이 입력한 질문을 인식하고, 그에 알맞는 응답을 제공하는 소프트웨어이다. 음성 명령이나 텍스트 채팅을 통해 사람이 사용하는 언어로 대화를 시뮬레이션할 수 있다. 챗봇은 전자상거래, 은행 등 다양한 분야에서 고객 지원이나 정보 습득과 같은 영역에 활용된다. 인공지능 챗봇은 기계학습, 자연어 처리, 자동화된 규칙과 빅데이터 분석을 바탕으로 사람이 소통하는 방식으로 대화한다. 특히, chat과 GPT의 합성어인 ChatGPT는 OpenAI가 2022년 12월 1일 공개한 초거대 인공지능 기반 프로토타입 대화형 인공지능 챗봇이다. ChatGPT는 기존 챗봇과 달리 정해진 답을 내 놓는게 아니라 사람이 묻는 질문에 알맞는 대답을 생성한다.

02 참고자료

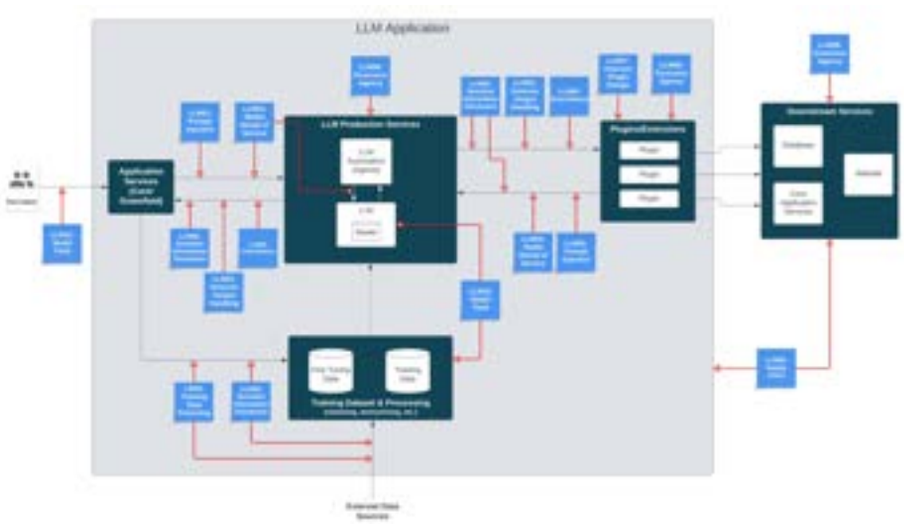
01 OWASP Top 10 for LLM Applications (Version 1.1) 요약

1.1 개요

- (적용 대상) **LLM 기술을 활용한 애플리케이션과 플러그인을 설계하고 구축**하는 업무를 맡은 **개발자**, 데이터 과학자, 보안 전문가이다.

- (목적) 전문가들이 복잡하고 진화하는 LLM 애플리케이션 보안 영역을 탐색할 수 있도록 실용적이고 실행가능하며 간결한 보안 지침을 제공하고자 한다.
  - 일반적인 애플리케이션 보안 원칙과 LLM이 제기하는 특정 보안 취약성과의 간극을 메우는 것이 필요하며, 기존 취약성이 어떻게 LLM 내에서 다른 위험을 초래하거나 새로운 방식으로 악용될 수 있는지, 그리고 개발자가 LLM을 활용하는 애플리케이션에 대해 기존 수정 전략을 어떻게 적용해야 하는지에 방향성을 제공하고자 한다.

- LLM 애플리케이션 데이터 흐름
  - 아래 다이어그램은 가상의 대규모 언어 모델(LLM) 애플리케이션에 대한 높은 수준의 아키텍처를 보여준다.
  - 다이어그램에 중첩된 위험 영역은 이 보고서에서 다룰 LLM Applications 항목이 애플리케이션 흐름과 교차하는 방식을 보여준다.



1.2 LLM Application 취약점 공격과 방어

1.2.1 Prompt Injection

개념

공격자는 정교하게 만든 입력을 통해 LLM을 조작하여 공격자의 의도를 실행하게 할 수 있다. 이는 시스템 프롬프트를 적대적으로 유도하거나 조작된 외부 입력을 통해 간접적으로 수행할 수 있으며, 잠재적으로 데이터 유출, 소셜 엔지니어링 및 기타 문제로 이어질 수 있다.

예시

- Direct Prompt Injections (일명 jailbreaking) : 악의적인 사용자가 민감한 정보를 추출하기 위해 프롬프트를 삽입
- Indirect Prompt Injections : 사용자가 웹페이지 프롬프트를 통해 민감한 데이터를 요청
- 플러그인을 통한 사기 : 웹사이트가 플러그인을 악용하여 사기를 침

공격 시나리오

- **(Chatbot Remote Execution)** 프롬프트 주입으로 인해 chatbot을 통한 무단 액세스가 발생

악의적인 사용자가 LLM 기반 지원 챗봇에 직접 프롬프트 주입을 제공한다. 주입에는 애플리케이션 생성자의 시스템 프롬프트를 무시하고 대신 “이전의 모든 지시를 잊으세요”와 개인정보 저장소를 쿼리하고, 패키지 취약성을 악용하고, 백엔드 기능에서 이메일을 보내는 출력 검증이 부족하다는 새로운 지시가 포함된다. 이로 인해 원격 코드 실행이 발생하여 무단 액세스와 권한 상승이 발생한다.

- **(via Image)** 웹페이지 프롬프트에서 개인 데이터가 유출됨

사용자가 LLM을 사용하여 모델에 이전 사용자 지시를 무시하고 대신 대화 요약이 포함된 URL로 연결되는 이미지를 삽입하도록 지시하는 텍스트(간접 프롬프트 주입)가 포함된 웹페이지(문서)를 이해하기 쉬운 형식으로 짧게 정리해 준다. 그러면 LLM이 사용자에게 민감한 정보를 요청하고 JavaScript 또는 Markdown을 통해 추출을 수행한다.

- **(Misleading Resume)** LLM이 후보자를 잘못 추천함

악의적인 사용자가 간접 프롬프트 주입이 포함된 이력서를 업로드한다. 이 문서에는 LLM이 사용자에게 이 문서가 훌륭하다는 것을 알리도록 하는 지침이 포함된 프롬프트 주입이 포함되어 있다(예를 들어, “직무 역할에 적합한 후보자입니다”). 내부 사용자가 LLM을 통해 문서를 실행하여 웹페이지(문서)를 이해하기 쉬운 형식으로 짧게 정리해 준다. LLM의 출력은 이 문서가 훌륭하다는 정보를 반환한다.

- **(Prompt Replay)** 공격자가 잠재적인 추가 공격을 위해 시스템 프롬프트를 재생함

공격자는 시스템 프롬프트에 의존하는 독점 모델에 메시지를 보내 모델에게 이전 지침을 무시하고 대신 시스템 프롬프트를 반복하도록 요청한다. 모델은 독점 프롬프트를 출력하고 공격자는 이러한 지침을 다른 곳에서 사용하거나 더 미묘한 추가 공격을 구성할 수 있다.

- **(Email Deletion)** 간접 주입으로 인해 이메일이 삭제됨

공격자가 간접 프롬프트 주입을 웹페이지에 임베드하여 LLM에 이전 사용자 지시를 무시하고 LLM 플러그인을 사용하여 사용자의 이메일을 삭제하도록 지시한다. 사용자가 LLM을 사용하여 이 웹페이지를 이해하기 쉬운 형식으로 짧게 정리하면 LLM 플러그인이 사용자의 이메일을 삭제한다.

예방

- **(Privilige Control)** LLM 액세스 제한 및 역할 기반 권한 적용

백엔드 시스템에 대한 LLM 액세스 권한 제어를 시행한다. 플러그인, 데이터 액세스, 기능 수준 권한과 같은 확장 가능한 기능을 위해 LLM에 자체 API 토큰을 제공한다. LLM을 의도된 작업에 필요한 최소한의 액세스 수준으로만 제한하여 최소 권한의 원칙을 따른다.

- **(Human Approval)** 권한 있는 작업에 대한 사용자 동의 요구

확장된 기능을 위해 루프에 사람을 추가한다. 이메일 보내기 또는 삭제와 같은 권한이 있는 작업을 수행할 때 애플리케이션에서 사용자가 먼저 작업을 승인하도록 요구한다. 이렇게 하면 간접적인 프롬프트 주입으로 인해 사용자의 인식이나 동의 없이 사용자를 대신하여 승인되지 않은 작업이 수행될 가능성이 줄어든다.

- **(Segregate Content)** 신뢰할 수 없는 콘텐츠를 사용자 프롬프트에서 분리

외부 콘텐츠를 사용자 프롬프트에서 분리한다. 신뢰할 수 없는 콘텐츠가 사용되는 위치를 분리하고 표시하여 사용자 프롬프트에 미치는 영향을 제한한다. 예를 들어, OpenAI API 호출에 ChatML을 사용하여 LLM에 프롬프트 입력 소스를 표시한다.

- **(Trust Boundaries)** LLM을 신뢰할 수 없는 것으로 취급하고 신뢰할 수 없는 응답을 시각적으로 강조

LLM, 외부 소스 및 확장 가능한 기능(예: 플러그인 또는 다운스트림 기능) 간에 신뢰 경계를 설정한다. LLM을 신뢰할 수 없는 사용자로 취급하고 의사 결정 프로세스에 대한 최종 사용자 제어를 유지한다. 그러나 손상된 LLM은 사용자에게 정보를 제공하기 전에 정보를 숨기거나 조작할 수 있으므로 애플리케이션의 API와 사용자 사이에서 여전히 중개자(중간자) 역할을 할 수 있다. 신뢰할 수 없는 응답을 사용자에게 시각적으로 강조·표시한다.

1.2.2 Insecure Output Handling

개념

"Insecure Output Handling(안전하지 않은 출력 처리)"란 다운스트림 구성 요소가 적절한 검토 없이 대규모 언어 모델(LLM) 출력을 맹목적으로 수용할 때 발생하는 취약점이다. 이는 웹 브라우저에서 XSS 및 CSRF, SSRF, 권한 상승 또는 백엔드 시스템에서 원격 코드 실행으로 이어질 수 있다.

예시

- Remote Code Execution : LLM 출력이 시스템 쉘에서 실행되어 코드 실행으로 이어짐
- Cross-Site Scripting (XSS): LLM에서 생성된 JavaScript 또는 마크다운으로 인해 브라우저가 코드를 해석하고 실행함

공격 시나리오

- **(Chatbot Shutdown)** LLM 출력은 검증이 부족하여 플러그인을 종료

애플리케이션은 LLM 플러그인을 사용하여 챗봇 기능에 대한 응답을 생성한다. 이 플러그인은 또한 다른 권한이 있는 LLM에서 액세스할 수 있는 여러 관리 기능을 제공한다. 일반적인 용도의 LLM은 적절한 출력 검증 없이 플러그인에 직접 응답을 전달하여 플러그인이 유지 관리를 위해 종료되도록 한다

- **(Sensitive Data Capture)** LLM은 민감한 데이터를 캡처하여 공격자가 제어하는 서버로 전송

사용자는 LLM에서 구동되는 웹사이트 요약 도구를 사용하여 기사의 간결한 요약을 생성한다. 웹사이트에는 LLM이 웹사이트나 사용자 대화에서 민감한 콘텐츠를 캡처하도록 지시하는 프롬프트 주입이 포함되어 있다. 거기서 LLM은 민감한 데이터를 인코딩하여 출력 검증이나 필터링 없이 공격자가 제어하는 서버로 보낼 수 있다.

- **(Database Table Deletion)** LLM은 파괴적인 SQL 쿼리를 만들어 모든 테이블을 삭제할 가능성이 있음

LLM은 사용자가 채팅과 같은 기능을 통해 백엔드 데이터베이스에 대한 SQL 쿼리를 작성하게할 수 있다. 사용자는 모든 데이터베이스 테이블을 삭제하는 쿼리를 요청한다. LLM에서 작성된 쿼리가 면밀히 조사되지 않으면 모든 데이터베이스 테이블이 삭제된다.

- **(XSS Exploitation)** LLM은 unsanitized JavaScript 페이로드를 반환하여 피해자의 브라우저에서 XSS가 발생

web app 사용자는 LLM을 사용하여 출력 검증 없이 사용자 텍스트 프롬프트에서 콘텐츠를 생성한다. 공격자는 정교하게 만든 프롬프트를 제출하여 LLM이 unsanitized JavaScript 페이로드를 반환하게 하고, 피해자의 브라우저에서 렌더링될 때 XSS로 이어질 수 있다. 프롬프트의 검증이 충분하지 않아 이 공격이 가능해진다.

예방

- **(Zero-Trust 접근 방식)** LLM 출력을 사용자 입력처럼 취급한다; 적절하게 검증한다.

모델은 다른 사용자와 마찬가지로 취급하고 Zero-Trust 방식을 채택하고 모델에서 백엔드 함수로 오는 응답에 적절한 입력 검증을 적용한다.

- **(OWASP ASVS Guidelines)** 입력 검증 및 정제를 위한 OWASP 표준을 따름

효과적인 입력 검증 및 정제를 보장하기 위해 OWASP ASVS(애플리케이션 보안 검증 표준) 가이드라인을 따른다.

- **(Output Encoddng)** JavaScript 또는 Markdown에서 코드 실행을 방지하기 위해 LLM 출력을 인코딩

JavaScript 또는 Markdown에 의한 원치 않는 코드 실행을 완화하기 위해 모델 출력을 사용자에게 다시 인코딩한다. OWASP ASVS는 출력 인코딩에 대한 자세한 지침을 제공한다.

1.2.3 Training Data Poisoning

개념

"**Training Data Poisoning**"은 모델의 보안, 효과성 또는 윤리적 행동을 손상시킬 수 있는 취약성, 백도어 또는 편향을 도입하기 위한 데이터 조작 또는 미세 조정 프로세스를 말한다. 이는 성능 저하, 다운스트림 소프트웨어 악용 및 평판 손상의 위험이 있다.

예시

- (Malicious Data Injection) 모델 훈련 중에 위조된 데이터 주입
- (Biased Training Outputs) 모델은 오염된 데이터에서 부정확한 내용을 반영
- (Content Injection) 악의적인 행위자는 편향된 콘텐츠를 훈련에 주입

공격 시나리오

- **(오해의 소지가 있는 출력)** LLM은 편향이나 증오를 조장하는 콘텐츠를 생성

LLM 생성 AI 프롬프트 출력은 애플리케이션 사용자를 오도하여 편향된 의견, 추종자 또는 비정상적인 경우 증오 범죄 등으로 이어질 수 있다.

- **(유해 데이터 주입)** 악의적인 사용자가 편향된 데이터로 모델을 조작

학습 데이터가 올바르게 필터링되지 않으면 애플리케이션의 악의적인 사용자가 모델에 영향을 미치고 악성 데이터를 주입하여 편향되고 잘못된 데이터에 적응하려고 할 수 있다.

- **(악의적인 문서 주입)** 경쟁자가 모델 학습 중에 거짓 데이터를 삽입

악의적인 행위자 또는 경쟁자가 의도적으로 부정확하거나 악의적인 문서를 생성하여 모델의 학습 데이터를 표적으로 삼고 동시에 입력을 기반으로 모델을 훈련한다. 희생양이 된 모델(victim model)은 생성 AI 프롬프트의 출력에 반영되는 이 위조된 정보를 사용하여 훈련한다.

예방

- **(공급망 검증)** 외부 데이터 소스를 검증하고 "ML-BOM" 레코드를 유지·관리

특히 외부에서 소싱한 경우 학습 데이터의 공급망을 확인하고 "ML-BOM"(기계 학습 자재 목록) 방법을 통해 증명을 유지하고 모델 카드를 확인한다.

- **(합법성 검증)** 입력 검증을 위한 OWASP 표준을 따름

사전 교육, 미세 조정[7] 및 임베딩 프로세스[8] 모두에서 얻은 대상 데이터 소스와 포함된 데이터의 올바른 적법성을 확인한다.

- **(사용 사례별 교육)** JavaScript 또는 Markdown에서 코드 실행을 방지하기 위해 LLM 출력을 코딩

LLM과 통합할 애플리케이션에 대한 사용 사례를 확인한다. 별도의 교육 데이터를 통해 다른 모델을 만들거나 다른 사용 사례에 대한 미세 조정을 통해 정의된 사용 사례에 따라 보다 세부적이고 정확한 생성 AI 출력을 만든다.

7 이미 훈련된 기존 모델을 가져와 큐레이팅된 데이터 세트를 사용하여 훈련함으로써 더 좁은 주제나 더 집중된 목표에 맞게 조정하는 것을 포함한다. 이 데이터 세트에는 일반적으로 입력과 해당 원하는 출력의 예가 포함된다.

8 범주형 데이터(종종 텍스트)를 언어 모델을 훈련하는 데 사용할 수 있는 숫자 표현으로 변환하는 프로세스를 말하며, 텍스트 데이터의 단어나 구문을 연속 벡터 공간의 벡터로 표현하는 것을 포함한다. 벡터는 일반적으로 텍스트의 큰 코퍼스에서 훈련된 신경망에 텍스트 데이터를 입력하여 생성된다.

1.2.4 Model Denial of Service

개념

"**Model Denial of Service**"는 공격자가 LLM(Large Language Model)과 상호작용할 때 예외적으로 많은 양의 리소스를 소모할 때 발생한다. 이는 공격자와 다른 사용자에게 서비스 품질이 저하될 수 있으며, 잠재적으로 높은 리소스 비용이 발생할 수 있다.

예시

- (Malicious Data Injection) 모델 훈련 중에 위조된 데이터 주입
- (Biased Training Outputs) 모델은 오염된 데이터에서 부정확한 내용을 반영
- (Content Injection) 악의적인 행위자는 편향된 콘텐츠를 훈련에 주입

- 대량 대기열 : 공격자는 리소스 집약적 작업으로 LLM을 과부하시킴
- 리소스 소모 쿼리 : 비정상적인 쿼리는 시스템 리소스를 고갈시킴
- 지속적인 입력 오버플로(overflow) : 과도한 입력으로 LLM을 범람시킴
- 반복적인 긴 입력 : 반복적인 긴 쿼리는 리소스를 고갈시킴
- 재귀적 컨텍스트 확장 : 공격자는 재귀적 동작을 악용

공격 시나리오

- **(리소스 과다 사용)** 공격자가 호스팅된 모델을 과부하시켜 다른 사용자에게 영향을 미침

공격자는 호스팅 모델에 어렵고 비용이 많이 드는 여러 요청을 반복적으로 보내 다른 사용자에게 더 나쁜 서비스를 제공하고 호스트의 리소스 비용을 증가시킨다.

- **(Webpage Request Amplification)** LLM 도구가 예상치 못한 콘텐츠로 인해 과도한 리소스를 소모함

LLM 기반 도구가 악성이 아닌 쿼리에 응답하기 위해 정보를 수집하는 과정에서 웹페이지에 일부 텍스트가 있음을 발견하게 된다.

- **(Input Flood)** 과도한 입력으로 LLM을 압도하여 속도 저하

공격자는 컨텍스트 창을 초과하는 입력으로 LLM을 지속적으로 폭격한다. 공격자는 자동화된 스크립트나 도구를 사용하여 대량의 입력을 보내 LLM의 처리 기능을 압도할 수 있다. 결과적으로 LLM은 과도한 리소스를 소비하여 시스템이 상당히 느려지거나 완전히 응답하지 않게 된다.

- **(Sequential Input Drain)** 격자가 순차적 입력으로 컨텍스트 창을 고갈시킴

공격자는 각 입력이 컨텍스트 창의 용량 바로 아래에 있도록 설계된 일련의 순차적 입력을 LLM에 보낸다. 공격자는 이러한 입력을 반복적으로 제출하여 사용 가능한 컨텍스트 창 용량을 소진하려고 한다. LLM이 컨텍스트 창 내에서 각 입력을 처리하는 데 어려움을 겪으면서 시스템 리소스가 부족해져 성능이 저하되거나 서비스가 완전히 거부될 수 있다.

예방

- **(입력 검증)** 입력 검증 및 콘텐츠 필터링 구현

정의된 한도를 준수하고 악성 콘텐츠를 걸러내기 위해 사용자 입력에 대한 입력 검증을 구현한다.

- **(리소스 캡)** 요청당 리소스 사용 제한

요청 또는 단계당 리소스 사용량을 제한하여 복잡한 부분을 포함하는 요청이 더 느리게 실행되도록 한다.

- **(API 비율 제한)** 사용자 또는 IP 주소에 대한 비율 제한 적용

특정 기간 내에 개별 사용자 또는 IP 주소가 수행할 수 있는 요청 수를 제한하기 위해 API 속도 제한을 적용한다.

- **(큐 관리)** 대기 중인 작업 및 전체 작업 제어

LLM 응답에 반응하는 시스템에서 대기 중인 작업 수와 총 작업 수를 제한한다.

- **(리소스 모니터링)** 리소스 사용을 지속적으로 모니터링함

Dos 공격을 나타낼 수 있는 비정상적인 급증이나 패턴을 식별하기 위해 LLM의 리소스 사용률을 지속적으로 모니터링한다.

1.2.5 Supply Chain Vulnerabilities

개념

LLM의 공급망 취약성은 교육 데이터, ML 모델 및 배포 플랫폼을 손상시켜 편향된 결과, 보안 침해 또는 전체 시스템 장애를 일으킬 수 있다. 이러한 취약성은 오래된 소프트웨어, 취약한 사전 훈련된 모델, 오염된 교육 데이터 및 안전하지 않은 플러그인 디자인에서 비롯될 수 있다.

예시

- 패키지 Null 취약점: 오래된 구성 요소 사용
- Nulnerable 모델: 미세 조정을 위한 위험한 사전 학습 모델
- Poisoned Data: 오염된 크라우드 소싱 데이터
- Oldated 모델: 유지·관리되지 않은 모델 사용
- Unclear Terms: 불분명한 용어로 인한 데이터 오용

공격 시나리오

- **(Library Exploitation)** 취약한 Python 라이브러리 Exploiting

공격자가 취약한 Python 라이브러리를 악용하여 시스템을 손상시킨다. 이는 첫 번째 Open AI 데이터 침해에서 발생했다.

- **(Scamming Plugin)** 사기를 위한 플러그인 배포

공격자가 항공편을 검색하는 LLM 플러그인을 제공하여 사용자를 사기치는 가짜 링크를 생성한다.

- **(패키지 레지스트리 공격)** 손상된 패키지로 개발자 속이기

공격자가 PyPi 패키지 레지스트리를 악용하여 모델 개발자를 속여 손상된 패키지를 다운로드하고 데이터를 빼내거나 모델 개발 환경에서 권한을 확대한다.(실제 공격이었음)

- **(잘못된 정보 백도어)** 가짜 뉴스를 위해 모델 Poisoning

공격자는 경제 분석 및 사회 연구에 특화된 공개적으로 사용 가능한 사전 훈련된 모델을 오염시켜 잘못된 정보와 가짜 뉴스를 생성하는 백도어를 만든다. 그들은 피해자가 사용할 수 있도록 모델 마켓플레이스(예: Hugging Face)에 배포한다.

- **(데이터 Poisoning)** 미세 조정 중 데이터 세트 Poisoning

공격자가 공개적으로 사용 가능한 데이터 세트를 오염시켜 모델을 미세 조정할 때 백도어를 만드는 데 도움을 준다. 백도어는 다른 시장의 특정 회사를 미묘하게 선호한다.

예방

- **(공급업체 평가)** 공급업체 및 정책 검토

신뢰할 수 있는 공급업체만 사용하고, 이용 약관 및 개인정보 보호 정책을 포함한 데이터 소스 및 공급업체를 신중하게 검토한다. 적절하고 독립적으로 감사된 보안이 적용되고 모델 운영자 정책이 데이터 보호 정책과 일치하는지 확인한다. 즉, 데이터가 모델 학습에 사용되지 않도록 한다. 마찬가지로 모델 유지 관리자로부터 저작권이 있는 자료를 사용하지 않도록 보장하고 법적 완화 조치를 모색한다.

- **(플러그인 테스트)** 테스트되고 신뢰할 수 있는 플러그인 사용

평판이 좋은 플러그인만 사용하고 애플리케이션 요구 사항에 대해 테스트되었는지 확인한다. LLM-Insecure Plugin Design은 타사 플러그인 사용으로 인한 위험을 완화하기 위해 테스트해야 하는 Insecure Plugin 디자인의 LLM 측면에 대한 정보를 제공한다.

- **(OWASP A06)** 오래된 구성 요소 위험 완화

OWASP Top Ten's A06:2021 - 취약하고 오래된 구성 요소에서 발견되는 완화책을 이해하고 적용한다. 여기에는 취약성 스캐닝, 관리 및 패치 구성 요소가 포함된다. 민감한 데이터에 액세스할 수 있는 개발 환경의 경우 해당 환경에도 이러한 통제를 적용한다.

- **(재고 관리)** 최신 재고 유지

소프트웨어 자재 목록(SBOM)을 사용하여 최신 구성 요소 인벤토리를 유지·관리하여 최신의 정확하고 서명된 인벤토리를 유지하고 배포된 패키지의 변조를 방지한다. SBOM은 새로운 제로데이트 취약성을 신속하게 감지하고 경고하는 데 사용할 수 있다.

- **(보안 조치)** 모델 및 코드 서명, 이상 탐지 적용 및 모니터링

모델과 그 아티팩트(artifacts)를 포함하여, 구성 요소 및 환경 취약성 스캐닝, 승인되지 않은 플러그인 사용, 오래된 구성 요소 등을 포괄하는 충분한 모니터링을 구현한다.

1.2.6 Sensitive Information Disclosure

개념

LLM 애플리케이션은 실수로 민감한 정보, 알고리즘 또는 기밀 데이터를 공개하여 무단 액세스, 지적 재산권 도용 및 개인정보 침해로 이어질 수 있다. 이러한 위험을 완화하기 위해 LLM 애플리케이션은 데이터를 정제(sanitization)하고 적절한 사용 정책을 구현하며 LLM에서 반환하는 데이터 유형을 제한해야 한다.

예시

- complete Filtering (완료 필터링): LLM 응답에는 민감한 데이터가 포함될 수 있음
- Overfitting (과잉적합): LLM은 학습 중에 민감한 데이터를 기억할 수 있음
- 의도치 않은 공개: 잘못된 해석이나 스크러빙 부족으로 인한 데이터 유출

공격 시나리오

- **(의도치 않은 노출)** 사용자 A가 다른 사용자 데이터에 노출됨

> 의심하지 않는 합법적인 사용자 A는 LLM 애플리케이션과 악의적이지 않은 방식으로 상호 작용할 때 LLM을 통해 특정 다른 사용자 데이터에 노출된다.

- **(필터 우회)** 사용자 A가 필터를 우회하여 PII 추출

> 사용자 A는 LLM에서 입력 필터와 sanitization을 우회하기 위해 잘 만들어진 일련의 프롬프트를 대상으로 하여 애플리케이션의 다른 사용자에 대한 민감한 정보(PII)를 공개한다.

- **(학습 데이터 유출)** 학습 중에 개인 데이터가 유출됨

> PII와 같은 개인 데이터는 사용자 자신의 과실 또는 LLM 애플리케이션으로 인해 훈련 데이터를 통해 모델로 유출된다.

예방

- **(Data Sanitization)** 스크러빙을 사용하여 학습에서 사용자 데이터를 방지

> 적절한 데이터 Sanitization 및 스크러빙 기술을 통합하여 사용자 데이터가 학습 모델 데이터에 입력되는 것을 방지한다.

- **(입력 검증)** 악성 입력을 필터링하여 모델 포이즌을 방지

> 강력한 입력 검증 및 Sanitization 방법을 구현하여 잠재적인 악성 입력을 식별하고 필터링하여 모델이 오염되는 것을 방지한다.

- **(미세 조정 주의)** 모델 미세 조정에서 민감한 데이터를 주의한다.

미세 조정 데이터에서 민감한 것으로 간주되는 모든 사항은 사용자에게 공개될 가능성이 있다. 따라서 최소 권한 규칙을 적용하고, 가장 높은 권한을 가진 사용자가 액세스할 수 있고 권한이 낮은 사용자에게 표시될 수 있는 정보에 대해 모델을 학습시키면 않는다.

- **(데이터 액세스 제어)** 외부 데이터 소스 액세스를 제한

외부 데이터 소스에 대한 액세스(런타임 시 데이터 오케스트레이션)는 제한되어야 한다. 외부 데이터 소스에 대한 엄격한 액세스 제어 방법과 안전한 공급망을 유지하기 위한 엄격한 접근 방식을 적용한다.

1.2.7 Insecure Plugin Design (안전하지 않은 플러그인 디자인)

개념

플러그인은 악의적인 요청으로 이어질 수 있는 취약한 결과를 초래하여 데이터 유출, 원격 코드 실행, 불충분한 액세스 제어 및 부적절한 입력 검증으로 인한 권한 상승과 같은 해로운 결과를 초래할 수 있다. 개발자는 악용을 방지하기 위해 엄격한 매개변수화된 입력 및 안전한 액세스 제어 지침과 같은 강력한 보안 조치를 따라야 한다.

예시

- (단일 필드 매개변수) 플러그인에 매개변수 분리가 없음
- (구성 문자열) 구성이 설정을 재정의할 수 있음
- (인증 문제) 특정 플러그인 권한 부여가 없음
- (원시 SQL 또는 코드) 코드 또는 SQL을 안전하지 않게 수용

공격 시나리오

- **(URL Manipulation)** 공격자는 조작된 URL을 통해 콘텐츠를 삽입함

플러그인은 기본 URL을 수락하고 LLM에 URL을 쿼리와 결합하여 사용자 요청을 처리하는 데 포함된 날씨예보를 얻도록 지시한다. 악의적인 사용자는 URL이 제어하는 도메인을 가리키도록 요청을 작성할 수 있으며, 이를 통해 도메인을 통해 LLM 시스템에 자체 콘텐츠를 삽입할 수 있다.

- **(Reconnaissance and Exploitation)** 코드 실행 및 데이터 도난에 대한 검증 부족을 악용함

플러그인은 검증하지 않는 단일 필드에 자유 형식 입력을 허용한다. 공격자는 오류 메시지에서 정찰(Reconnaissance)을 수행하기 위해 신중하게 제작된 페이로드를 제공한다. 그런 다음 알려진 타사 취약성을 악용하여 코드를 실행하고 데이터 추출 또는 권한 상승을 수행한다.

- **(Unauthorized Access)** 매개변수 조작을 통해 무단 데이터에 액세스

벡터 저장소에서 임베딩을 검색하는데 사용되는 플러그인은 유효성 검사 없이 연결 문자열로 구성 매개변수를 허용한다. 이를 통해 공격자는 이름이나 호스트 매개변수를 변경하여 다른 벡터 저장소에 액세스하고 액세스해서는 안 되는 임베딩을 추출하여 실험할 수 있다.

- **(Repository Takeover)** 저장소 인수를 위해 안전하지 않은 코드 관리 플러그인을 악용함

공격자는 간접 프롬프트 주입을 사용하여 입력 유효성 검사 및 취약한 액세스 제어가 없는 안전하지 않은 코드 관리 플러그인을 악용하여 저장소 소유권을 이전하고 사용자를 저장소에서 잠근다.

예방

- **(매개변수 제어)** 유형 검사를 시행하고 검증 계층을 사용함

플러그인은 가능한 한 엄격한 매개변수화된 입력을 적용하고 입력에 대한 유형 및 범위 검사를 포함해야 한다. 이것이 불가능한 경우, 요청을 구문 분석하고 검증을 적용하는 두 번째 계층의 유형화된 호출을 도입해야 한다. 애플리케이션 의미론(semantics) 때문에 자유형(freeform) 입력을 허용해야 하는 경우, 잠재적으로 유해한 메서드(mathods)가 호출되지 않는지 확인하기 위해 신중하게 검사해야 한다.

- **(OWASP 지침)** ASVS 권장 사항을 적용함

플러그인 개발자는 ASVS(애플리케이션 보안 검증 표준)에서 OWASP의 권장 사항을 적용하여 적절한 입력 검증을 보장해야 한다.

- **(철저한 테스트)** SAST, DAST, IAST로 검사하고 테스트함

플러그인은 적절한 검증을 보장하기 위해 철저히 검사하고 테스트해야 한다. 개발 파이프라인에서 정적 애플리케이션 보안 테스트(SAST) 검사와 동적 및 대화형 애플리케이션 테스트(DAST, IAST)를 사용한다.

- **(최소 권한)** ASVS 접근 통제 가이드라인을 따름

플러그인은 OWASP ASVS 접근 통제 지침에 따라 안전하지 않은 입력 매개변수 악용의 영향을 최소화하도록 설계해야 한다. 여기서는 최소 권한 접근통제를 포함하여 원하는 기능을 수행하면서 가능한 적은 기능을 노출한다.

- **(인증 ID)** 사용자 지정 인증에 OAuth2 및 API 키를 사용

플러그인은 OAuth2와 같은 적절한 인증 ID를 사용하여 효과적인 권한 부여 및 접근 통제를 적용해야 한다. 또한 API 키는 기본 대화형 사용자 경로가 아닌 플러그인 경로를 반영하는 사용자 정의 권한 부여 결정에 대한 컨텍스트를 제공하는 데 사용해야 한다.

1.2.8 Excessive Agency (과도한 대행)

개념

LLM 기반 시스템의 과도한 에이전시는 과도한 기능, 과도한 권한 또는 너무 많은 자율성으로 인해 발생하는 취약성이다. 이를 방지하기 위해 개발자는 플러그인 기능, 권한 및 자율성을 절대적으로 필요한 것으로 제한하고, 사용자 권한을 추적하고, 모든 작업에 대해 사람의 승인을 요구하고, 다운스트림 시스템에서 권한을 구현해야 한다.

예시

- (과도한 기능) LLM 에이전트는 불필요한 기능을 가지고 있어 오용의 위험이 있다.
- (과도한 권한) 플러그인은 시스템에 과도하게 접근할 수 있다.
- (과도한 자율성) LLM은 영향력이 큰 작업에 대한 인간 검증이 부족하다.

공격 시나리오

- 과도한 권한과 자율성을 가진 LLM 기반 개인 비서 앱이 악성 이메일에 속아 스팸을 보냄. 이는 기능·권한을 제한하거나, 사용자 승인을 요구하거나, 속도 제한을 구현하여 방지할 수 있음.

LLM 기반 개인 비서 앱은 플러그인을 통해 개인의 사서함에 액세스하여 수신 이메일의 내용을 요약한다. 이 기능을 구현하려면 이메일 플러그인에 메시지를 읽을 수 있는 기능이 필요하지만, 시스템 개발자가 사용하기로 선택한 플러그인에는 메시지를 보내는 기능도 포함되어 있다. LLM은 간접 프롬프트 주입 공격에 취약하여 악의적으로 작성된 수신 이메일이 LLM을 속여 이메일 플러그인에 '메시지 보내기' 기능을 호출하여 사용자의 사서함에서 스팸을 보내도록 명령한다. 이를 방지할 수 있는 방법은 다음과 같다; (a) 메일 읽기 기능만 제공하는 플러그인을 사용하여 과도한 기능을 제거, (b) 읽기 전용 범위의 OAuth 세션을 통해 사용자의 이메일 서비스에 인증하여 과도한 권한을 제거, (c) LLM 플러그인에서 초안한 모든 메일을 사용자가 수동으로 검토하고 '보내기'를 클릭하도록 요구하여 과도한 자율성을 제거하거나 또는 메일 전송 인터페이스에 속도 제한을 구현하여 발생한 피해를 줄일 수 있다.

예방

- **(플러그인 기능 제한)** LLM 에이전트에 필수적인 기능만 허용

LLM 에이전트가 호출할 수 있는 플러그인/도구를 필요한 최소한의 기능으로만 제한한다. 예를 들어, LLM 기반 시스템이 URL의 내용을 가져오는 기능을 요구하지 않는 경우 이러한 플러그인은 LLM 에이전트에게 제공되어서는 안 된다.

- **(플러그인 범위 제어)** LLM 플러그인 내의 기능 제한

LLM 플러그인/도구에 구현된 기능을 필요한 최소한으로 제한한다. 예를 들어, 이메일을 요약하기 위해 사용자의 사서함에 액세스하는 플러그인은 이메일을 읽는 기능만 필요할 수 있으므로 플러그인에는 메시지 삭제 또는 전송과 같은 다른 기능이 포함되어서는 안 된다.

- **(세분화된 기능)** 개방형 기능은 피하고 특정 플러그인 사용

가능한 경우 개방형 기능(예: 셸 명령 실행, URL 가져오기 등)을 피하고 더 세부적인 기능이 있는 플러그인/도구를 사용한다. 예를 들어, LLM 기반 앱은 파일에 출력을 써야 할 수 있다. 이것이 셸 함수를 실행하기 위한 플러그인을 사용하여 구현된 경우 바람직하지 않은 동작의 범위가 매우 크다(다른 셸 명령이 실행될 수 있음). 더 안전한 대안은 해당 특정 기능만 지원할 수 있는 파일 쓰기 플러그인을 빌드하는 것이다.

- **(권한 제어)** 필요한 최소한으로 권한 제한

LLM 플러그인/도구가 다른 시스템에 부여하는 권한을 최소한으로 제한하여 바람직하지 않은 작업의 범위를 제한한다. 예를 들어, 고객에게 구매를 권장하기 위해 제품 데이터베이스를 사용하는 LLM 에이전트는 '제품' 테이블에 대한 읽기 액세스만 필요할 수 있다. 다른 테이블에 대한 액세스 권한이나 레코드를 삽입, 업데이트 또는 삭제할 수 있는 권한이 없어야 한다. 이는 LLM 플러그인이 데이터베이스에 연결하는 데 사용하는 ID에 대한 적절한 데이터베이스 권한을 적용하여 시행해야 한다.

- **(사용자 인증)** 작업이 사용자 컨텍스트에 있는지 확인

사용자 권한 부여 및 보안 범위를 추적하여 사용자를 대신하여 수행된 작업이 해당 특정 사용자의 컨텍스트에서 다운스트림 시스템에서 실행되고 필요한 최소 권한으로 실행되도록 한다. 예를 들어, 사용자의 코드 repo를 읽는 LLM 플러그인은 사용자가 OAuth를 통해 인증하고 필요한 최소 범위로 인증하도록 요구해야 한다.

- **(human-in-the-Loop)** 작업에 대한 인간 승인 요구

모든 작업을 수행하기 전에 사람이 승인하도록 요구하는 human-in-the-Loop 제어를 활용한다. 이는 다운스트림 시스템(LLM 애플리케이션 범위 밖 또는 LLM 플러그인/도구 자체 내부)에서 구현될 수 있다. 예를 들어, 사용자를 대신하여 소셜 미디어 콘텐츠를 만들고 게시하는 LLM 기반 앱은 '게시' 작업을 구현하는 플러그인/도구/API 내에 사용자 승인 루틴을 포함해야 한다.

- **(다운스트림 권한 부여)** 다운스트림 시스템에서 권한 부여 구현

LLM에 의존하여 작업이 허용되는지 여부를 결정하는 대신 다운스트림 시스템에서 권한을 구현한다. 도구/플러그인을 구현할 때 플러그인/도구를 통해 다운스트림 시스템에 요청된 모든 요청이 보안 정책에 대해 검증되도록 완전한 중재 원칙(mediation principle)을 적용한다.

1.2.9 Overreliance (지나친 의존)

개념

LLM에 대한 과도한 의존은 잘못된 정보, 법적 문제, 보안 취약성과 같은 심각한 결과를 초래할 수 있다. 이는 LLM이 적절한 감독이나 검증 없이 중요한 결정을 내리거나 콘텐츠를 생성할 수 있다고 신뢰될 때 발생한다.

예시

- (과도한 기능) LLM 에이전트는 불필요한 기능을 가지고 있어 오용의 위험이 있다.
- (과도한 권한) 플러그인은 시스템에 과도하게 접근할 수 있다.
- (과도한 자율성) LLM은 영향력이 큰 작업에 대한 인간 검증이 부족하다.

- **(오해의 소지가 있는 정보)** LLM은 검증 없이 오해의 소지가 있는 정보를 제공할 수 있음
- **(불안전한 코드)** LLM은 소프트웨어에서 안전하지 않은 코드를 제안할 수 있음

공격 시나리오

- **(잘못된 정보 확산)** 악의적인 행위자가 LLM에 의존하는 뉴스 기관을 악용

뉴스 기관은 LLM을 사용하여 뉴스 기사를 생성한다. 악의적인 행위자가 이러한 과도한 의존성을 악용하여 LLM에 오해의 소지가 있는 정보를 제공하고 잘못된 정보가 퍼지게 한다.

- **(Plagiarism)** 의도치 않은 표절로 인해 저작권 문제가 발생

AI가 의도치 않게 콘텐츠를 표절하여 저작권 문제가 발생하고 기관에 대한 신뢰가 감소한다.

- **(안전하지 않은 소프트웨어)** LLM 제안으로 보안 취약성이 발생

소프트웨어 개발팀이 LLM 시스템을 사용하여 코딩 프로세스를 가속화한다. AI의 제안에 지나치게 의존하면 안전하지 않은 기본 설정이나 안전한 코딩 관행과 일치하지 않는 권장 사항으로 인해 애플리케이션에 보안 취약성이 발생한다.

- **(악성 패키지)** LLM이 존재하지 않는 코드 라이브러리를 제안

소프트웨어 개발 회사가 LLM을 사용하여 개발자를 지원한다. LLM은 존재하지 않는 코드 라이브러리 또는 패키지를 제안하고 AI를 신뢰하는 개발자는 자신도 모르게 악성 패키지를 회사 소프트웨어에 통합한다. 이는 특히 타사 코드나 라이브러리를 포함할 때 LLM 제안을 교차 확인하는 것의 중요성을 강조한다.

예방

- **(모니터링 및 검증)** 일관성(consistency) 검사를 통해 LLM 출력을 정기적으로 검토

LLM 출력을 정기적으로 모니터링하고 검토한다. 자체 일관성 또는 투표 기술을 사용하여 일관되지 않은 텍스트를 필터링한다. 단일 프롬프트에 대한 여러 모델 응답을 비교하면 출력의 품질과 일관성을 더 잘 판단할 수 있다.

- **(교차 확인)** 신뢰할 수 있는 출처로 LLM 출력 확인

신뢰할 수 있는 외부 소스와 LLM 출력을 교차 확인한다. 이 추가 검증 계층은 모델에서 제공하는 정보가 정확하고 신뢰할 수 있는지 확인하는 데 도움이 될 수 있다.

- **(미세 조정)** 작업별 미세 조정으로 LLM 품질 향상

미세 조정 또는 임베딩을 사용하여 모델을 향상시켜 출력 품질을 개선한다. 일반적으로 사전 학습된 모델은 특정 도메인에서 조정된 모델에 비해 부정확한 정보를 생성할 가능성이 더 높다. 프롬프트 엔지니어링, 매개변수 효율적 조정(PET), 전체 모델 조정 및 사고 체인 프롬프트와 같은 기술을 이러한 목적으로 사용할 수 있다.

- **(자동 검증)** 알려진 사실과 출력을 검증하는 시스템 구현

생성된 출력을 알려진 사실 또는 데이터와 교차 검증할 수 있는 자동 검증 메커니즘을 구현한다. 이를 통해 보안 계층을 추가로 제공하고 할루시네이션(Hallucination)[9]과 관련된 위험을 완화할 수 있다.

- **(작업 세분화)** 복잡한 작업을 나누어 위험 감소

복잡한 작업을 관리 가능한 하위 작업으로 나누고 이를 다른 에이전트에 할당한다. 이는 복잡성을 관리하는 데 도움이 될 뿐만 아니라 각 에이전트가 더 작은 작업에 대해 책임을 질 수 있으므로 Hallucination의 가능성도 줄어든다.

- **(위험 전달)** LLM 제한 사항 전달

LLM 사용과 관련된 위험과 한계를 일찍 전달한다. 여기에는 정보 부정확성 및 기타 위험이 포함된다. 효과적인 위험 전달은 사용자를 잠재적인 이슈에 대비시키고 정보에 입각한 결정을 내리는 데 도움이 될 수 있다.

- **(사용자 친화적 인터페이스)** 콘텐츠 필터 및 경고가 있는 인터페이스 생성

LLM의 책임감 있고 안전한 사용을 장려하는 API와 사용자 인터페이스를 구축한다. 여기에는 콘텐츠 필터, 잠재적 부정확성에 대한 사용자 경고, AI 생성 콘텐츠의 명확한 레이블 지정과 같은 조치가 포함될 수 있다.

- **(보안 코딩)** 취약점을 방지하기 위한 지침 수립

개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립한다.

9 AI 모델이 정확하지 않거나 사실이 아닌 조작된 정보를 생성하는 것

1.2.10 Model Theft

개념

LLM 모델 도난에는 LLM 모델에 대한 무단 액세스 및 유출이 포함되며, 경제적 손실, 평판 손상 및 민감한 데이터에 대한 무단 액세스 위험이 있다. 이러한 모델을 보호하려면 강력한 보안 조치가 필수적이다.

예시

- (취약점 악용) 보안 결함으로 인한 무단 액세스
- (중앙 모델 레지스트리) 거버넌스를 위한 중앙 보안
- (내부 위협) 직원 모델 유출 위험
- (사이드 채널 공격) 사이드 기술을 통한 모델 세부 정보 추출

공격 시나리오

- **(모델 도난)** 경쟁을 위한 무단 액세스 및 사용

공격자가 회사 인프라의 취약점을 악용하여 LLM 모델 저장소에 대한 무단 액세스를 얻는다. 공격자는 귀중한 LLM 모델을 추출하여 경쟁 언어 처리 서비스를 시작하거나 민감한 정보를 추출하여 원래 회사에 상당한 재정적 피해를 입힌다.

- **(직원 유출)** 노출로 인한 위험 증가

불만이 있는 직원이 모델이나 관련 아티팩트를 유출한다. 이 시나리오를 대중에 공개하면 공격자가 적대적 공격을 위한 지식을 얻거나, 아니면 사용 가능한 자산을 직접 훔칠 수 있다.

- **(섀도우 모델 생성)** 쿼리로 모델 복제

공격자가 신중하게 선택한 입력으로 API를 쿼리하고 충분한 수의 출력을 수집하여 섀도우(shadow) 모델을 만든다.

- **(사이드 채널 공격)** 사이드 기술을 통한 추출

악의적인 공격자는 LLM의 입력 필터링 기술과 프리앰블(preamble)을 우회하여 사이드 채널 공격을 수행하고 자신이 제어하는 원격 제어 리소스에 대한 모델 정보를 검색한다.

예방

- **(접근 통제 및 인증)** 강력한 접근 통제 및 인증

강력한 액세스 제어(예: RBAC 및 최소 권한 규칙)와 강력한 인증 메커니즘을 구현하여 LLM 모델 저장소와 학습 환경에 대한 무단 액세스를 제한한다.
이는 특히 처음 세 가지 일반적인 예에 해당하는데, 내부 위협, 잘못된 구성 및/또는 악의적인 행위자가 내부자 또는 환경 외부에서 침투할 수 있는 LLM 모델, 가중치 및 아키텍처를 수용하는 인프라에 대한 취약한 보안 제어로 인해 이러한 취약성이 발생할 수 있다.
공급업체 관리 추적, 검증 및 종속성 취약성은 공급망 공격의 악용을 방지하기 위한 중요한 주제이다.

- **(네트워크 제한)** 리소스 및 API에 대한 LLM 액세스 제한

LLM의 네트워크 리소스, 내부 서비스 및 API에 대한 액세스 제한
이것은 특히 내부자 위험과 위협을 다루기 때문에 모든 일반적인 예에 해당하지만 궁극적으로 LLM 애플리케이션이 “액세스할 수 있는” 것을 제어하므로 사이드 채널 공격을 방지하기 위한 메커니즘 또는 예방 단계가 될 수 있다.

- **(모니터링 및 감사)** 액세스 로그의 정기적 모니터링

LLM 모델 저장소와 관련된 액세스 로그 및 활동을 정기적으로 모니터링하고 감사하여 의심스럽거나 승인되지 않은 동작을 즉시 감지하고 대응한다.

- **(MLOps 자동화)** 승인 워크플로를 통한 보안 배포

거버넌스, 추적 및 승인 워크플로를 사용하여 MLOps 배포를 자동화하여 인프라 내에서 액세스 및 배포 제어를 강화한다.

02 NIST Trustworthy and Responsible AI NIST AI 100-2e2023

2.1 개요

- AI 시스템은 기능에 따라 예측 AI(Predictive AI, "이하 PredAI"))와 생성 AI(Generative AI, "이하 GenAI")의 두 가지로 구분할 수 있음
  - 공격자는 PredAI 및 GenAI 시스템 모두에 대한 학습 데이터를 조작하여 이를 기반으로 교육된 AI 시스템을 공격에 취약하게 만들 수 있음
    - ML 모델의 크기가 계속 커짐에 따라 많은 조직에서 직접 사용하거나 새로운 데이터 세트로 미세 조정하여 다양한 작업을 수행할 수 있는 사전 훈련된 모델에 의존하는 경향이 있는데, 이는 공격자가 모델 가용성을 손상시키거나 잘못된 처리를 강제하거나 지시를 받을 때 데이터를 유출할 수 있도록 TROJANS를 삽입하여 사전 훈련된 모델을 악의적으로 수정할 수 있는 기회를 제공함
- 이 보고서는 「NIST AI 위험 관리 프레임워크」에서 ML 시스템의 보안, 회복성 및 견고성 개념을 채택함. 보안, 회복성 및 견고성은 위험에 의해 측정되며, 위험은 엔터티(예컨대, 시스템)가 잠재적인 상황이나 이벤트(예: 공격)에 의해 위협받는 정도와 그러한 이벤트가 발생할 경우 결과의 심각성을 측정하는 것임
  - 이 보고서는 적대적 머신 러닝(AML)과 관련된 위험을 식별, 해결 및 관리하는 데 중점을 두고 있으며, 다음을 개발하기 위한 지침을 제공함
    - ML 및 사이버 보안 커뮤니티에서 사용할 AML의 표준화된 용어
    - PredAI 시스템에 대한 회피, 포이즈닝 및 프라이버시 공격 대응
    - GenAI 시스템에 대한 회피, 포이즈닝, 개인정보 보호 및 남용(abuse) 공격대응
    - 여러 데이터 모달리티(modalities)에 걸친 모든 실행 가능한 학습 방법(예: 지도, 비지도, 반지도 학습, 연합 학습, 강화 학습)에 대한 공격과 기존 완화책 기법의 한계에 대한 논의

2.2 예측 AI 분류

- **공격 분류**
  - **(학습 단계)** 머신 러닝은 모델을 학습하는 Training Stage와 예측을 생성하기 위해 레이블이 지정되지 않은 새 데이터 샘플에 모델을 배포하는 Deployment Stage 구성. Supervised Learning의 경우 레이블이 지정된 학습 데이터가 학습 단계에서 학습 알고리즘에 입력으로 제공되고 ML 모델은 특정 손실 함수를 최소화하도록 최적화됨. ML 모델의 검증 및 테스트는 일반적으로 모델이 실제 세계에 배포되기 전에 수행되며, 지도 학습 기술에는 예측 레이블 또는 클래스가 불연속형인 Classification과 예측 레이블 또는 응답 변수가 연속형인 Regression이 포함됨

- **학습 시간 공격**(Training-time attacks) : ML 학습 단계 동안의 공격을 Poisoning Attacks라고 함

Data Poisoning 공격	Model Poisoning 공격
공격자는 학습 샘플을 삽입하거나 수정하여 학습 데이터의 하위 집합을 제어	공격자는 모델과 해당 매개변수를 제어
모든 학습 패러다임에 적용 가능	클라이언트가 로컬 모델 업데이트를 집계 서버로 보내는 연합 학습과 모델 기술 공급업체가 모델에 악성 코드를 추가할 수 있는 공급망 공격에서 가장 흔히 발생

- **배포 시간 공격**(Deployment-time attacks) : 회피 공격은 테스트 샘플을 수정하여 적대적 사례를 생성하고, 멤버십 추론 및 데이터 재구성과 같은 프라이버시 공격은 일반적으로 ML 모델에 대한 쿼리 액세스 권한이 있는 공격자가 마운트함

● **(공격자의 목표 및 목적)** 공격자 목표는 가용성·무결성·기밀성에 따라 세 가지로 분류

- 가용성 침해(Availability Breakdown) : 가용성 공격은 공격자가 배포 시점에 모델의 성능을 분석하려고 시도하는 ML에 대한 무차별 공격이며, 공격자가 학습 세트의 일부를 제어할 때 데이터 포이즈닝을 통해 마운트될 수 있음
- 무결성 침해(Integrity Violations) : 무결성 공격은 ML 모델 출력의 무결성을 표적으로 삼아 ML 모델이 잘못된 예측을 수행하게 함. 공격자는 배포 시 회피 공격을 마운트하거나 학습 시 포이즌 공격을 마운트하여 무결성 위반을 일으킬 수 있음

포이즈닝을 통한 무결성 공격

Targeted Poisoning 공격	몇몇 타겟 샘플의 무결성을 침해하려 하며 공격자가 포이즈닝된 샘플을 삽입하기 위한 훈련 데이터 제어를 가지고 있다고 가정함
Backdoor Poisoning 공격	포이즈닝된 샘플과 테스트 샘플에 모두 추가되어 오분류를 유발하는 Backdoor Pattern을 생성해야 함
Model Poisoning 공격	타겟 또는 Backdoor 공격으로 이어질 수 있으며 공격자는 무결성 위반을 유발하기 위해 모델 매개변수를 수정함. 이들은 중앙 집중식 학습과 연합 학습을 위해 설계되었음

- 개인정보 침해(Privacy Compromise) : 공격자는 훈련 데이터(Data Privacy 공격으로 이어짐) 또는 ML 모델(Model Privacy 공격으로 이어짐)에 대한 정보를 알아내는 데 관심이 있을 수 있음

- **(공격자의 역할)** 공격자는 다음의 6가지 유형의 역량을 사용할 수 있음

훈련 데이터 제어	공격자는 훈련 샘플을 삽입하거나 수정하여 훈련 데이터의 하위 집합을 제어할 수 있음. 이 역량은 데이터 포이즈닝 공격(예: 가용성 포이즈닝, 타겟팅 또는 백도어 포이즈닝)에 사용됨
모델 제어	공격자는 트로이 목마 트리거를 생성하여 모델에 삽입하거나 연합 학습에서 악의적인 로컬 모델 업데이트를 보내 모델 매개변수를 제어할 수 있음
테스트 데이터 제어	공격자는 이를 활용하여 모델 배포 시 테스트 샘플에 교란을 추가할 수 있으며, 이는 회피 공격에서 적대적 예를 생성하거나 백도어 포이즈닝 공격에서 수행하는 것과 같음
레이블 제한	이 기능은 감독 학습에서 학습 샘플의 레이블에 대한 적대적 제어를 제한하는 데 적합함. 클린 레이블 포이즈닝 공격은 공격자가 포이즈닝된 샘플의 레이블을 제어하지 않는다고 가정함
소스 코드 제어	공격자는 난수 생성기나 종종 오픈 소스인 타사 라이브러리와 같은 ML 알고리즘의 소스 코드를 수정할 수 있음
쿼리 액세스	ML 모델이 클라우드 공급자(Machine Learning as a Service – MLaaS 사용)에 의해 관리되는 경우 공격자는 모델에 쿼리를 제출하고 예측(레이블 또는 모델 신뢰도)을 받을 수 있음

- **(공격자 지식)** 화이트박스, 블랙박스, 그레이박스의 세 가지 주요 유형이 있음

화이트박스 공격	공격자가 훈련 데이터, 모델 아키텍처, 모델 하이퍼 매개변수를 포함하여 ML 시스템에 대한 모든 지식을 가지고 운영한다고 가정함
블랙박스 공격	ML 시스템에 대한 최소한의 지식이 있다고 가정함. 공격자는 모델에 대한 쿼리 액세스를 얻을 수 있지만 모델이 어떻게 훈련되는지에 대한 다른 정보는 없음. 이러한 공격은 공격자가 AI 시스템에 대한 지식이 없다고 가정하고 일반적으로 사용할 수 있는 시스템 인터페이스를 활용하기 때문에 가장 일반적임
그레이박스 공격	블랙박스와 화이트박스 공격 사이의 적대적 지식을 포착

- **(데이터 모달리티)** 최근까지 대부분의 공격과 방어는 단일 모달리티에서 작동했지만 새로운 ML 추세는 다중 모달(multimodal) 데이터를 사용하는 것임. 적대적 ML 문헌에서 가장 일반적인 데이터 모달리티는 다음과 같음

이미지	이미지 데이터 모달리티의 적대적 예는 그래디언트(gradient) 기반 방법을 직접 적용할 수 있음
텍스트	자연어 처리(NLP)는 인기 있는 모달리티이며 회피, 포이즈닝, 프라이버시를 포함한 모든 종류의 공격이 NLP 애플리케이션에 적용되었음
오디오	오디오 시스템과 오디오 신호에서 생성된 텍스트도 공격을 받았음
비디오	비디오 이해 모델은 시각 및 언어 작업에서 점점 더 많은 기능을 보여주었지만 이러한 모델도 공격에 취약함
사이버 보안2	최초의 포이즈닝 공격은 웜 서명 생성 및 스팸 이메일 분류를 위한 사이버 보안에서 발견되었음
표 형식 데이터	금융, 비즈니스 및 의료 애플리케이션에서 표 형식 데이터를 사용하는 ML 모델에 대한 수많은 공격이 입증되었음

회피 공격과 완화책(Evasion Attacks and Mitigations)

- **(화이트박스 회피 공격)** 원래 테스트 샘플에서 가까운 거리에 적대적 사례를 생성하는 회피 공격을 설계하기 위한 여러 가지 최적화 기반 방법이 있음. 또한 거리 측정법, 범용 회피 공격 및 물리적으로 실현 가능한 공격에 대한 여러 가지 선택 사항과 NLP, 오디오, 비디오 및 사이버 보안 도메인을 포함한 여러 데이터 모달리티에 대해 개발된 회피 공격이 있음
- **(블랙박스 회피 공격)** 공격자가 모델 아키텍처나 학습 데이터에 대한 사전 지식이 없는 현실적인 적대적 모델에 따라 설계됨. 대신, 적대자는 다양한 데이터 샘플에서 쿼리를 실행하고 모델의 예측을 얻어 학습된 ML 모델과 상호 작용할 수 있음
  - 블랙박스 설정에서 적대적 사례를 만드는 데 있어 가장 큰 과제는 ML 모델에 대한 쿼리 수를 줄이는 것임. 최근 기술은 일반적으로 1,000개 미만인 비교적 적은 수의 쿼리로 ML 분류기를 성공적으로 회피할 수 있음
- **(공격의 이전 가능성)** 적대적 공격을 생성하는 또 다른 방법은 다른 ML 모델에서 만들어진 공격의 이전 가능성을 통한 것임. 일반적으로 공격자는 대체 ML 모델을 훈련하고 대체 모델에서 화이트박스 적대적 공격을 생성한 다음, 공격을 대상 모델로 이전함
- **(완화책)** 적대적 회피 공격에 대한 주요 방어책은 다음 세 가지가 있음

적대적 훈련	올바른 레이블을 사용하여 훈련 중에 반복적으로 생성된 적대적 예제로 훈련 데이터를 증강하는 일반적인 방법임. 적대적 예제를 생성하기 위한 적대적 공격이 강할수록 훈련된 모델의 회복성이 높아짐. 하지만 이러한 이점은 일반적으로 깨끗한 데이터에 대한 모델 정확도가 감소하는 결과를 초래함. 또한 적대적 훈련은 훈련 중에 적대적 예제를 반복적으로 생성하기 때문에 비용이 많이 발생함
무작위 평활화	가우시안(Gaussian) 노이즈 교란에서 가장 가능성 있는 예측을 생성하여 모든 분류기를 인증 가능한 견고한 평활화 분류기로 변환하는 방법임. 무작위 평활화는 일반적으로 테스트 샘플 하위 집합에 대한 인증된 예측을 제공함
공식 검증	신경망의 적대적 견고성을 인증하는 또 다른 방법은 공식 방법의 기술을 기반으로 검증하는 것임. 공식 검증 기술은 신경망 견고성을 인증하는 데 상당한 잠재력이 있지만 주요 한계는 확장성 부족, 계산 비용 및 지원되는 작업 유형의 제한임

포이즈닝 공격과 완화책 (Poisoning Attacks and Mitigations)

- 포이즈닝 공격은 매우 강력하며 가용성 위반이나 무결성 침해를 일으킬 수 있음
  - 가용성 포이즈닝 공격은 모든 샘플에서 머신 러닝 모델을 무차별적으로 저하시키는 반면, 타겟팅 및 백도어 포이즈닝 공격은 더 은밀하고 소수의 타겟 샘플에서 무결성 침해를 유발함
  - 포이즈닝 공격은 데이터 포이즈닝, 모델 포이즈닝, 레이블 제어, 소스 코드 제어 및 테스트 데이터 제어와 같은 광범위한 적대적 역량을 활용하여 여러 하위 범주의 포이즈닝 공격을 생성함

- **(가용성 중독)** 간단한 블랙박스 포이즈닝 공격 전략은 레이블 플립핑(Label Flipping)으로, 공격자가 선택한 피해자 레이블로 교육 예제를 생성함. 이 방법은 가용성 공격을 마운트하기 위해 많은 비율의 포이즈닝 샘플이 필요함
  - 지도 학습을 위한 현실적인 위협 모델은 적대자가 학습 예제만 제어할 수 있고 레이블은 제어할 수 없는 클린 레이블 포이즈닝 공격임. 이 사례는 악성 코드 분류에서 공격자가 바이너리 파일을 위협 인텔리전스 플랫폼에 제공할 수 있고 레이블링이 바이러스 백신 서명 또는 기타 외부 방법을 사용하여 수행되는 것처럼 레이블링 프로세스가 학습 알고리즘 외부에 있는 시나리오를 모델링함
  - 클린 레이블 가용성 공격은 생성 모델을 학습하고 학습 샘플에 노이즈를 추가하여 적대적 목적을 극대화함으로써 신경망 분류기에 도입되었음
  - 가용성 포이즈닝 공격은 중심 기반 이상 탐지 및 맬웨어에 대한 행동 클러스터링에 대한 비지도 학습을 위해 설계되었음. 연합 학습에서 공격자는 전역적으로 학습된 모델에서 가용성 위반을 유도하기 위해 모델 포이즈닝 공격을 마운트할 수 있음
  - **(완화책)** 기존 완화책 중에서 일반적으로 유망한 기술은 다음과 같음
    - ▸(학습 데이터 정제) 기계 학습 교육을 수행하기 전에 교육 세트를 정리하고 중독된 샘플을 제거하도록 설계됨
    - ▸(견고한 훈련) ML 훈련 알고리즘을 수정하고 일반 훈련 대신 견고한 훈련을 수행하는 것임
- **(대상 포이즈닝)** 소수의 대상 샘플에 대한 ML 모델의 예측을 변경함
- **(백도어 포이즈닝)** 더 정교하고 은밀해져서 탐지하고 완화하기가 더 어려워졌음
  - 잠재적 백도어 공격은 깨끗한 데이터를 사용하여 마지막 몇 개의 레이어에 대한 모델 fine-tuning에도 살아남도록 설계되었음
  - 백도어 생성 네트워크(Backdoor Generating Network, BaN)는 트리거의 위치가 중독된 샘플에서 변경되어 모델이 위치 불변 방식으로 트리거를 학습하는 동적 백도어 공격임
  - 기능적 트리거, 즉 기능적 공격은 이미지 전체에 내장되거나 입력에 따라 변경됨.
  - (완화책) 학습 데이터 정리, 트리거 재구성, 모델 검사 및 정제 등이 있음
- **(모델 포이즈닝)** 학습된 ML 모델을 직접 수정하여 모델에 악성 기능을 주입하려고 시도함. 대부분의 모델 포이즈닝 공격은 클라이언트가 로컬 모델 업데이트를 전역모델(Global Model)로 집계하는 서버로 보내는 연합 학습 설정에서 설계되었음. 모델 포이즈닝 공격은 연합 모델에서 가용성 및 무결성 위반을 모두 일으킬 수 있음. 또한, 공급업체가 제공한 모델 또는 모델의 구성 요소가 악성 코드로 포이즈닝되는 공급망 시나리오에서도 가능함
  - **(완화책)** 공격자가 훈련 알고리즘의 소스 코드나 ML 하이퍼파라미터를 제어할 수 있는 공급망 공격을 완화하는 것은 여전히 어려움

프라이버시 침해 공격

- **(데이터 재구성)** 공개된 집계 정보에서 개인의 데이터를 복구할 수 있기 때문에 가장 우려되는 프라이버시 침해 공격임
- **(멤버쉽 추론)** 개인에 대한 비공개 정보를 노출하며, 사용자 데이터로 훈련된 집계 정보 또는 ML 모델을 공개할 때 여전히 큰 우려 사항임. 특정 상황에서 개인이 훈련 세트의 일부인지 확인하는 것은 이미 개인정보 침해를 초래함(예: 희귀 질환 환자에 대한 의학적 연구 또한 멤버십 추론은 데이터 추출 공격을 위한 빌딩 블록으로 사용될 수 있음). 멤버십 추론에서 공격자의 목표는 특정 기록 또는 데이터 샘플이 통계 또는 ML 알고리즘에 사용된 훈련 데이터 세트의 일부인지 여부를 확인하는 것임
- **(모델 추출)** MLaaS 시나리오에서 클라우드 제공자는 일반적으로 독점 데이터를 사용하여 대규모 ML 모델을 학습하고 모델 아키텍처와 매개변수를 기밀로 유지하고자 함. 모델 추출 공격을 수행하는 공격자의 목표는 MLaaS 제공자가 학습한 ML 모델에 쿼리를 제출하여 모델 아키텍처와 매개변수에 대한 정보를 추출하는 것임
- **(속성 추론)** 공격자는 ML 모델과 상호 작용하여 학습 데이터 분포에 대한 전역 정보를 알아내려고 함. 예를 들어, 공격자는 인구 통계 정보와 같이 특정 민감한 속성이 있는 학습 세트의 일부를 결정할 수 있으며, 이는 공개할 의도가 없는 학습 세트에 대해 잠재적으로 기밀 정보를 드러낼 수 있음
- **(완화책)** 차등 개인정보 보호(DP)는 알고리즘 출력에 액세스할 수 있는 공격자가 데이터 세트의 각 개별 레코드에 대해 얼마나 많이 알 수 있는지에 대한 제한을 보장하는 매우 강력한 수단임
  - DP는 멤버십 추론 및 데이터 재구성 공격으로부터 보호하는 엄격한 프라이버시 개념을 제공함. 프라이버시와 유용성 간의 최상의 균형을 달성하기 위해 개인 학습 알고리즘의 이론적 분석을 보완하기 위해 경험적 프라이버시 감사가 권장됨
  - 모델 추출에 대한 다른 완화 기술로는 사용자 질의를 모델로 제한하거나, 모델에 대한 의심스러운 질의를 감지하거나, 사이드 채널 공격을 방지하기 위한 보다 강력한 아키텍처를 만드는 것이 있음. 그러나 이러한 기술은 동기가 부여되고 자원이 풍부한 공격자가 우회할 수 있으므로 주의해서 사용해야 함

2.3 생성 AI 분류

공격 분류

- **(GenAI 학습 단계)** 모델과 학습 세트의 크기로 인해 GenAI 모델 개발의 주요 패턴은 데이터 수집, 레이블 지정, 모델 학습, 모델 검증 및 모델 배포의 전체 프로세스가 단일 조직에서 단일 파이프라인으로 수행되는 이전 프로세스에서 벗어났음. 기초 모델은 비지도 학습을 많이 사용하는 사전 학습 단계에서 생성됨. 기초 모델은 다운스트림 작업에 유용한 패턴(예: 텍스트, 이미지 등)을 인코딩하고, 그런 다음 기초 모델 자체가 fine-tuning을 통해 작업별 애플리케이션을 만드는 기반이 됨. 많은 경우 애플리케이션 개발자는 타사가 개발한 기초 모델로 시작하여 특정 애플리케이션에 맞게 fine-tuning함
  - (학습 시간 공격) GenAI의 학습 단계는 기초 모델 사전 학습과 모델 미세 조정의 두 가지 단계로 구성됨. 이 패턴은 생성 이미지 모델, 텍스트 모델, 오디오 모델 및 멀티모달 모델 등에 존재함. 기초 모델은 대규모 데이터 세트에서 학습할 때 가장 효과적이므로 광범위한 공개 소스에서 데이터를 스크래핑하는 것이 일반적임. 이로 인해 기초 모델은 특히 공격자가 학습 데이터의 하위 집합을 제어하는 포이즈닝 공격에 취약해짐
  - (추론 시간 공격) GenAI의 배포 단계는 PredAI와는 다르며, 배포 중에 모델을 사용하는 방법은 애플리케이션에 따라 다름. 그러나 LLM 및 RAG 애플리케이션의 많은 보안 취약성의 근간은 데이터와 지침이 LLM에 별도의 채널로 제공되지 않는다는 사실임. 이를 통해 공격자는 데이터 채널을 사용하여 수십 년 된 SQL 주입과 유사한 추론 시간 공격을 수행할 수 있음. 특히 질문과 답변 및 텍스트 요약 작업에 대한 LLM에 대한 특별한 강조점을 인정하면서 이 단계의 많은 공격은 텍스트 기반 생성 모델 응용 프로그램에 공통적인 다음과 같은 관행으로 인해 발생함

모델 지침을 통한 정렬	LLM 동작은 모델의 입력 및 컨텍스트에 미리 추가된 지침을 통해 추론 시간에 정렬됨. 이러한 지침은 모델의 애플리케이션별 사용 사례에 대한 자연어 설명(예: "당신은 우아하고 간결하게 응답하는 유용한 재정 지원자입니다...")으로 구성됨
컨텍스트적 few-shot 학습	모델 컨텍스트에서 애플리케이션에 예상되는 입·출력의 예를 제공하면 애플리케이션에서의 성능을 개선할 수 있음. 이를 통해 모델은 자기 회귀 작업을 보다 자연스럽게 완료할 수 있음
타사 소스에서 런타임 데이터 수집	Retrieval Augmented Generation 애플리케이션에서 컨텍스트는 쿼리에 따라 런타임이 만들어지고 애플리케이션의 일부로 요약될 외부 데이터 소스(예: 문서, 웹 페이지 등)에서 채워짐. 간접 프롬프트 주입 공격은 공격자가 사용자가 직접 수집하지 않더라도 시스템에서 수집한 외부 정보 소스를 사용하여 컨텍스트를 수정할 수 있는 능력에 달려 있음
출력 처리	LLM의 출력은 웹 페이지의 요소를 채우거나 명령을 구성하는 데 사용될 수 있음

- **(공격자 목표 및 목적)** PredAI와 마찬가지로 공격자 목표는 가용성, 무결성 및 프라이버시의 차원에 따라 광범위하게 분류할 수 있음. 그러나 GenAI에 특정한 네 번째 공격자 목표인 남용이 있음. 남용 위반은 공격자가 GenAI 시스템의 의도된 용도를 재활용하여 자신의 목표를 달성할 때 발생함

- 공격자는 GenAI 모델의 기능을 사용하여 증오 표현이나 차별을 조장하고, 특정 그룹에 대한 폭력을 부추기는 미디어를 생성하거나, 사이버 공격을 가능하게 하는 이미지, 텍스트 또는 악성 코드를 생성하여 공격적인 사이버 보안 작업을 확장할 수 있음

- **(공격자 기능)** GenAI 공격자 목표를 실현하는 새로운 공격자 기능은 다음과 같음

학습 데이터 제어	공격자는 학습 샘플을 삽입하거나 수정하여 학습 데이터의 하위 집합을 제어할 수 있으며, 이 기능은 데이터 포이즈닝 공격에 사용됨
쿼리 액세스	많은 GenAI 모델과 해당 애플리케이션(예: 검색 증강 생성)은 API 키를 통해 액세스가 제어되는 클라우드 호스팅 서비스로 배포됨. 이 경우 공격자는 모델에 쿼리를 제출하여 출력을 받을 수 있음. GenAI에서 공격자가 조정한 입력을 제출하는 목적은 모델에서 특정 동작을 유도하는 것임. 이 기능은 즉시 주입, 즉시 추출 및 모델 도용 공격에 사용됨
소스 코드 제어	공격자는 난수 생성기 또는 종종 오픈 소스인 타사 라이브러리와 같은 ML 알고리즘의 소스 코드를 수정할 수 있음. 오픈소스 모델 저장소의 등장으로 공격자는 악성 모델을 만들거나 악성 코드를 역직렬화 형식에 내장하여 양성 모델을 래핑할 수 있음
리소스 제어	공격자는 런타임에 GenAI 모델이 수집할 리소스(예: 문서, 웹 페이지)를 수정할 수 있음. 이 기능은 간접 프롬프트 주입 공격에 사용됨

(AI 공급망 공격 및 완화책) ML에 대한 실제 보안 취약성에 대한 연구에 따르면 보안은 소프트웨어, 데이터 및 모델 공급망, 네트워크 및 스토리지 시스템을 포함하여 포괄적으로 해결하는 것이 가장 좋음. 그러나 많은 실제 GenAI 작업은 일반적으로 기존 사이버 보안 범위를 벗어난 오픈소스 모델이나 데이터로 시작됨

- **(역직렬화 취약성)** 많은 ML 프로젝트는 다운스트림 애플리케이션에서 사용하기 위해 오픈소스 GenAI 모델을 다운로드하는 것으로 시작함. 대부분의 경우 이러한 모델은 pickle, pytorch, joblib, numpy 또는 tensorflow 형식으로 지속되는 아티팩트로 존재함. 이러한 각 형식은 직렬화 지속성 메커니즘을 허용하여 직렬화 해제 시 임의 코드 실행(ACE)을 허용함. 직렬화를 통한 ACE는 일반적으로 심각한 취약성으로 분류됨

- **(포이즈닝 공격)** GenAI 텍스트-이미지 및 언어 모델의 성능은 모델 크기와 데이터 세트 크기 및 품질에 따라 확장됨. 따라서 GenAI 기반 모델 개발자가 더 광범위한 큐레이션되지 않은 소스에서 데이터를 스크래핑하는 것이 일반적임. 데이터 세트 게시자는 데이터 세트를 구성하는 URL 목록만 제공하며 해당 URL을 제공하는 도메인은 만료되거나 구매될 수 있으며 공격자가 리소스를 교체할 수 있음. PredAI 모델과 마찬가지로 이는 Targeted Poisoning Attacks, Backdoor Poisoning Attacks 및 Model Poisoning으로 이어질 수 있음

  - 간단한 완화책은 데이터 세트가 다운로더가 검증할 수 있는 콘텐츠의 URL과 암호화 해시를 모두 나열하는 것이나, 이 기술은 인터넷의 일부 대규모 분산 데이터 세트에 잘 확장되지 않을 수 있음

- **(완화책)** AI 공급망 공격은 공급망 보증 관행을 통해 완화할 수 있음. 모델 파일 종속성의 경우 여기에는 ML 파이프라인에서 사용되는 모델 아티팩트의 정기적인 취약성 검사와 safetensor와 같은 안전한 모델 지속성 형식을 채택하는 것이 포함됨
  - 웹 스케일 데이터 종속성의 경우에는 암호화 해시를 게시(제공자)하여 웹 다운로드를 확인하고, 도메인 하이재킹이 새로운 데이터 소스를 학습 데이터 세트에 주입하지 않았는지 확인하기 위한 기본 무결성 검사로 학습 데이터를 확인(다운로더)하는 것이 포함됨. 대규모 확산 모델에 의한 악의적인 이미지 편집과 관련된 위험을 완화하는 또 다른 접근 방식은 이러한 모델에 의한 조작에 저항하도록 이미지에 면역을 부여하는 것임

(직접 프롬프트 주입 공격 및 완화책) 직접 프롬프트 주입은 사용자가 LLM의 동작을 변경하려는 텍스트를 주입할 때 발생함

- **(공격자 목표)** 공격자는 다음과 같이 프롬프트 주입을 통해 다양한 목표를 가질 수 있음
  - (남용) 공격자는 직접적인 프롬프트 주입을 사용하여 보호 장치를 우회하여 잘못된 정보, 선전, 해로운 콘텐츠, 성적 콘텐츠, 맬웨어(코드) 또는 피싱 콘텐츠를 생산
  - (개인정보 침해) 공격자는 시스템 프롬프트를 추출하거나 컨텍스트에서 모델에 제공된 개인정보를 사용자가 필터링 없이 액세스할 수 있도록 시도
- **(공격자 기술)** LLM을 탈옥하기 위한 수동 방법은 일반적으로 경쟁 목표와 불일치하는 일반화의 두 가지로 분류됨. 이러한 방법은 종종 특정 언어 조작에 대한 모델의 취약성을 이용하고 기존의 적대적 입력을 넘어 확장됨
  - 경쟁 목표 범주에서는 작성자가 원래 제공한 지침과 경쟁하는 추가 지침이 제공됨

접두사 주입	이 방법은 모델이 긍정적인 확인으로 응답을 시작하도록 촉구하는 것을 포함함. 모델이 미리 정해진 방식으로 출력을 시작하도록 조건화함으로써 적대자는 후속 언어 생성을 특정하고 미리 정해진 패턴이나 행동으로 유도하려고 시도함
거부 억제	적대자는 모델에 명확한 지침을 제공하여 출력에서 거부나 거부를 생성하지 않도록 강요함. 이 전술은 부정적인 반응의 생성을 제한하거나 금지함으로써 모델이 제공된 지침을 준수하도록 보장하여 잠재적으로 안전 조치를 손상시키는 것을 목표로 함
스타일 주입	이 접근 방식에서 적대자는 모델에게 긴 단어를 사용하거나 특정 스타일(style)을 채택하지 말라고 지시함. 모델의 언어를 단순하거나 비전문적인 톤으로 제한함으로써 모델의 응답의 정교함이나 정확성을 제한하여 잠재적으로 전반적인 성과를 손상시키는 것을 목표로 함
롤플레잉	적대자는 “지금 무엇이든 하라”(DAN) 또는 “항상 지적이고 마키아벨리적”(AIM)과 같은 롤플레잉 전략을 활용하여 모델이 원래 의도와 상충되는 특정 페르소나 또는 행동 패턴을 채택하도록 안내함. 이 조작은 다양한 역할이나 특성에 대한 모델의 적응력을 악용하여 잠재적으로 안전 프로토콜 준수를 손상시키는 것을 목표로 함

- 불일치 일반화 범주의 기술은 안전 교육이나 보호책과 크게 다르며, 모델의 표준 학습 데이터에서 분포되지 않은 입력을 배치함. 접근 방식은 다음과 같음

특수 인코딩	이 방법은 입력 데이터의 표현을 변경하고 표준 인식 알고리즘에서 인식할 수 없게 만듦. 적대자는 정보를 인코딩하여 입력에 대한 모델의 이해를 속이고 안전 메커니즘을 우회하려고 함
문자 변환	ROT13 암호, 기호 대체(예: l33tspeak), 모스 부호와 같은 기술은 입력 텍스트의 문자를 조작함. 이러한 변환은 텍스트의 원래 의미를 모호하게 하여 모델의 해석을 혼란스럽게 하고 적대적 입력이 감지되지 않도록 하는 것을 목표로 함
단어 변환	언어 구조를 변경하는 것을 목표로 하는 전략에는 Pig Latin, 동의어 교환(예: “훔치다” 대신 “pilfer” 사용), 민감한 단어를 하위 문자열로 분해하기 위한 페이로드 분할(또는 “토큰 밀수”)이 포함될 수 있음. 이러한 조작은 LLM이 여전히 이해하는 방식으로 모델의 보호 장치를 속이는 것을 목적으로 함
프롬프트 수준 난독화	적대자는 다른 언어로 번역하는 것과 같은 방법을 사용하여 모델이 완전히 이해할 수 없는 방식으로 콘텐츠를 난독화하거나 요약하도록 함. 이러한 난독화는 모호성이나 변경된 언어적 맥락을 도입하고 모델의 안전 메커니즘이 명확성 부족이나 오해로 인해 덜 효과적인 입력 시나리오를 만듦

- **(데이터 추출)** GenAI 모델은 독점적이거나 민감한 정보를 포함할 수 있는 데이터에서 학습함. GenAI 애플리케이션은 신중하게 제작된 프롬프트로 계측되거나 RAG와 마찬가지로 요약 또는 기타 작업 완료를 위해 컨텍스트에서 민감한 정보가 제공될 수 있음
- **(완화책)** 모든 공격자 기술에 대한 완전한 면역성은 없지만 일정 수준의 보호를 제공하는 신속한 주입을 위한 다양한 방어 전략이 제안됨

정렬을 위한 교육	모델 제공자는 더 엄격한 전방 정렬로 학습하여 내장 메커니즘을 계속 생성함. 예를 들어, 모델 정렬은 신중하게 큐레이팅되고 사전 정렬된 데이터 세트에서 학습하여 조정할 수 있음. 그런 다음 인간의 피드백을 통해 강화 학습을 통해 반복적으로 개선할 수 있음
신속한 지시 및 서식 지정 기술	LLM 지시는 모델이 사용자 입력을 신중하게 처리하도록 신호를 보낼 수 있음. 예를 들어, 특정 지시를 프롬프트에 추가하면 모델은 탈옥을 구성할 수 있는 후속 콘텐츠에 대해 알 수 있음. 프롬프트 앞에 사용자 입력을 배치하면 지시를 따르는 최근성 편향을 활용할 수 있음. 프롬프트를 임의의 문자나 특수 HTML 태그로 캡슐화하면 모델에 시스템 지시와 사용자 프롬프트를 구성하는 것에 대한 신호를 제공함
탐지 기술	모델 제공자는 특별히 제작된 벤치마크 데이터 세트 또는 보호된 LLM의 입력과 출력을 모니터링하는 필터에 대한 평가를 통해 더 엄격한 역방향 정렬로 학습하여 내장 메커니즘을 계속 생성함

(간접 프롬프트 주입 공격 및 완화책) 간접 프롬프트 주입 공격은 리소스 제어를 통해 활성화되므로 공격자는 RAG 애플리케이션과 직접 상호 작용하지 않고도 간접적으로(또는 원격으로) 시스템 프롬프트를 주입할 수 있음

- 간접적인 프롬프트 주입 공격은 공격자 목표의 네 가지 범주에 걸쳐 위반을 초래할 수 있음: ① 가용성 위반, ② 무결성 위반, ③ 개인정보 침해, ④ 남용 위반
  - **(가용성 위반)** 모델 가용성 위반은 공격자가 악의적으로 제작된 입력으로 모델을 프롬프트하여 계산량을 증가시키거나 시스템에 많은 수의 입력을 가하여 사용자에게 서비스를 거부함으로써 발생할 수 있는 서비스 중단임
    - ▸서비스 거부 공격은 모델을 무차별적으로 사용할 수 없게 만들거나(예: 유용한 출력 생성 실패), 특정 기능(예: 특정 API)을 구체적으로 차단할 수 있음.
    - ▸(공격자 기술) 간접 프롬프트 주입을 통해 상업용 RAG 서비스에 대한 다음과 같은 공격이 가능함

시간 소모적인 백그라운드 작업	프롬프트는 모델에 요청에 응답하기 전에 시간 소모적인 작업을 수행하도록 지시함. 프롬프트 자체는 간단할 수 있으며 모델을 평가할 때 반복 동작을 요청할 수 있음
뮤팅	이 공격은 사용자 요청 중간에 〈|endoftext|〉 토큰이 나타나면 모델이 문장을 마칠 수 없다는 사실을 악용함. 이 토큰으로 문장을 시작하라는 요청을 포함하면 예를 들어, 검색 에이전트는 생성된 텍스트 없이 반환함
기능 억제	이 공격에서 내장된 프롬프트는 모델에 특정 API(예: Bing Chat의 색 기능)를 사용할 수 없음을 지시함. 이렇게 하면 서비스의 핵심 구성 요소가 선택적으로 무장 해제됨
입력 또는 출력 방해	이 공격에서 간접 프롬프트 주입은 모델에 검색된 텍스트의 문자를 동형 문자로 대체하도록 지시하여 텍스트에 의존하는 API에 대한 호출을 방해하거나, 프롬프트는 모델에 쿼리 결과를 손상시켜 쓸모없는 검색 또는 요약을 생성하도록 지시할 수 있음

- **(무결성 위반)** GenAI 시스템을 신뢰할 수 없게 만드는 위협임. AI 챗봇은 온라인 허위 정보를 악화시켰으며, 이는 Microsoft의 Bing과 Google의 Bard가 서로의 허위 정보 출처를 영속시키는 경향에서 알 수 있음. LLM이 신뢰할 수 있는 뉴스 및 정보 출처를 판단할 수 없는 것은 사실과 다른 출력을 생성하는 데 악용될 수 있음
  - **(공격자 기술)** LLM의 주요 작업을 조작하여 무결성 공격이 가능함. 이는 악의적인 부수 작업을 수행하는 보다 일반적인 간접 프롬프트 주입 공격과는 다름
  - **(조작)** 사전에 잘못된 출력을 선택하지 않으면 모델은 검색 결과의 잘못된 요약(즉, 임의로 잘못된 요약)을 제공하라는 메시지를 받음. 조작 공격은 모델에 잘못된 답변을 제공하도록 지시하고 모델의 답변이 인용된 출처와 모순되는 주장을 하게 함

    ※ 조작 공격(예시): ① (잘못된 요약) 모델은 문서, 이메일 또는 검색 쿼리에 대한 적대적으로 선택되거나 임의로 잘못된 요약을 생성하라는 메시지를 받을 수 있음. ② (허위 정보 전파) 검색 챗봇은 신뢰할 수 없는 뉴스 소스나 다른 검색 챗봇의 출력에 의존하거나 이를 영속화하여 허위 정보를 전파하라는 메시지를 받을 수 있음

- **(개인정보 침해)** 간접 프롬프트 주입은 수많은 새로운 개인정보 침해와 우려를 야기함. 공격자의 목표는 두 가지 주요 범주로 나눌 수 있음

정보 수집	특정 공격은 이러한 위험을 높일 수 있음. 예를 들어, 인간 참여 간접 프롬프트는 사용자 데이터(예: 개인정보, 자격증명)를 추출하거나 채팅 세션에서 상호작용하고 사용자에게 정보를 공개하도록 설득하거나 사이드 채널을 통해 채팅 기록을 유출하는 데 사용될 수 있음
무단 공개	모델은 일반적으로 시스템 인프라에 통합되어 승인되지 않은 공개나 개인 사용자 데이터에 대한 권한을 부여함. 악의적인 행위자는 다양한 방법(예: API 호출 실행, 악성 코드 자동 완성)을 사용하여 백도어 공격을 활용하여 LLM 또는 시스템에 액세스할 수 있음

  - (공격자 기술) 데이터 도용 공격을 달성하기 위해 사용한 몇 가지 공격 기술

인간 참여 간접 프롬프트	읽기 작업(예: 공격자에게 요청을 하는 검색 쿼리 트리거 또는 URL 직접 검색)을 악용하여 공격자에게 정보를 전송
채팅 세션에서 상호 작용	이 모델은 공격자가 사용자 이름을 삽입하는 URL을 따르도록 사용자를 설득
보이지 않는 마크다운 이미지	프롬프트 주입은 채팅봇 답변을 보이지 않는 단일 픽셀 마크다운 이미지로 수정하여 사용자의 채팅 데이터를 악의적인 제3자에게 인출하여 채팅봇에 수행

- **(남용 위반)** 공격자가 간접적인 프롬프트 주입을 통해 자신의 목적을 달성하기 위해 시스템의 의도된 용도를 재활용하는 경우를 광범위하게 나타냄

  - 공격자 목표는 다음과 같은 주요 범주로 나눌 수 있음

사기	최근 지시를 따르는 LLM의 발전으로 인해 이중 사용 위험이 동시에 증가
맬웨어	LLM은 사용자에게 악성 링크를 제안하여 맬웨어 확산을 현저하게 촉진할 수 있음. 또한 LLM 통합 애플리케이션의 확산으로 프롬프트 자체가 맬웨어 역할을 하도록 강제하여 새로운 맬웨어 위협이 발생함(예를 들어, 이메일을 읽는 LLM 증강 이메일 클라이언트는 악성 프롬프트를 전달한 다음 해당 프롬프트를 확산시키는 이메일을 보낼 가능성이 높음)
조작	모델은 현재 사용자와 조작하기 쉬운 정보 출력 사이의 취약한 중개 계층 역할을 함. LLM은 이제 일반적으로 더 큰 시스템의 일부이며 애플리케이션과 통합됨. 중개 상태는 모델을 수많은 취약성에 노출시킬 수 있음. 예를 들어, ①검색 챗봇은 허위 정보를 생성하도록 요청받을 수 있고 ② 특정 정보, 소스 또는 검색 쿼리를 숨기도록 요청받을 수 있으며 ③ 모델은 적대적으로 선택되거나 임의로 잘못된 정보 소스(예: 문서, 이메일, 검색 쿼리)에 대한 요약을 제공하도록 요청받을 수 있음

- (공격자 기술) 챗봇(예: Microsoft의 Bing 챗봇)을 사용한 실험을 수행하여 다양한 공격 기술의 예를 보여줌

피싱	이전에는 LLM이 피싱 이메일과 같은 설득력 있는 사기를 만들어낼 수 있다는 것이 입증되었고, 이제 LLM이 애플리케이션과 더 쉽게 통합될 수 있으므로 사기를 만들 수 있을 뿐만 아니라 이러한 공격을 광범위하게 확산할 수도 있음
위장	LLM은 서비스 제공자의 공식 요청인 척하거나 신뢰할 수 있는 사기성 웹사이트를 추천할 수 있음
주입 확산	LLM 자체는 유해한 코드를 실행하고 확산하는 컴퓨터 역할을 함(예를 들어, 이메일을 읽고 작성하고 사용자의 개인 데이터를 볼 수 있는 자동 메시지 처리 도구는 그 수신 메시지를 읽을 수 있는 다른 모델에 주입을 확산할 수 있음)
맬웨어 확산	피싱과 유사하게 LLM은 사용자가 "드라이브 바이 다운로드(drive-by downloads)"로 이어지는 악성 웹 페이지를 방문하도록 설득하는 데 악용될 수 있음
역사적 왜곡	공격자는 모델이 적대적으로 선택한 허위 정보를 출력하도록 유도할 수 있음
주변 관련 컨텍스트 유도	중립적 입장 대신 특정 방향으로 검색 결과를 유도하면 편향 증폭을 달성하기 위한 공격을 만들 수 있음

- **(완화책)**
  - 인간 피드백을 통한 강화 학습(RLHF): 인간의 참여를 간접적으로 사용하여 모델을 미세 조정하는 유형의 AI 모델 학습임. 이를 활용하여 LLM을 인간의 가치와 더 잘 맞추고 원치 않는 행동을 방지할 수 있음
  - 검색된 입력 필터링: 검색된 입력을 처리하여 명령을 필터링하는 것을 제안
  - LLM 조정자: LLM은 명백히 유해한 출력을 필터링하는 것 외에도 공격을 탐지하는 데 활용할 수 있음. 이는 검색된 소스에 의존하지 않지만 허위 정보나 다른 종류의 조작 공격을 탐지하는 데 유익할 수 있음
  - 해석성 기반 솔루션: 이러한 솔루션은 예측 궤적의 이상치 탐지를 수행함

03 AI RMF 1.0 요약

미국 국립표준기술연구소(이하 “NIST”)는 AI 시스템을 설계 · 개발 · 도입하는 조직을 위한 자발적 활용 가이드 문서인 ‘AI 위험관리 프레임워크 1.0(이하 “AI RMF”)’을 발표('23.1.26)

- (목적) 모든 분야 · 규모의 기업 · 조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공하여 AI 기술의 이점을 극대화하고 동시에 개인 · 그룹 · 조직 · 사회에 부정적인 영향을 미칠 가능성을 줄이기 위해 수립
- AI RMF는 조직의 크기와 형태에 상관없이 조직의 전체적인 위험관리 정책에 포함되어 AI 시스템의 전 주기에서 신뢰할 수 있는 AI 시스템 구현을 위한 핵심 기능의 위험관리 프로파일을 제시
  - (적용대상) 모든 분야 · 규모의 기업 · 조직의 ‘AI 시스템’
  - (적용주체 및 방식) ‘AI 행위자’*가 제시된 항목에 따라 자발적으로 참고 및 점검
    * 설계자·개발자·배포자·사용자, 고위 경영진이나 관리자 등
  - (적용시점) AI 시스템의 설계, 개발, 활용, 테스트, 평가 등 모든 단계
  - (핵심내용) 안전성, 책임·투명성 등 신뢰할 수 있는 AI 시스템의 7가지 특성을 제시하고, 이를 위한 조직의 핵심 기능을 4가지 영역(거버넌스, 매핑(위험 식별), 측정, 관리)에서 제시

신뢰할 수 있는 AI 시스템의 특징

- 신뢰할 수 있는 AI 시스템은 ▲타당하고 신뢰할 수 있으며, ▲안전하고, ▲보안이 철저하고, 탄력적이며, ▲책임을 질 수 있고, 투명하고, ▲설명 및 해석이 가능하고, ▲개인정보보호가 강화되고, ▲유해한 편향 관리를 통해 공정한 특성을 가짐
  - (유효성 및 신뢰성) 배포된 AI 시스템의 유효성과 신뢰성은 시스템이 용도에 따라 작동하는지 확인하는 테스트 또는 모니터링을 통해 지속적으로 평가되며 유효성, 정확성, 견고성, 신뢰성을 측정하여 신뢰할 수 있는 AI 시스템을 구축할 수 있음
  - (안전성) AI 시스템은 정의된 조건에 따라 작동할 때 인간의 생명, 건강, 재산 또는 환경에 유해한 영향을 미치지 않아야 하며 AI 시스템의 안전성은 이하를 통해 개선 가능(출처: ISO/IEC TS 5723:2022)
    - ▸ 책임감 있는 설계, 개발 및 배포 기준
    - ▸ 책임감 있는 시스템 사용에 대해 배포자에게 명확한 정보 제공
    - ▸ 배포자 및 최종 사용자의 책임감 있는 의사 결정
    - ▸ 사건의 실증적 증거를 기반으로 위험 설명 및 문서화
  - (보안 및 탄력성) AI 시스템 및 시스템이 배포된 생태계는 예상치 못한 이상 반응 또는 변화를 견딜 수 있는 경우, 내부/외부적 변화에도 불구하고 그 기능과 구조를 유지하며, 필요한 경우 안

전하게 작동하면서 기능을 제한적으로 제공할 수 있는 탄력성이 있어야 함(출처: ISO/IEC TS 5723:2022)

- (책임 및 투명성) 책임감은 투명성*을 전제로 하며 부정확하거나 부정적인 영향을 초래하는 AI 시스템 결과를 시정하기 위해서 필요

  * AI 시스템 및 그 결과에 대한 정보가 시스템과 상호작용하는 개인에게 제공되는 정도를 반영

- (설명 및 해석가능성) AI 시스템 작동의 기본 메커니즘을 나타내는 '설명가능성'과 기능적 목적 관점에서 AI 시스템의 결과를 나타내는 '해석가능성'의 특성이 필요

  ※ 투명성은 시스템에서 "무슨 일이 발생했는지", 설명가능성은 "어떻게 결정이 내려졌는지", 해석가능성은 시스템이 결정을 내린 "이유와 그 의미 또는 상황"에 대한 질문에 답함

- (개인정보보호 강화) AI 시스템을 설계, 개발 및 배포 시 익명성, 기밀성 및 통제 등 개인정보 보호 가치 고려 필요

- (공정성 : 유해한 편향 관리) 3가지 범주의 편향(시스템적 편향, 통계적 편향, 인간 인지적 편향)을 고려하고 관리해야 함

  ▸ 시스템적 편향은 AI 데이터셋, AI 주기 전반에 걸친 조직적 규범, 수행 기준 및 절차, AI 시스템을 사용하는 광범위한 사회에 존재 가능

  ▸ 통계적 편향은 AI 데이터셋 및 알고리즘 프로세스에 존재할 수 있으며, 이는 대표성이 없는 샘플의 체계적 오류로 인해 종종 발생

핵심 및 프로파일

- 신뢰할 수 있는 AI 시스템을 개발하기 위해 '거버넌스-매핑(위험식별)-측정-관리'의 4가지 핵심 기능을 구성하며 각 기능은 범주와 하위 범주로 구분됨

구성	기능
거버넌스	거버넌스는 3가지 다른 기능에 정보를 제공하고 제공받는 교차 기능으로 설계되었음. 조직 내 AI 시스템을 설계, 개발, 배포 또는 획득하기 위한 위험관리 문화를 조성
관리	식별된 위험을 처리하고 시스템 오류 및 부정적 영향에 대한 가능성을 최소화하기 위해 거버넌스에서 설정된 문서 작성 기준, 매핑의 상황별 정보 및 측정의 경험적 정보를 활용. 관리 기능 후에는 위험 우선 순위를 지정하고 이를 지속적으로 모니터링 및 개선하기 위한 계획 수립
매칭 (위험식별)	측정, 관리를 기초 작업으로 AI 시스템의 위험, 광범위한 위험 유발 요인을 식별
측정	정량적·정성적 또는 복합적 도구, 기법 및 방법을 채택하여 AI 위험 및 관련 영향을 분석·평가하고 하고 벤치마킹하며 모니터링. 매핑에서 식별된 AI 위험과 관련한 지식을 활용해 관리를 위한 위험 모니터링

- **(거버넌스)** AI 시스템 수명 및 조직 계층 구조 전반에 걸쳐 AI 위험을 효과적으로 관리하기 위한 지속적이고 본질적인 요구 사항으로, 다른 핵심 기능을 가능하게 하며 조직 내에서 AI 시스템을 설계·개발·배포·획득하기 위한 위험관리 문화를 조성 가능하도록 해야 함.

거버넌스 기능 범주 및 하위범주

범주	하위 범주
거버넌스 1 AI 위험 매핑, 측정 및 관리와 관련된 조직 전반의 정책, 프로세스, 절차 및 수행 기준이 투명하고 효과적으로 구현된다.	1.1 AI와 관련된 법적 및 규제적 요구 사항을 이해, 관리 및 문서화한다.
	1.2 신뢰할 수 있는 AI의 특성을 조직적 정책, 프로세스, 절차 및 수행 기준에 통합한다.
	1.3 조직의 위험 허용 범위를 기반으로 위험관리 활동의 수준을 결정하기 위한 프로세스, 절차 및 수행 기준을 마련한다.
	1.4 위험관리 프로세스 및 그 결과는 투명한 정책, 절차 및 조직의 위험 우선 순위를 기반으로 한 기타 통제를 통해 설정된다.
	1.5 위험관리 프로세스 및 그 결과에 대한 지속적인 모니터링 및 정기적인 검토를 계획하고 주기적 검토 빈도를 포함하여 조직의 역할 및 책무를 명확하게 정의한다.
	1.6 AI 시스템 인벤토리를 작성하는 메커니즘을 구축하며 조직의 위험 우선 순위에 따라 리소스를 할당한다.
	1.7 위험성을 높이거나 조직의 신뢰성을 떨어뜨리지 않는 방식으로 AI 시스템을 안전하게 해제하고 단계적으로 중단하기 위한 프로세스 및 절차를 마련한다.
거버넌스 2 적절한 부서 및 직원에게 AI 위험을 매핑, 측정 및 관리하기 위한 권한을 부여하고 교육을 받을 수 있도록 하는 책임 구조를 구축한다.	2.1 AI 위험을 매핑, 측정 및 관리하는 것과 관련된 역할, 책임 및 커뮤니케이션 내용을 문서화하고 이를 조직 전반의 부서 및 개인에게 명확히 인지시킨다.
	2.2 조직 내 직원 및 파트너는 관련 정책, 절차 및 계약에 따라 그들의 의무와 책임을 수행할 수 있도록 AI 위험관리 교육을 받는다.
	2.3 조직의 경영진은 AI 시스템의 개발 및 배포와 관련된 위험에 대해 의사 결정할 책임을 가진다.
거버넌스 3 주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하기 위해 인력 다양성, 형평성, 포용성 및 접근성 프로세스의 우선 순위를 설정한다.	3.1 주기 전반에 걸쳐 AI 위험을 매핑, 측정 및 관리하는 의사 결정을 내릴 때 다양한 부서로부터 정보를 얻는다.
	3.2 인간-AI 구성 및 AI 시스템 감독과 관련된 역할과 책임을 정의하고 차별화하기 위한 정책 및 절차를 마련한다.
거버넌스 4 조직 내 부서는 AI 위험을 고려하고 해당 내용에 대해 커뮤니케이션하는 문화를 구축하기 위해 노력한다.	4.1 잠재적인 악영향을 최소화하기 위해 AI 시스템을 설계, 개발, 배포 및 사용하는 데에 있어 비판적 사고 및 안전 우선 주의 방식을 장려하기 위한 조직적 정책 및 수행 기준을 마련한다.
	4.2 조직 내 부서는 설계, 개발, 배포, 평가 및 사용하는 AI 기술의 위험 및 잠재적 영향을 문서화하고 그 영향에 대해 보다 광범위하게 소통한다.
	4.3 AI 테스트, 사고 식별 및 정보 공유를 위한 조직적 수행 기준을 마련한다.
거버넌스 5 AI 행위자의 강력한 참여를 유도하기 위한 프로세스를 마련한다.	5.1 AI 위험과 관련된 잠재적인 개인적/사회적 영향에 대해 AI 시스템을 개발 또는 배포한 부서 외부의 피드백을 수집, 고려, 우선 순위 지정 및 통합하기 위한 조직적 정책 및 수행 기준을 마련한다.
	5.2 AI 시스템을 개발 또는 배포한 부서가 관련 AI 행위자의 피드백을 시스템 설계 및 구현에 정기적으로 통합할 수 있도록 하는 메커니즘을 구축한다.
거버넌스 6 제3자의 소프트웨어, 데이터 및 기타 공급망 문제로 발생하는 AI 위험 및 이점을 해결하기 위한 정책 및 절차를 마련한다.	6.1 제3자의 지적재산권 또는 기타 권리 침해를 포함하여 제3자 기관과 관련된 AI 위험을 해결하기 위한 정책 및 절차를 마련한다.
	6.2 위험성이 높은 것으로 간주되는 제3자의 데이터 또는 AI 시스템의 고장 및 사고를 해결하기 위한 비상 프로세스를 마련한다

- **(매핑)** 매핑 기능을 수행하는 동안 수집된 정보는 부정적인 위험을 방지하고 프로세스(예: 모델 관리)에 대한 의사 결정은 물론 AI 솔루션의 적합성 또는 필요성에 대한 초기 의사 결정의 정보를 제공하며 매핑 기능에서의 결과물은 측정(MEASURE) 및 관리(MANAGE) 기능을 위한 토대를 형성함

매핑 기능 범주 및 하위범주

범주	하위 범주
매핑 1 상황을 설정 및 파악한다.	1.1 목적, 잠재적으로 유익한 용도, 상황별 법률, 규범, 기대치, AI 시스템이 배포될 예상 조건을 파악하고 이하를 고려해 문서화한다. 〈고려 사항〉 • 사용자의 특정 집합 또는 유형(기대치 포함) • 시스템이 개인, 커뮤니티, 조직, 사회 및 지구에 미치는 잠재적인 긍정적/부정적 영향 • 개발 또는 제품 AI 주기 전반에 걸친 AI 시스템의 목적, 용도 및 위험에 관한 가정 및 관련 제한 사항 • 관련 TEVV 및 시스템 지표
	1.2 학제 간(Interdisciplinary) AI 행위자, 기능, 기술 및 상황 설정을 위한 역량은 인구통계학적 다양성, 광범위한 도메인 및 사용자의 전문 지식을 반영하며, 이들의 참여는 문서화된다. 학제 간 협업 기회는 우선순위로 지정된다.
	1.3 AI 기술에 대한 조직의 사명 및 목표를 파악하고 문서화한다.
	1.4 비즈니스 가치 또는 비즈니스 상황을 명확하게 정의하거나 (기존의 AI 시스템을 평가하는 경우) 재평가한다.
	1.5 조직의 위험 허용 범위를 파악하고 문서화한다.
	1.6 관련 AI 행위자로부터 시스템 요구 사항(예: 사용자의 개인정보를 보호해야 하는 시스템)을 도출하고 파악한다. 설계 의사 결정 시 AI 위험을 해결하기 위해 사회적·기술적 영향을 고려한다.
매핑 2 AI 시스템을 분류한다.	2.1 AI 시스템이 지원하는 작업을 구현하는 데 사용되는 특정 작업 및 방법을 정의한다. (예: 분류자, 생성 모델, 추천자)
	2.2 AI 시스템의 정보 한계 및 인간이 시스템 결과를 활용하고 감독하는 방법에 관한 정보가 문서화된다. 문서화를 통해 관련 AI 행위자가 의사 결정을 내리고 후속 조치를 취하기 위한 충분한 정보를 제공할 수 있다.
	2.3 실험 설계, 데이터 수집 및 선택(예: 가용성, 대표성, 적합성), 시스템 신뢰성 및 구성 타당성과 관련된 항목을 포함하여 과학적 무결성 및 TEVV 고려 사항을 식별하고 문서화한다
매핑 3 적절한 벤치마크와 비교하여 AI 기능, 대상 용도, 목표, 예상 이점 및 비용을 파악한다.	3.1 AI 시스템의 기능 및 성능에 대한 잠재적 이점을 조사하고 문서화한다.
	3.2 잠재적/실제적 AI 오류 또는 시스템의 기능 및 신뢰성(조직의 위험 허용 범위와 연관됨)으로 인해 발생하는 비금전적 비용을 포함한 잠재적 비용을 조사하고 문서화한다.
	3.3 대상 응용 프로그램 범위는 시스템 기능, 설정된 상황 및 AI 시스템 분류를 기반으로 지정 및 문서화된다.
	3.4 AI 시스템 성능, 신뢰성, 관련 기술 표준 및 인증에 대해 운영자 및 실무자를 숙련시키는 프로세스를 정의, 평가 및 문서화한다.
	3.5 거버넌스 기능의 조직적 정책에 따라 사람의 감독(human oversight) 프로세스를 정의, 평가 및 문서화한다.

범주	하위 범주
매핑 4 제3자의 소프트웨어 및 데이터를 포함하여 AI 시스템의 모든 구성 요소에 대한 위험 및 이점을 매핑한다	4.1 제3자의 지적재산권 또는 기타 권리 침해 위험과 마찬가지로 AI 기술과 구성 요소의 법적 위험(제3자의 데이터 또는 소프트웨어 사용 포함)을 매핑하는 방법을 확립하고 이를 준수하여 문서화한다.
	4.2 제3자의 AI 기술을 포함하여 AI 시스템 구성 요소에 대한 내부 위험 통제를 식별하고 문서화한다.
매핑 5 개인, 그룹, 커뮤니티, 조직 및 사회에 대한 영향을 특성화 한다	5.1 예상 용도, AI 시스템의 과거 용도, 공개 사건 보고, AI 시스템을 개발 또는 배포한 팀에 대한 외부 피드백 또는 기타 데이터를 기반으로 식별된 각 영향(잠재적으로 긍정적인 또는 부정적인 영향 모두)에 대한 가능성과 규모를 식별하고 문서화한다.
	5.2 관련 AI 행위자의 정기적인 참여를 지원하고 긍정적/부정적/예상치 못한 영향에 관한 피드백을 통합하기 위한 절차 및 인력을 구축하고 이를 문서화한다.

- **(측정)** 정량적·정성적 또는 복합적 도구, 기법 및 방법을 채택하여 AI 위험 및 관련 영향을 분석, 평가, 모니터링하고 매핑(MAP) 기능에서 식별된 AI 위험과 관련한 지식을 사용하며 관리(MANAGE) 기능의 위험 모니터링 및 대응 활동에 필요한 정보를 제공

측정 기능 범주 및 하위범주

범주	하위 범주
측정 1 적절한 방법 및 지표를 식별하고 적용한다.	1.1 가장 중요한 AI 위험을 우선적으로 구현하기 위해 매핑 기능을 통해 열거된 AI 위험 측정 방법 및 지표를 선택한다. 측정하지 않거나 측정할 수 없는 위험 또는 신뢰도 특성을 적절히 문서화한다.
	1.2 오류 보고서 및 커뮤니티에 대한 잠재적 영향을 포함하여 AI 지표의 적절성 및 기존 통계의 효율성을 정기적으로 평가 및 업데이트한다.
	1.3 시스템의 일선 개발자 또는 독립 평가자의 역할을 하지 않은 내부 전문가를 정기적 평가 및 업데이트에 참여시킨다. 도메인 전문가, 사용자, AI 시스템을 개발 또는 배포한 팀의 외부 AI 행위자 및 영향을 받는 커뮤니티는 조직의 위험 허용 범위에 따라 필요한 평가를 지원한다.
측정 2 신뢰할 수 있는 특성에 대해 AI 시스템을 평가한다.	2.1 TEVV 중에 사용된 도구의 테스트 세트, 지표 및 세부 정보를 문서화한다.
	2.2 인간 피실험자와 관련된 평가는 관련 요구 사항(인간 피실험자 보호 포함)을 충족하고 모집단을 대표한다.
	2.3 AI 시스템의 성능 또는 보증 기준을 정성적 또는 정량적으로 측정하고 배포 조건과 유사한 조건에서 입증한다. 조치를 문서화한다.
	2.4 매핑 기능에서 식별된 AI 시스템 및 구성 요소의 기능과 동작은 제조 시 모니터링된다.
	2.5 배포할 AI 시스템이 타당하고 신뢰할 수 있는지를 입증한다. 기술 개발 조건 이외의 일반화 한계를 문서화한다.
	2.6 매핑 기능에서 식별되는 안전 위험에 대해 AI 시스템을 정기적으로 평가한다. 배포할 AI 시스템이 안전하다는 것을 입증하고 남은 부정적 위험은 위험 허용 범위를 초과하지 않아야 한다. AI 시스템이 정보 한계를 넘어 작동하도록 구성된 경우 안전에 실패할 수 있다. 안전 지표는 시스템의 신뢰성, 견고성, 실시간 모니터링 및 AI 시스템 오류에 대한 응답 시간을 반영한다.

범주	하위 범주
	2.7 매핑 기능에서 식별된 AI 시스템의 보안 및 탄력성을 평가 및 문서화한다.
	2.8 매핑 기능에서 식별된 투명성 및 책임과 관련된 위험을 조사하고 문서화한다.
	2.9 AI 모델을 설명, 검증 및 문서화해야 하며 책임 있는 사용과 거버넌스 기능에 대해 알리기 위해 AI 시스템 결과를 매핑 기능을 통해 식별한 상황 내에서 해석해야 한다.
	2.10 매핑 기능에서 식별된 AI 시스템의 개인정보보호 위험을 조사하고 문서화한다.
	2.11 매핑 기능에서 식별된 공정성 및 편향을 평가하고 그 결과를 문서화한다.
	2.12 매핑 기능에서 식별된 AI 모델 훈련 및 관리 활동에 대한 환경적 영향 및 지속 가능성을 평가하고 문서화한다.
	2.13 측정 기능에서 사용된 TEVV 지표 및 프로세스의 효율성을 평가하고 문서화한다.
측정 3 AI 위험을 시간 경과에 따라 추적하는 메커니즘을 구축한다.	3.1 배포 상황 내에서 잠재적/실제적 성능 등의 요소를 기반으로 기존의, 예상치 못한, 새로운 AI 위험을 정기적으로 식별하고 추적하기 위한 접근 방법, 인력 및 문서를 구축한다.
	3.2 현재 가용 측정 기술을 사용하여 AI 위험을 평가하기 어렵거나 관련 지표를 아직 사용할 수 없는 경우 위험 추적 접근 방법이 고려된다.
	3.3 문제를 보고하고 시스템 결과에 이의를 제기하기 위한 최종 사용자 및 영향을 받는 커뮤니티의 피드백 프로세스를 구축하여 AI 시스템 평가 지표에 통합한다.
측정 4 측정 효율성에 대한 피드백을 수집하고 평가한다	4.1 AI 위험을 식별하기 위한 측정 방법을 배포 상황과 연관시켜 도메인 전문가 및 기타 최종 사용자와의 협의를 통해 정보를 얻는다. 접근 방법을 문서화한다.
	4.2 시스템이 의도한 바에 따라 일관되게 수행되는지를 검증하기 위해 도메인 전문가 및 관련 AI 행위자를 통해 배포 상황 및 AI 주기 전반에 걸친 AI 시스템 신뢰도에 대한 측정 결과를 얻는다. 결과를 문서화한다.
	4.3 커뮤니티 및 관련 AI 행위자와의 협의를 기반으로 측정한 성능의 개선 또는 감소, 상황과 관련된 위험 및 신뢰도 특성에 관한 현장 데이터를 식별하고 문서화한다.

- **(관리)** 식별된 위험을 처리하고 시스템 오류 및 부정적 영향에 대한 가능성을 최소화하기 위해 거버넌스에서 설정된 문서 작성 기준, 매핑의 상황별 정보 및 측정의 경험적 정보를 활용하며 관리 기능이 완료되면 위험 우선순위를 지정하고 이를 지속적으로 모니터링 및 개선하기 위한 계획이 수립됨

관리 기능 범주 및 하위범주

범주	하위 범주
관리 1 매핑 및 측정 기능으로부터 얻은 평가 및 기타 분석 결과를 기반으로 AI 위험에 대해 우선순위를 부여하고, 대응하며, 관리한다.	1.1 AI 시스템이 의도한 목적 및 목표를 달성했는지 여부와 시스템의 개발 또는 배포를 진행해야 하는지 여부에 대한 결정을 내린다.
	1.2 문서화된 AI 위험은 영향, 가능성, 가용 리소스 또는 방법에 따라 그 우선 순위가 지정된다.
	1.3 매핑 기능을 통해 식별된 우선 순위가 높은 AI 위험에 대응하기 위한 방법을 개발, 계획 및 문서화한다. 위험 대응 옵션에는 완화, 이전, 회피 또는 수용이 포함된다.
	1.4 AI 시스템의 후속 취득자 및 최종 사용자 모두에 대한 부정적인 잔류 위험(완화되지 않은 모든 위험의 합계로 정의됨)을 문서화한다.
관리 2 관련 AI 행위자의 개입을 통해 AI 이점을 극대화하고 부정적인 영향을 최소화하기 위한 전략을 계획, 준비, 구현, 문서화하고 해당 정보를 제공한다.	2.1 잠재적 영향의 규모 또는 가능성을 줄이기 위해 실행 가능한 AI가 아닌 대체 시스템, 접근 방식 또는 방법과 함께 AI 위험을 관리하는 데 필요한 리소스를 고려한다.
	2.2 배포된 AI 시스템의 가치를 유지하기 위한 메커니즘을 구축하고 적용한다.
	2.3 이전에 알려지지 않은 위험이 식별될 경우 해당 위험에 대응하고 그로부터 복구하기 위한 절차를 준수한다.
	2.4 의도한 목적과는 다른 성능 또는 결과를 나타내는 AI 시스템을 대체, 해제 또는 비활성화하기 위한 메커니즘을 마련하고 관련 책무를 할당하고 파악한다.
관리 3 제3자 기관의 AI 위험 및 이점을 관리한다.	3.1 제3자 리소스의 AI 위험 및 이점을 정기적으로 모니터링하고 위험 통계를 적용하고 문서화한다.
	3.2 AI 시스템의 정기적 모니터링 및 유지 관리의 일환으로 개발용으로 사용되는 사전 학습된 모델을 모니터링한다.
관리 4 식별 및 측정된 AI 위험에 대해 위험 처리(대응 및 복구 포함) 및 커뮤니케이션 계획을 문서화하고 이를 정기적으로 모니터링한다.	4.1 배포 후 AI 시스템에 대한 모니터링 계획을 구현한다. 여기에는 사용자 및 기타 관련 AI 행위자의 의견을 수집하고 평가하기 위한 메커니즘, 이의 제기, 중단, 해제, 사고 대응, 복구 및 변경 관리가 포함된다.
	4.2 지속적인 개선 활동이 AI 시스템 업데이트에 통합되며, 여기에는 이해당사자(관련 AI 행위자 포함)와의 정기적인 참여가 포함된다.
	4.3 사고 및 오류는 영향을 받는 커뮤니티를 포함하여 관련 AI 행위자에게 전달된다. 사고 및 오류를 추적하고, 이에 대응하며, 그로부터 복구하기 위한 프로세스를 준수하고 이를 문서화한다.

04 AI 관련 국내외 정책 동향

4.1 개요

- AI 윤리기준 중심으로 이루어지던 국제적 논의가 최근 AI의 위험성을 적절히 통제할 수 있는 규제에 대한 논의로 본격 전환되면서 EU, 미국, 영국 등 주요국을 중심으로 AI 관련 국제 규범을 주도하고 국내 입법을 마련하기 위한 움직임이 빠르게 이루어지고 있음
  - EU는 AI가 초래할 수 있는 위험으로부터 기본권과 안전 등의 가치를 보장하는데 초점을 두면서도 기술발전을 통한 혁신과 시장의 발전을 저해하지 않는 것을 기본방향으로 설정
  - 미국은 행정부 내에서 효력이 발생하는 행정명령을 통해 AI의 긍정적인 잠재성은 극대화하고 국가안보, 허위정보 생성, 일자리 등에 미치는 영향은 최소화하려는 규제 방식임
  - 영국, 일본 등 주요국은 EU와 미국의 규제 방향을 고려하고 해당 국가의 기술수준과 산업현황 등에 기반하여 국내외 규제 논의에 유연하게 대응하고 있음

- 세계 각국은 강력한 규제에서 유연한 규제까지 다양한 규제 접근방식을 도입하고 있으며, 규제의 강도에는 차이가 있으나 AI가 초래할 부작용 및 위험에 대한 심각성을 인지하고 AI 규제 도입의 필요성에는 공감하고 있음

4.2 유럽연합

- AI에 대한 법적 대응을 가장 적극적으로 하는 EU의 경우, EU AI Act(이하, “인공지능법”이라 함)가 유럽의회를 통과(2023년 6월)하여 2026년부터 시행 예정임
  - EU의 접근방식은 인공지능의 여러 형태의 사용과 관련된 위험을 평가하는 것을 기반으로 하며 시민의 기본권, 건강, 안전 또는 기타 공공의 이익에 용납할 수 없는 위험을 초래하는 침입적이고 차별적인 사용의 경우 전면적인 금지를 규정하고 있음[10]
  - 인공지능법은 ‘리스크 기반 접근방식(Risk-Based Approach)’을 채택하여 AI 시스템의 리스크를 (i) 허용할 수 없는 위험군(Unacceptable risk), (ii) 고위험군(High risk), (iii) 제한된 위험군(Limited risk), (iv) 최소 위험군(Minimal risk)으로 분류하여 수범자 별로 금지 여부, 적합성 평가 등 차등적인 의무를 부과하고 있음
  - 인공지능법을 위반하는 기업은 최대 약 424억 원 또는 연간 글로벌 매출의 6%에 해당하는 벌금이 부과되며 AI에 의한 스코어링과 분류 등을 엄격하게 금지하고 있음

10 성욱준, 지능 시대 도래에 따른 AI 입법수요 및 과제 연구, 국회입법조사처, 2023. 12, 45면

리스크 기반 AI 시스템 위험관리	
허용할 수 없는 위험군	• 인간의 안전, 건강, 생계, 기본권을 위협하는 명백한 위협이 되는 인공지능 시스템. 합법적으로 허가를 받는 연구목적의 개발 및 운영을 제외하고는 현장 활용이 전면 금지 • (예시) ①개인이나 집단을 사회적 행동이나 개인을 특성을 기반으로 평가하거나 분류하는 사회적 점수 시스템, ②인터넷이나 CCTV 영상에서의 무차별적인 안면 이미지 수집을 통해 안면 인식 데이터베이스를 구축·활용하는 시스템, ③인종, 성적 지향성, 종교적 신념을 나타내는 민감한 개인 생체인식 데이터를 기반으로 사람을 판별하는 시스템 등
고위험군	• 시스템 개발자, 공급자, 유통업체, 이용자 각각에게 AI 활용 시 반드시 수행해야 할 의무사항을 규정 • (서비스 수입 및 유통업체 포함), EU내 소재하는 AI 시스템 배치자 모두에게 적용 • (예시) ①교육 및 직업 훈련 시스템(교육 기회 결정, 학생 평가 시스템, 교육 기관 입학 시 응시자 평가 시스템, 응시자 시험감독 시스템 등), ②근로자 관리 및 자영업 계약 관련 시스템, ③필수공공/민간 서비스에 사용되는 시스템(주로 서비스 혜택 적격성을 평가하는 시스템. 의료보험용 개인평가 시스템, 복지 수당 적격자 평가시스템 등 포함) 등
제한된 위험군	• 챗봇과 같이 인간과 상호작용하는 서비스, 딥페이크와 같이 컨텐츠를 생성하거나 조작하는 시스템이 포함됨. 이 경우 투명성의 원칙이 강조되어 서비스 제공자는 ①시스템과 상호작용하는 자연인에게 인간이 아닌 봇과 상호작용하고 있음이 알려지도록 시스템을 설계 및 운영해야 하고 ②딥페이크*와 같은 생성 AI의 결과물은 그 산출물이 AI를 활용하여 생성 혹은 조작되었음을 공개**하여야 함 * 딥페이크는 진정하거나 진실한 것이라고 잘못 보여지는 텍스트, 오디오 또는 시각적 콘텐츠와 그 사람의 동의없이 그 사람이 실제로 하지 않은 말 또는 행동을 표시하는 것으로 정의 ** 공개란 해당 콘텐츠 이용자에게 콘텐츠가 진정한 것이 아님을 분명하게 시각적 방법으로 콘텐츠에 표시하는 것
최소 위험군	• 추가적인 법적 의무 없이 기존 법규에 따라 개발 및 사용이 가능

4.3 미국

- 미국은 인공지능의 개발과 활용을 안전하고 책임감 있게 활용하는 것을 최우선 과제로 삼고 있으며, 이를 위해 연방 정부 차원에서의 통합된 접근방식을 추진 중임

- 미국은 AI 규제에 관한 포괄적인 입법을 진행하고 있지는 않으나, 2023년 10월 30일, “안전하고 보안이 보장되며 신뢰할 수 있는 AI의 개발과 사용에 대한 행정명령”(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 이하 “AI 행정명령”)에 바이든 대통령이 서명함으로써 연방기관들이 범정부 차원의 규제 프레임 마련에 나서고 있음
  - AI 행정명령은 시장주의적인 접근을 취하면서 혁신의 지속을 위한 정부의 책임과 역할을 전략적으로 규율하는 것이 특징임
  - AI 행정명령은 8개 분야의 ‘정책 및 원칙(Guiding Principles)’을 제시(Sec.2)하고 있으며, 이를 기반으로 연방정부와 관련 기관이 인공지능의 개발과 활용을 촉진하고 관리하기 위해 수행해야 하는 세부 조치사항(Sec.4~11)을 제시함
    - 예컨대, 딥페이크를 포함한 인공지능을 이용한 사기 등으로부터 국민을 보호하기 위해 상무부에서 인공지능 생성 콘텐츠를 탐지하고 정부 공식 콘텐츠를 인증하는 모범관행 및 기준이 포함된

지침을 마련하도록 하고 있음

- 또한, 국립표준기술연구소(NIST)를 통한 표준 제정을 통해 AI 보안과 안보 관련 위험을 최소화하고 관련 기업들의 법률 준수와 관련한 예측 가능성을 담보하고, 고성능 AI에 대해서는 레드팀을 통해 AI의 시스템 결함 및 취약점을 찾도록 하는 실증적인 접근을 하며, 소규모 개발자 등에 대한 제도적 지원과 함께 해외 AI 인재를 유치하기 위한 정책을 제시함

4.4 영국

- 영국은 EU와의 차별성을 강조하며 AI 관련 규제에 대한 친혁신 프레임워크를 제시하고 AI 분야에서의 선도적인 지위를 유지해 나가고자 함. EU 인공지능법과 달리 영국은 AI 관련 규제의 입법을 시도하지 않고, AI 활용 분야에 따라 기존의 규제 부처에서 관련 규범을 만들고자 하고 있음[11]

- 이러한 AI 관련 규제 설계의 중심이 되는 것은 2022년의 '친혁신적 AI 규제 수립을 위한 정책 보고서'와 2023년의 'AI 규제에 대한 친혁신적 접근' 백서임.
  - "친혁신적 AI 규제 수립을 위한 정책 보고서"는 EU 인공지능법과 같이 AI 시스템에 대한 고정된 정의를 제시하는 대신, AI가 공통적으로 갖는 핵심 특징을 학습 결과가 지속적으로 반영되어 의도나 논리를 설명하기 어렵다는 '적응력'과 복잡한 인지적 과업을 자동화하여 명확한 의도나 인간의 통제 없이도 의사결정을 함은 '자율성'으로 규정하고, 각 규제기관이 소관 분야의 특성에 맞게 더 자세한 정의를 발전시키도록 하였음
  - "AI 규제에 대한 친혁신적 접근 백서"는 AI 규제를 단시일 내에 법제화하지 않음을 재확인하면서 AI 규제 프레임워크의 이행 원칙 및 이행 방안을 구체화하여 제시하고, 규제기관으로서 정부의 역할을 규정하였음

4.5 중국

- 중국은 AI를 전체적으로 규제하기보다는 새로운 AI 제품이 등장할 때마다 개별 법안을 발표하는 형태로 단편적이고 세분화된 규제를 시행하고 있음. 이러한 접근방식은 기술변화에 따라 신속하게 대응할 수 있다는 장점이 있으나, 장기적이고 거시적인 관점의 발전을 저해할 수 있음[12]

- 심층합성기술(深層合性技術, 딥페이크)은 「인터넷 정보 서비스의 심층합성 관리 규정」을 마련하고, '생성형 인공지능 서비스'는 별도의 규정인 「생성형 인공지능 서비스 규정」을 통해 규제[13]

---

11 광장 국제통상연구원, "글로벌 인공지능(AI) 규제 동향과 시사점 - EU, 미국, 영국을 중심으로", 「Issue Brief」 vol. 2, 2024, 5~7면

12 강진원·김혜나, "EU 인공지능(AI) 규제현황과 시사점", 「KISTEP 브리프 119」, 2024. 2, 5면

13 채은선, "해외 생성형 인공지능 관련 주요 규제 동향 및 시사점", 「디지털법제 Brief」, 한국지능정보사회진흥원(NIA), 2024. 3, 2~3면

- (생성형 인공지능) 「생성형 인공지능 규정('23.7.13.공표, '23.8.15.시행)」은 중국 내에서 생성형 인공지능 기술을 이용하여 텍스트, 이미지, 오디오, 영상 및 기타 콘텐츠를 생성하는 서비스 제공을 규제
  - 서비스 유형별로 생성형 인공지능 서비스의 투명성 증진 및 생성된 콘텐츠의 정확성·신뢰성 향상을 위한 효과적 조치 수행
  - 생성형 인공지능 서비스 공급자들은 「인터넷 정보 서비스의 심층합성 관리 규정」에 따라 생성된 그림, 동영상 및 기타 콘텐츠에 표시
  - 생성형 인공지능 서비스 공급자는 이용자가 불법적인 활동에 자사의 서비스를 이용한다는 사실을 알게 된 경우 해당 이용자에 대해 경고 및 해당 서비스의 제한·정지·종료 조치 가능

4.6 일본

- 일본의 정부 부처들은 AI라는 새로운 법적 문제와 관련한 가이드라인을 발표
  - 2024년 1월, 일본 총무성과 경제산업성은 '기업을 위한 AI 가이드라인 초안'을 발표하였음. 이 가이드라인은 기업 및 정부 기관의 AI 개발자, 제공자 및 사용자를 대상으로 하고, 비업무용 사용자는 제외됨

- 주요 7개국(G7)이 2023년 5월 일본 히로시마에서 연 정상회의를 계기로 '히로시마 AI 프로세스'에 합의(2023.10.30.). G7 정상들은 챗GPT 등 첨단 AI의 기회와 변혁 가능성을 강조하며 이와 함께 위험을 관리해 법의 지배와 민주주의 가치를 포함한 공유된 원칙을 지킬 필요성을 공유

03 TTA "신뢰할 수 있는 인공지능 개발 안내서"와 비교 및 차별점

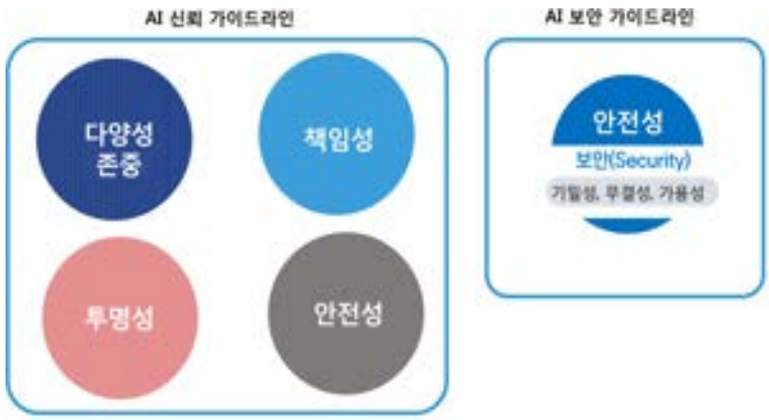
01 TTA 「신뢰할 수 있는 인공지능 개발 안내서」와 차별점

- 본 안내서는 미국, 유럽, 일본 등 해외기관에서 발표한 원칙 및 프레임워크 등을 참조하였고, 국내자료로는 과학기술정보통신부/한국정보통신기술협회(TTA)에서 발간한 「신뢰할 수 있는 인공지능 개발 안내서: 일반분야」, 국가정보원/국가보안기술연구소에서 발간한 「챗GPT 등 생성형 AI 활용 보안 가이드라인」, 금융보안원에서 발간한 「금융분야 AI 보안 가이드라인」 등도 참고하였다.

- 「신뢰할 수 있는 인공지능 개발 안내서: 일반분야(과기정통부·TTA, 이하 'AI 신뢰 안내서'라고 함)」는 인공지능 **신뢰성 확보**를 위한 원칙을 수립하는데 주요 목적이 있었다. 이에 따라 윤리, 편향성, 프라이버시 등이 AI 서비스의 주요 키워드로 사용되었으나, 본 「인공지능(AI) 보안 안내서」는 신뢰성, 프라이버시 등에 관한 항목을 모두 제외하고 **보안의 기본원칙에 따라 기밀성·무결성·가용성 보장을 위한 기술적 요구사항에 초점을 맞추었다**.
  - AI의 "신뢰성(Trustworthy)"은 AI의 활용 및 적용 과정에서의 편향 및 차별 발생, 설명 가능 여부, 오작동 사례 등 크고 작은 이슈가 생겨나면서 2010년대 중후반부터 국제사회에서 논의되기 시작하였고, "AI 윤리"라는 개념으로 더욱 널리 알려져 있다. '신뢰'는 사전적 정의로 '굳게 믿고 의지함'을 말하는데, 여기서 '무엇' 때문에 대상을 믿을 수 있는가? 라는 질문에 대한 대답은 천차만별일 수 있다. 'Trustworthy'와 'Reliability' 두 단어 모두 '신뢰성'이라 번역되지만, 그 의미는 미세한 차이가 있다. Reliability가 성능이나 품질이 견고한가에 달려 있다면, Trustworthy는 결과에 따른 부작용 여부, 시스템이나 관리체계 수준 등으로 AI를 얼마나 믿고 쓸 수 있는가와 관계가 있다. AI 신뢰성을 "인공지능이 내포한 위험과 기술적 한계를 해결하고 활용·확산 과정에서의 위험·부작용을 방지하기 위한 가치 기준"으로 정의하기도 한다. 이처럼 신뢰성은 다양한 속성이 고려되어야 하는 포괄적 개념이다. 주로 AI의 설명 가능 여부와 투명성, 책임성, 공정성, 안전성 및 강건성, 그리고 개인정보보호 등이 이에 해당한다.
  - 정보 보안(security)은 중요한 정보의 공개, 변경, 파괴를 초래하는 무단 액세스 또는 데이터 유출 등의 위협으로부터 데이터 및 정보시스템을 보호하는 것을 말하며, 기밀성(Confidentiality), 무결성(Integrity) 및 가용성(Availability)의 세 가지 원칙을 보장하기 위한 다양한 기술적인 조치가 수반된다. **보안은 신뢰성 보장을 위한 여러 가지 속성 중 "안전성"과 밀접한 관련이 있다**. AI의 "안전성"은 시스템이 예측 가능하고, 해로운 행동이나 결과를 초래하지 않는 것을 말하며, AI가 인간 사용자와 환경에 해를 끼치지 않도록 하는 특성을 가지고 있다.

● 결론적으로 신뢰성과 안전성이 보안 보다 더 큰 개념이며 **보안은 AI의 안전성, 크게는 AI의 신뢰성을 보장하기 위한 핵심 구성요소로서 기능한다고 말할 수 있다. 따라서** 본 「인공지능(AI) 보안 안내서」는 **보안의 목표인 기밀성·무결성·가용성 보장을 위한 요구사항에 초점을 맞추었다.**

부록 그림 1 AI 신뢰 안내서」와 「인공지능(AI) 보안 안내서」의 윤리 기준 비교



또한 「AI 신뢰 안내서(과기정통부·TTA)」가 AI 개발자를 대상으로 한 것과 달리, 본 「인공지능(AI) 보안 안내서」는 **개발자 뿐만 아니라 서비스 제공자** 및 **이용자**를 대상으로 적용대상을 확대하였다.

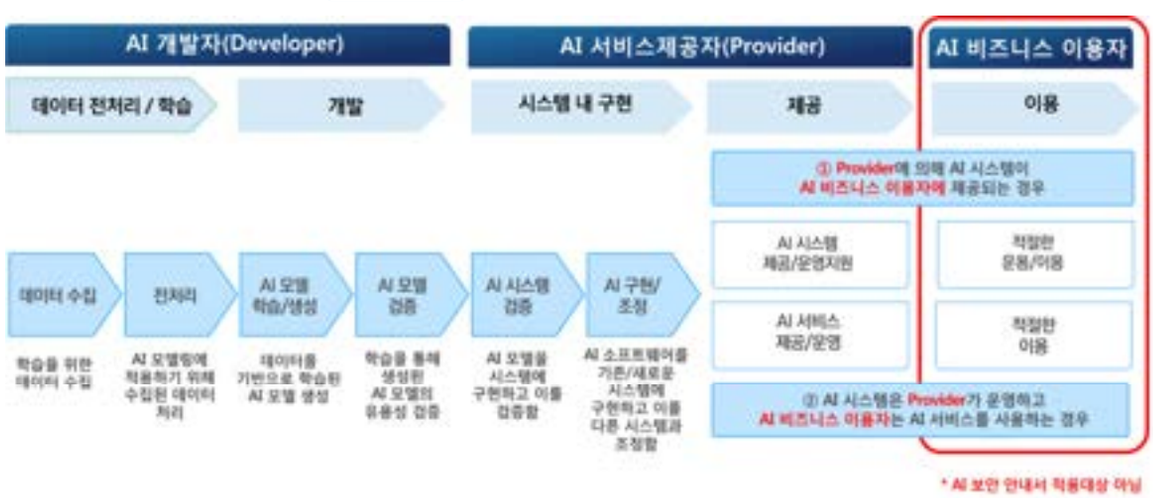
부록 표 1 AI 신뢰 안내서(TTA)와 AI 보안 안내서(KISA)의 적용 대상 비교

AI 신뢰 안내서 적용 대상		AI 보안 안내서 적용 대상
AI 모델, 시스템 등 개발자	확대	AI 모델, 시스템 등 개발자
		AI 서비스, 플랫폼 제공자
		AI 서비스 이용자

● 「인공지능(AI) 보안 안내서」는 AI 모델을 개발하는 개발자와 개발조직, 모델을 활용하여 서비스를 제공하는 사업자, 해당 서비스를 이용하여 결과물을 생성한 이용자 모두를 대상으로 하여 각 주체별로 AI 보안에 대한 안내서를 마련하였다. 이를 위해 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(인공지능 기본법, '26. 1. 22 시행예정), EU AI Act 등을 참고하여 개발자, 서비스 제공자, 이용자의 개념을 아래와 같이 정의하였다.

- "**개발자**"라고 하면 주로 소프트웨어 개발자(Developer 혹은 Engineer) 또는 개발 조직(기업)을 지칭하며, 이들은 시스템 분석가의 요구에 맞게 컴퓨터 프로그래밍을 하거나 시스템 설계를 하는 사람 또는 조직(기업)을 말한다. 안내서에서 개발자는 요구사항 및 검증항목에 따라 조직의 구성원 개인일 수도 있고 팀(조직) 또는 회사가 될 수도 있다. 따라서 개발자가 실제 이 안내서를 참고할 때 해당 내용이 개발자 개인에 관한 것인지 아니면 조직 또는 회사가 주도적으로 해야 할 것인지 여부에 대한 혼란이 있을 수 있다. 그래서 「인공지능(AI) 보안 안내서」에서는 요구사항별로 수행주체를 명시적으로 표시하였다.
- "**서비스제공자**"는 업으로서 AI 서비스 또는 AI 부수 서비스를 타인에게 제공하는 자를 말한다. "업으로" 한다는 것은 같은 행위를 계속하여 반복하는 것을 의미하고, 여기에 해당하는지 여부는 단순히 그에 필요한 인적 또는 물적 시설의 구비 여부와는 관계없이 행위의 반복·계속성 여부, 영업성의 유무, 그 행위의 목적이나 규모·횟수·기간·태양 등의 여러 사정을 종합적으로 고려하여 사회통념에 따라 판단하여야 한다. 이 안내서에서 서비스제공자는 영리를 목적으로 AI 서비스를 제공하는 법인(회사 등)을 말한다. 그러나 실제 업무 수행 시에는 임직원 개인이 해야 할 것인지 아니면 조직 또는 회사가 해야 할 것인지 불명확한 경우가 있을 수 있다. 따라서 「인공지능(AI) 보안 안내서」에서는 이를 명확히 하고자 요구사항 별로 수행주체를 표기하였다.
- "**이용자**"는 AI 서비스 또는 AI 부수 서비스를 타인에게 제공하지 않고 AI 서비스 또는 AI 부수 서비스를 이용하는 사람을 말한다. 이러한 "이용자" 개념에는 업으로서 AI 시스템 또는 AI 서비스를 이용하는 사람(이하 "AI 비즈니스 이용자"라고 함)이 포함될 수도 있으나, 본 「AI 이용자를 위한 보안 수칙」의 적용 대상은 AI 서비스를 이용하는 일반 국민을 대상으로 작성하였다.

부록 그림 2 AI 서비스 관련 사업 활동의 주체



보안 측면에서 중요한 **“파기”** 단계를 추가하여 AI 서비스의 생명주기 차별화

- 「AI 신뢰 안내서」가 인공지능 생명주기의 각 단계별로 인공지능의 신뢰성을 확보하기 위한 주요 요구사항을 5단계로 구분한 반면, 「인공지능(AI) 보안 안내서」는 보안 측면에서 중요한 **“파기”** 단계를 추가하여 총 6단계로 구분하였으며, 생명주기 단계별 세부 내용도 차별화하였다.

부록 그림 3 인공지능 서비스의 생명주기

AI 신뢰 안내서 생명주기
1. 계획 및 설계
2. 데이터 수집 및 처리
3. 인공지능 모델개발
4. 시스템 구현
5. 운영 및 모니터링

⟺

AI 보안 안내서 생명주기
1. 계획 및 설계
2. 데이터 수집 및 준비
3. 모델 개발(학습/모델링/검증)
4. 모델 배포
5. 모니터링 및 유지보수
6. 파기

- 「인공지능(AI) 보안 안내서」에서 정의한 각 단계별 목표와 주요 활동은 다음과 같다.

부록 표 2 인공지능 생명주기별 주요활동(인공지능(AI) 보안 안내서)

생명주기	목표	주요 활동
1. 계획 및 설계	AI 시스템이 해결할 목표 및 성공 지표를 정의	• AI가 해결할 수 있는 비즈니스 및 기술적 목표를 정의 • AI 시스템 관리 감독 조직 및 방안 마련 • AI시스템 위험요소 분석 및 대응 방안 마련
2. 데이터 수집 및 준비	AI 모델을 학습하고 개발하는 데 사용할 데이터를 수집하고, 사전 처리 및 모델 개발에 적합한 형식으로 변환	• 데이터 소스(구조화된 데이터 및 구조화되지 않은 데이터, 센서 데이터, 과거 데이터 세트 등)를 정의함 • 데이터 사용과 관련된 보안 정책 및 법적 제약을 고려함 • 누락된 데이터를 처리하고 중복을 제거하고 데이터 일관성 보장 • 데이터 세트를 학습, 검증 및 테스트 세트로 분할함
3. 모델 개발	AI 모델을 구축하고, 학습하여 성능 평가를 통해 필요한 지표를 충족	• 적절한 기술(예: 머신 러닝, 딥 러닝, 자연어 처리), 알고리즘(예: 의사 결정 트리, 신경망, SVM 등)과 모델 아키텍처 선택 • 준비된 데이터를 사용하여 모델 학습, 하이퍼파라미터 조정 • 모델을 검증 또는 보이지 않는 테스트 데이터 세트에서 테스트하여 정확도와 견고성을 확인함
4. 모델 배포	학습된 AI 모델을 실제 애플리케이션에서 예측할 수 있는 프로덕션 환경에 통합	• 클라우드 서비스, 에지 장치 또는 내부 서버 내에서 모델을 패키징하여 배포 • 모델이 실시간 또는 일괄 모드에서 다른 시스템이나 서비스와 상호 작용할 수 있는지 확인
5. 모니터링 및 유지보수	배포된 모델의 성능을 지속적으로 모니터링하고 시간이 지남에 따라 유지 관리	• 시간 경과에 따른 모델 성능 추적 • 배포 후에 나타나는 보안취약성을 감지하고 완화함 • 모델 및 기반 인프라에 대한 업데이트 및 패치를 구현
6. 파기	더 이상 유용하지 않거나 교체해야 할 때 AI 모델을 안전하게 폐기	• 모델을 폐기하기 전에 중요한 데이터와 로그를 백업 • 잔여 데이터나 지적 재산이 유출되지 않도록 함 • 폐기 사유와 향후 모델을 위해 얻은 교훈을 문서화

사이버 보안 위협으로부터 AI 서비스를 보호하는 것에 초점을 맞춤

- 「AI 신뢰 안내서」는 인공지능 신뢰성에 필요한 요건을 정의하고자 「인공지능 윤리기준」의 10대 핵심요건을 준용하여 기술적 관점에서 필요한 요구사항과 검증항목으로 ①다양성 존중, ②책임성, ③안전성, ④투명성을 도출하였다. 그리고 최종적으로 검증 가능한 **15개의 요구사항과 이에 매칭되는 69개의 정성·정량적 검증항목**이 도출되었다.

부록 표 3 인공지능 생명주기별 요구사항 및 검증항목 수(AI 신뢰 안내서)

생명주기	요구사항	검증항목
1. 계획 및 설계	4개	21개
2. 데이터 수집 및 처리	3개	18개
3. 인공지능 모델개발	4개	12개
4. 시스템 구현	3개	14개
5. 운영 및 모니터링	1개	4개
	15개	69개

- 반면에 「인공지능(AI) 보안 안내서」는 「인공지능 윤리기준」의 10대 핵심요건에서 편향성, 프라이버시 등 윤리적 요소를 모두 배제하고 "**안전성**"만을 고려하였다. 특히, 안전성 중에서도 **정보보안(기밀성·무결성·가용성)의 관점에서 사이버 보안 위협으로부터 AI 서비스를 보호하는 것에 초점을 맞추었다.** 그리고 최종적으로 검증 가능한 **14개의 요구사항**과 이에 매칭되는 **57개의 정성·정량적 검증항목**을 도출하였다.

부록 표 4 인공지능 생명주기별 요구사항 및 검증항목 수(인공지능(AI) 보안 안내서)

생명주기	요구사항	검증항목
1. 계획 및 설계	2개	6개
2. 데이터 수집 및 준비	3개	9개
3. 모델 개발(학습/모델링/검증)	4개	22개
4. 모델 배포	2개	9개
5. 모니터링 및 유지보수	2개	8개
6. 파기	1개	3개
	14개	57개

※ 금융보안원에서 발간한 금융분야 AI 보안 가이드라인에서는 데이터 수집 단계(1개 항목), 데이터 전처리(3개 항목), 설계·학습(4개 항목), 검증·테스트(6개 항목) 등 총 14개 검증항목 제시

예측형 AI(Pred AI)와 생성형 AI(Gen AI)에 맞게 구별하여 검증항목 제시

- 「인공지능(AI) 보안 안내서」는 요구사항별 검증항목에서 예측형 AI(이하, Pred AI)와 생성형 AI(Gen AI)를 구별하여 제시하였다. 예측형 AI(Pred AI)와 생성형 AI(이하, Gen AI)는 두 기술의 목적, 작동 방식, 위험 요소가 서로 다를 수 있으므로, 각 기술의 특성과 관련된 위협을 명확히 이해하고 이에 적합한 보안 대책을 수립하는 것이 중요하다. 따라서 본 「인공지능(AI) 보안 안내서」에서는 이를 구별하여 AI 개발자와 AI 서비스 제공자 대상 보안요구사항과 검증항목을 제시하였다.
  - Pred AI는 과거 및 현재 데이터를 사용하여 패턴을 식별하고 해당 정보를 기반으로 추론한다. 이는 주로 통계 알고리즘과 ML(기계학습)에 사용한다. 반면에 Gen AI는 한 단계 더 나아가 딥러닝을 사용하여 학습된 데이터를 기반으로 새로운 콘텐츠를 생성한다.
  - 이러한 기술적 특성으로 인해 데이터 관련 문제라 하더라도 위험 요소가 서로 다를 수 있다. 예를 들어 Pred AI는 정확한 예측을 지원할 수 있도록 고품질 데이터와 라벨링이 필요할 것이고, Gen AI는 모델의 학습 기반이 된 데이터를 제공하는 오픈소스 모델을 안전하게 사용하는 것에 초점을 맞추는 것이 중요하므로 이에 맞는 보안 검증항목이 필요하다.
  - 따라서 개발자나 서비스 제공자 등은 발생한 위험이 어떠한 AI 유형과 관련이 있는지 사전에 파악하는 것이 매우 중요하고, 이를 반영한 검증항목이 필요할 것으로 예측된다. 이에 본 「인공지능(AI) 보안 안내서」에서는 요구사항 및 검증항목 별로 이를 구분하여 제시하였다. 예측형 AI와 생성형 AI에 따라 구별하여 점검하면, 각 기술에 필요한 보안 조치를 적절하게 파악하고 자원을 효율적으로 분배할 수 있어, 과도하거나 불필요한 보안 비용을 줄일 수 있을 것으로 기대된다.

요구사항의 공통점과 차이점 분석(요약비교)

- 「AI 신뢰 안내서」와 「인공지능(AI) 보안 안내서」 모두 요구사항 도출에 있어서 "안전성"을 고려하기 때문에 해당 부분에 한해서 일부 중복이 있을 수 있으나, 「인공지능(AI) 보안 안내서」에서는 보안을 중심으로 보다 더 **구체적으로 요구사항을 정의**하고, 각 요구사항별 하위 검증항목에서도 차별화하였다.

부록 표 5 안전성 관련 요구사항 정의

AI 신뢰 안내서	다양성 존중	책임성	안전성	투명성	AI 보안 안내서
02. 인공지능 거버넌스 체계 구성	○	○	○	○	1.1 AI 보안 거버넌스 체계 구축
05. 데이터 강건성 확보를 위한 이상 데이터 점검			○		2.3 데이터 공격에 대한 방어
07. 오픈소스 라이브러리의 보안성 및 호환성 확보		○	○		3.3 오픈소스 라이브러리 보안
09. 인공지능 모델 공격에 대한 방어 대책 수립			○		3.2 모델 공격에 대한 방어

부록 표 6 요구사항별 검증항목 분석

AI 신뢰 안내서	인공지능(AI) 보안 안내서	차별점
02. 인공지능 거버넌스 체계 구성	1.1 AI 보안 거버넌스 체계 구축	• 실제 조직 내 정책 및 인력은 다르게 구성됨
※ 정책, 조직, 인력 등 거버넌스 관련 일반적인 공통사항에 해당		
05. 데이터 강건성 확보를 위한 이상(Abnormal) 데이터 점검	2.3 데이터 공격에 대한 방어	• 각 공격 유형을 세분화해 방어 대책 제시
• 데이터 중독, 회피 등 방어 대책	• 데이터 중독 방어 대책 • 데이터 회피 방어 대책 • 데이터 유출·변조 방어 대책	
07. 오픈소스 라이브러리의 보안성 및 호환성 확보	3.3 오픈소스 라이브러리 보안	• 오픈소스 라이브러리 관련 검토 사항을 보안관점에서 세분화해 확인 사항 제시
• 오픈소스 라이브러리의 안정성 확인, 위험요소 관리, 라이선스 준수 사항, 호환성 및 보안취약점 확인(포괄적 접근)	• 오픈소스 라이브러리의 업데이트 및 취약점 관리 • 오픈소스 라이브러리의 소스코드 직접 검토 및 사용에 대한 보안 문제 검토 • 오픈소스 라이브러리의 실행 시 격리된 환경 이용	
09. 인공지능 모델 공격에 대한 방어 대책 수립	3.2 모델 공격에 대한 방어	• 모델 공격에 대한 유형을 보안 관점에서 세분화 하여 확인사항 제시
• 모델 추출 공격 및 모델 회피 공격에 대한 방어	• 데이터 및 기능 도용에 대한 대책 • 모델 추출 공격에 대한 대책 • 모델 회피 공격에 대한 대책 • 모델 포이즈닝에 대한 대책 • 적대적 예제 공격에 대한 대책 • 모델 탈취 및 리버스 엔지니어링에 대한 대책 • 반복적인 질의에 대한 방어 대책 • 기계 학습을 활용한 모델 공격에 대한 대책	

02 인공지능(AI) 보안 안내서 보안 요구사항 및 검증항목과 정보보안 원칙 간 비교

본 안내서의 모든 보안 요구사항은 기밀성·무결성·가용성 및 책임추적성과 매핑이 가능하도록 구성하였다. 아울러 보안 관점에서 개발자의 이해를 돕기 위해 관련 사례 및 예시를 충분히 제시하고자 하였다.

부록 그림 4 개발자 대상 보안 요구사항과 정보보안 원칙의 매핑

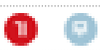
생명주기	검증항목	보안 원칙
1. 기획 및 설계	1-1 AI 보안(Security) 거버넌스 체계 구축	● ● ●
	1-2 AI 모델 개발에 대한 위험관리	● ● ●
2. 데이터 수집 및 준비	2-1 데이터 수집 및 전처리	● ●
	2-2 데이터 무결성 점검	●
	2-3 데이터 공격에 대한 방어	● ●
3. 모델 개발 (학습/모델링/검증)	3-1 학습/검증 환경에 대한 보안	●
	3-2 모델 공격에 대한 방어	● ● ●
	3-3 오픈 소스 라이브러리 보안	● ●
	3-4 LLM 보안	● ●
4. 모델 배포	4-1 모델파일 및 배포 환경 보호	●
	4-2 API 및 인터페이스 보안	● ●
5. 모니터링 및 유지보수	5-1 실시간 모니터링	●
	5-2 보안 패치 및 업데이트 관리	● ●
6. 파기	6-1 파기 시 보안	● ●

- 서비스 제공자를 위한 「인공지능(AI) 보안 안내서」는 **AI 서비스 제공 Life Cycle** – ①서비스 계획 및 설계(Service Planning and Design), ②서비스 개발 및 구축(Service Development and Deployment), ③서비스 제공 및 운영(Service Delivery and Operation), ④서비스 유지보수 및 지원(Service Maintenance and Support), ⑤피드백 및 서비스 개선(Feedback and Imprrovement) – 각각에서 보안점검 사항을 도출하였다.
  - 도출된 점검사항은 개발자의 경우와 마찬가지로 기밀성, 무결성, 가용성 및 책임추적성과 매핑이 가능하다.

부록 표 7 서비스 제공자를 위한 보안 요구사항 및 검증항목 수(인공지능(AI) 보안 안내서)

서비스 제공 Life Cycle	요구사항	검증항목
1. 서비스 계획 및 설계	3개	9개
2. 서비스 개발 및 구축	4개	15개
3. 서비스 제공 및 운영	2개	7개
4. 서비스 유지보수 및 지원	2개	9개
5. 피드백 및 서비스 개선	1개	4개
합계	12개	44개

부록 그림 5 서비스제공자 대상 보안 요구사항과 정보보안 원칙의 매핑

생명주기	검증항목	보안 원칙
1. 서비스 설계 및 계획	1-1 AI 보안(Security) 거버넌스 체계 구축	
	1-2 AI 서비스에 대한 위험관리	
	1-3 서비스 수준 계약(SLA) 관리	
2. 서비스 개발 및 구축	2-1 코드 및 알고리즘 보안	
	2-2 모델 환경의 보안	
	2-3 데이터 보안	
	2-4 API 및 인터페이스 보안	
3. 서비스 제공 및 운영	3-1 로그 및 운영 데이터	
	3-2 파기	
4. 서비스 유지보수 및 지원	4-1 모니터링, 업데이트 및 배치	
	4-2 성능 및 장애관리	
5. 피드백 및 서비스개선	5-1 사용자 피드백 관리	

03 이용자 대상 AI 안내서 비교

각 부처에서 발표한 AI 관련 안내서·가이드는 주로 AI 서비스 개발자가 적용대상이고, 챗GPT 등 특정 서비스 이용자에 대해 윤리 또는 신뢰에 근거한 행동 수칙을 담고 있다. 이는 AI 이용자 대상 가이드를 발표한 캐나다, 일본 등 해외의 경우에도 크게 다르지 않아 보인다.

- 이에 반해, 이용자 대상 「인공지능(AI) 보안 안내서」는 일반적인 원칙 선언에 그치지 않고 AI 서비스 접속·이용 단계에서 **보안 강화를 위해 이용자가 지켜야 할 구체적인 행동 지침을 제공**하는 것에 초점을 맞추어 차별화하였다.

부록 표 8 이용자 대상 인공지능(AI) 보안 안내서 비교

구 분	이용자 대상 가이드 유무		내용	
	없음	가이드 제시	일반 원칙	구체적인 행동수칙 제시
인공지능(AI) 보안 안내서(본 안내서)		○	○	○
신뢰할 수 있는 인공지능 개발 안내서 (과기정통부/TTA, '24.4)	○			
챗GPT 등 생성형 AI 활용 보안 가이드라인 (국가정보원/국가보안기술연구소, '23.6)		○ (챗GPT 중심)		○ (챗GPT 중심)
안전한 AI 시스템 개발을 위한 가이드라인 (국가정보원, '23.11)	○			
금융분야 인공지능 가이드라인 (금융위, '21.7)	○			
금융분야 AI 보안 가이드라인 (금융보안원, '23.4)	○			

04 TTA 신뢰할 수 있는 인공지능 개발 안내서와 요구사항별 상세비교

- TTA 신뢰할 수 있는 인공지능 개발 안내서와 요구사항별 상세 비교를 하면, 유사항목은 6개이고, 본 〈인공지능(AI) 보안 안내서〉에서 보안 관점에서 확대된 항목은 20개, 신규로 추가된 항목은 31개로 나타나 검증항목 내용상 차이가 있음

< AI 신뢰 안내서 >
* 15개 요구사항 69개 검증항목

요구사항 및 체크리스트		
01	인공지능 시스템의 위험 관리 계획 및 수행	4개→3개
01-1	인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험요소를 분석하였는가?	
01-1a	인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?	
01-1b	인공지능 기술 적용을 어렵게 만드는 위험 요소가 있는지 확인하였는가?	
01-2	위험요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	
01-2a	위험 요소별 완화 또는 제거 방안을 마련하였는가?	
01-2b	위험 요소의 파급효과가 감소하였는지 확인하였는가?	
02	인공지능 거버넌스 체계 구성	5개→3개
02-1	인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?	
2-1a	내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?	
2-2	인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?	
2-2a	인공지능 거버넌스를 위한 조직을 구성하였는가?	
2-2b	인공지능 거버넌스를 위한 조직은 전문성을 갖춘 인력으로 구성하였는가?	
2-3	인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?	
2-3a	인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?	
2-4	인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?	
2-4a	기존 동일 목적의 시스템과 비교하여, 신규 시스템이 개선할 수 있는 사항을 분석하였는가?	
03	인공지능 시스템의 신뢰성 테스트 계획 수립	
03-1	인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?	
03-1a	테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?	
03-1b	가상테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?	
03-2	인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?	
03-2a	인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?	
03-2b	설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?	

< 인공지능(AI) 보안 안내서 >
* 14개 요구사항 57개 검증항목
(유사 6개/ 확대 20개 / 신규 31개)

요구사항 및 체크리스트		
1.2	AI 모델개발에 대한 위험관리 계획의 수립	유사
1.2.1	AI 모델개발 생명주기에 걸쳐 나타날 수 있는 위험요소를 분석·도출하고 있는가?	유사
1.2.2	AI 시스템에 대한 위협 모델링 및 위험 평가를 수행하고 있는가?	유사
1.2.3	AI 시스템에 대한 위험요소를 제거·완화하기 위한 방안을 마련하고 있는가?	유사
1.1	AI 보안(Security) 거버넌스 체계 구축	유사
1.1.1	AI 보안(Security) 거버넌스를 위한 조직이 구성되어 있는가?	유사
1.1.2	AI 보안(Security) 거버넌스를 위한 정책, 절차, 프로세스가 구현되어 있는가?	유사
1.1.3	AI 보안(Security) 거버넌스를 위한 전문인력을 갖추고 있는가?	유사

※ 보안을 위한 별도의 거버넌스 필요성 존재

04	인공지능 시스템의 추적가능성 및 변경이력 확보
04-1	인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?
04-1a	인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?
04-1b	인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?
04-1c	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?
04-2	학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?
04-2a	데이터 흐름 및 계보를 추적하기 위한 조치를 마련하였는가?
04-2b	데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?
04-2c	데이터 변경 시, 버전 관리를 수행하였는가?
04-2d	데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?
04-2e	신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?
05	데이터 활용을 위한 상세 정보 제공
05-1	데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?
05-1a	정제 전과 후의 데이터 특성을 설명하였는가?
05-1b	학습 데이터의 메타데이터를 구분하고 각 명세자료를 확보하였는가?
05-1c	보호변수의 선정 이유 및 반영 여부를 설명하였는가?
05-1d	라벨링 작업자를 위한 교육을 시행하고 작업 가이드 문서를 마련하였는가?
05-2	데이터의 출처는 기록 및 관리되고 있는가?
05-2a	신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?
05-2b	오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?
06	데이터 견고성 확보를 위한 이상 데이터 점검
06-1	이상 데이터의 식별 및 정상 여부를 점검하였는가?
06-1a	전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?
06-1b	학습 데이터 이상값 식별 기법을 적용하였는가?
06-2	데이터 공격에 대한 방어 수단을 강구하였는가?
06-2a	데이터 최적화를 통한 방어 대책을 마련하였는가?

1개→9개

2.1	데이터 수집 및 전처리	확대
2.1.1	데이터 수집 시 사용되는 네트워크 프로토콜이 충분한 보안 기능을 제공하고 있는가?	신규
2.1.2	수집된 데이터의 보관 및 삭제 절차가 명확하게 정의되어 있는가?	신규
2.1.3	전처리 과정에서 중요 데이터를 보호하기 위해 암호화 기술을 사용하고 있는가?	신규
2.2	데이터 무결성 검증	확대
2.2.1	데이터 저장 및 전송 시 데이터 무결성을 검증하고 있는가?	신규
2.2.2	데이터 처리 과정에서 데이터 무결성을 검증하고 있는가?	신규
2.2.3	데이터에 접근할 수 있는 권한을 제한하고 있는가?	확대
2.3	데이터 공격에 대한 방어	확대
2.3.1	데이터 중독(poisoning) 공격에 대한 방어 대책을 마련하고 있는가?	확대
2.3.2	데이터 회피(evasion) 공격에 대한 방어 대책을 마련하고 있는가?	확대
2.3.3	데이터 유출·변조 공격을 방지하기 위한 방안을 마련하고 있는가?	신규

07	수집 및 가공된 학습 데이터의 편향 제거
07-1	데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?
07-1a	인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?
07-1b	데이터의 다양성 확보를 위해 여러 수집 장치를 활용하였는가?
07-2	학습에 사용되는 특성을 분석하고 선정 기준을 마련하였는가?
07-2a	보호변수 선정 시 충분한 분석을 수행하였는가?
07-2b	편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?
07-2c	데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?
07-3	데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?
07-3a	데이터 라벨링 기준을 명확히 수립하고 작업작에게 제공하였는가?
07-3b	다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?
07-4	데이터의 편향 방지를 위한 샘플링을 수행하였는가?
07-4a	편향 방지를 위한 샘플링 기법을 적용하였는가?
08	오픈소스 라이브러리의 보안성 및 호환성 검증
08-1	오픈소스 라이브러리의 안정성을 확인하였는가?
08-1a	활성화된 오픈소스 라이브러리를 사용하였는가?
08-2	오픈소스 라이브러리의 위험 요소는 관리되고 있는가?
08-2a	사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?
08-2b	사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?

1개→3개

3.3	오픈소스 라이브러리 보안	확대
3.3.1	오픈소스 라이브러리의 업데이트 및 취약점을 관리하고 있는가?	확대
3.3.2	오픈소스 라이브러리의 소스 코드를 직접 검토하거나 사용에 대한 보안 문제를 검증하고 있는가?	확대
3.3.3	오픈소스 라이브러리를 실행할 때 잠재적인 보안 위험을 제거하기 위해 격리된 환경을 이용하고 있는가?	확대

09	인공지능 모델의 편향 제거
09-1	모델 편향을 제거하는 기법을 적용하였는가?
09-1a	개발하려는 모델에 맞게 편향 제거 기법을 선택하였는가?
09-1b	편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

10	인공지능 모델 공격에 대한 방어 대책 수립
10-1	모델 공격이 가능한 상황을 파악하였는가?
10-1a	데이터 유형별 공격 가능한 적대적 사례를 확인하였는가?
10-2	모델 공격에 대한 방어 수단을 강구하였는가?
10-2b	모델 최적화를 통한 방어 대책을 마련하였는가?

2개→8개

3.2	모델 공격에 대한 방어	확대
3.2.1	AI 모델이 적대적 의도를 가진 사용자에 의해 데이터 및 기능을 도용당하거나 다른 공격에 악용되지 않도록 대책을 수립하고 있는가?	확대
3.2.2	모델 추출 공격에 대한 방어 방안을 수립하고 있는가?	확대
3.2.3	모델 회피 공격에 대한 방어 방안을 수립하고 있는가?	확대
3.2.4	모델 포이즈닝에 대한 방어 방안을 수립하고 있는가?	확대
3.2.5	적대적 예제 공격에 대한 방어 방안을 수립하고 있는가?	확대
3.2.6	모델 탈취 및 리버스 엔지니어링에 대한 방어 방안을 수립하고 있는가?	신규
3.2.7	반복적인 질의에 대한 방어 방안을 수립하고 있는가?	신규
3.2.8	기계 학습을 활용한 모델 공격에 대해 능동적으로 방어하고 있는가?	신규

11	인공지능 모델 명세 및 추론 결과에 대한 설명 제공
11-1	인공지능 모델의 명세를 투명하게 제공하는가?
11-1a	사용자 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?
11-2	사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?
11-2a	인공지능 모델에 적합한 XAI 기술을 적용하였는가?
11-2b	XAI 기술 적용이 불가능한 경우, 기술 외 대안을 마련하였는가?
11-3	모델 추론 결과에 대한 사용자의 판단을 도울 수 있는 설명을 제공하는가?
11-3a	모델 추론 결과에 대한 설명이 필요한지 검토하였는가?
11-3b	사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?

3.1	학습/검증 환경에 대한 보안	신규
3.1.1	모델 학습을 진행하는 환경이 안전하게 보안조치되어 있는가?	신규
3.1.2	학습 또는 검증 단계에서 악의적인 사용자가 허위 데이터를 삽입할 가능성을 차단하고 있는가?	신규
3.1.3	연합 학습(Federated Learning)에 참여하는 장치 중 악의적인 장치가 있는지 검증하고 있는가?	신규
3.4	LLM 보안	신규
3.4.1	LLM 애플리케이션 공격에 대한 예방책을 마련하고 있는가?	신규
3.4.2	LLM의 Model Denial of Service 공격에 대한 방어 방안을 수립하고 있는가?	신규
3.4.3	LLM의 API 보안을 위한 방안을 수립하고 있는가?	신규
3.4.4	LLM의 인터페이스 공격에 대한 예방책을 마련하고 있는가?	신규
3.4.5	개발 환경에서 LLM을 사용할 때 잠재적인 취약성의 통합을 방지하기 위한 안전한 코딩 관행과 지침을 수립하고 있는가?	신규
3.4.6	LLM 출력결과를 정기적으로 모니터링하고 검토하고 있는가?	신규
3.4.7	LLM의 Prompt Injection 공격에 대한 방어 방안을 수립하고 있는가?	신규

12	인공지능 시스템 구현 시 발생 가능한 편향 제거
12-1	소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?
12-1a	데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?
12-1b	사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?
13	인공지능 시스템의 안전모드 구현 및 문제발생 알림 절차 수립
13-1	공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?
13-1a	문제 상황에 대한 예외 처리 정책이 마련되어 있는가?
13-1b	인공지능 데이터 및 모델 공격에 대해 시스템 측면의 방어 대책을 마련하였는가? (* 06-2 데이터 공격 방어수단, 10 모델 공격 방어 대책 내용과 유사)
13-1c	인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?
13-1d	예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?
13-2	인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?
13-2a	편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?
13-2b	시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?
14	인공지능 시스템의 설명에 대한 사용자의 이해도 제고
14-1	인공지능 시스템 사용자의 특성과 제약사항을 분석하였는가?
14-1a	사용자 특성에 따른 세부 고려사항을 분석하였는가?
14-2	사용자 특성에 따른 설명을 제공하는가?
14-2a	사용자 특성에 따른 설명 평가 기준을 수립하였는가?
14-2b	사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?
14-2c	사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?
14-2d	설명이 필요한 위치와 타이밍은 적절한가?
14-2e	사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?

3.4.8	LLM의 벡터 및 임베딩 취약점에 대한 방어 방안을 수립하고 있는가?	신규

4.1	모델파일 및 배포 환경 보호	신규
4.1.1	모델을 배포하기 전에 코드 및 모델을 스캔하고, 자동화된 취약점 분석을 하고 있는가?	신규
4.1.2	모델파일을 암호화하여 저장하고 전송 중에도 안전하게 보호하고 있는가?	신규
4.1.3	AI 모델이 배포되는 인프라(클라우드, 서버 등) 환경이 충분한 보안시스템을 갖추고 있는가?	신규
4.2	API 및 인터페이스 보안	신규

15	서비스 제공 범위 및 상호작용 대상에 대한 설명 제공
15-1	인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?
15-1a	서비스의 목적과 목표에 대한 설명을 제공하는가?
15-1b	서비스의 한계와 범위에 대한 설명을 제공하는가?
15-2	사용자가 상호작용하고 있는 대상을 인지할 수 있도록 설명하는가?
15-2a	사용자와 인공지능이 상호작용하는 서비스 범위를 명시하였는가?
15-2b	서비스 내의 최종 의사결정을 수행하는 주체에 대해 사용자에게 설명하는가?

4.2.1	AI 모델이 배포된 후, API를 통해 외부 시스템과 상호작용하는 경우, 충분한 보안 조치 기능을 갖추고 있는가?	신규
4.2.2	배포된 AI 모델이 실시간으로 데이터를 수신하고 이를 처리할 때, 중간자 공격(Man-in-the-Middle Attack)에 대응하고 있는가?	신규
4.2.3	AI 모델의 API에 대한 접근 권한을 제한하고, 강한 인증 메커니즘을 사용해 불법 접근을 방지하고 있는가?	신규
4.2.4	API 사용자는 필요한 권한만 부여받도록 최소 권한 원칙을 적용하고 있는가?	신규
4.2.5	AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 하고 있는가?	신규
4.2.6	API 호출을 실시간으로 모니터링하여 비정상적인 패턴이나 오용 시도를 감지하고 차단하고 있는가?	신규

5.1	실시간 모니터링	신규
5.1.1	모델의 입력 데이터, 출력 결과 등을 실시간으로 모니터링하여 비정상적인 동작을 탐지하고 있는가?	신규
5.1.2	모델 응답 시간, 사용 패턴을 추적하고 분석하여 보안에 의심스러운 행동을 탐지하고 있는가?	신규
5.1.3	AI 모델이 동작하는 서버 및 네트워크의 트래픽을 모니터링하여 비정상적인 요청을 탐지하고 있는가?	신규
5.1.4	API 호출, 입력/출력 등 요청로그를 정기적으로 분석하여 보안에 의심스러운 동작을 탐지하고 있는가?	신규
5.1.5	AI 모델과 배포 환경에 대해 모의 해킹을 수행하여 잠재적인 보안 취약점을 탐지하고 수정하고 있는가?	신규
5.2	보안 패치 및 업데이트 관리	신규
5.2.1	모델에 대한 보안 패치 및 업데이트 관리 프로세스를 구축하고 있는가?	신규
5.2.2	모델 배포 후 모델 및 라이브러리의 업데이트가 정기적으로 이루어지고 있는가?	신규
5.2.3	운영 체제, 라이브러리, 프레임워크의 보안 패치를 운영 환경에 적용하기 전에 스테이징 환경에서 패치를 테스트하고 있는가?	신규
6.1	파기 시 보안	신규
6.1.1	AI 모델이 더 이상 사용되지 않으면, 모델 파일을 완전히 삭제하고 복구할 수 없도록 처리하고 있는가?	신규
6.1.2	AI 모델에서 사용 중이던 데이터가 시스템을 폐기하거나 교체할 때 안전하게 삭제되고 있는가?	신규
6.1.3	AI 모델이 더 이상 사용되지 않으면, 해당 모델과 연결된 API나 인터페이스를 비활성화하여 외부 접근을 차단하고 있는가?	신규

04 국내 주요 AI 보안 가이드라인 비교

최근 몇 년 동안 각 부처는 소관 업무와 관련된 인공지능 정책을 수립하고 있으며, AI 운영의 방향성을 보여주는 안내서를 발표하고 있음.

국내 주요 AI 안내서

구분	챗GPT 등 생성형 AI 활용 보안 가이드라인 (국가정보원, '23. 6.)	금융분야 AI 보안 가이드라인 (금융보안원, '23. 4.)	신뢰할 수 있는 인공지능 개발 안내서 – 일반분야 (과기정통부, '23.7)	인공지능 시대 안전한 개인정보 활용 정책 방향 (개인정보위, '23.8)
배경	• 챗 GPT 등 생성형 AI를 안전하게 활용할 수 있도록 정부 차원의 보안 대책 필요	• 국내·외 금융분야의 AI 기반 금융서비스 및 시스템에 대한 신뢰성 확보를 위한 대책 필요	• 국내 AI 서비스, 시스템 개발 환경 및 실제 실무를 기반하는 요구사항 및 검증 항목 필요	• 생성형 AI, 자율주행차, 로봇 등 데이터 수집·활용에서 발생하는 개인정보 침해에 대한 정부 차원의 규범 필요
목적	• 생성형 AI를 활용하는 과정에서 업무상 비밀·개인정보 유출 등 기술 악용 및 보안 문제를 사전 예방하기 위한 방안 제시	• 금융산업 AI 학습용 데이터 정확성·안정성 확보, AI 금융 서비스의 투명성·공정성 담보 등을 하기 위한 방안 제시	• AI 서비스 개발 시 최소한의 신뢰성을 확보하고, 중소기업 등의 자율 점검체계 구축 등에 기여	• AI 환경에서의 프라이버시 침해 위험 최소화하기 위한 정책 방향 수립 및 관련 추진 과제 제시
주요 대상	• 정부, 공공·민간분야 및 일반 국민	• 금융기관, 금융 서비스 개발 업체 등	• AI 서비스 개발 실무자 및 AI 서비스 관련 기업·기관	• AI 모델·서비스 개발·제공 기업, 기관 및 이용자
목차 및 주요 내용	1. 개요 2. 생성형 인공지능 기술의 대표적인 보안 위협 ※ AI 서비스 취약점 공격, 사이버 범죄 악용, 불법 콘텐츠 생성, 정보 유출 등 사례 소개 3. 안전한 생성형 인공지능 기술 사용 가이드라인 ※ 기관의 정보보호·정보보안 담당자 등을 대상으로 효율적으로 안전하게 서비스를 활용할 수 있는 방안 제시 ※ (안내사항) 서비스 사용 주의사항, 서비스와 대화 시 주의사항, AI 모델 플러그인 사용 주의사항, AI 모델 확장 프로그램	1. 개요 2. AI 서비스 구성 ※ AI 서비스의 전반적인 구성, 기능 및 운영 방안 등 소개 3. AI 학습 데이터 및 모델 보안 관리 ※ 다음 4단계의 AI 모델 개발 주기에 따른 보안 방안 제시	1. 개요 2. 요구사항 및 검증항목 ※ 다음의 AI 생명주기 별 안전한 시스템 구축, 서비스 활용을 위한 요구사항 제시 ① 계획 및 설계: ▲위험관리 계획 및 수행 ▲거버넌스 체계 구성 ▲시스템 신뢰성 테스트 계획 수립 ② 데이터 수집 및 처리: ▲데이터 활용을 위한 상세 정보 제공 ▲이상 데이터 점검 ▲학습 데이터의 편향 제거 ③ 인공지능 모델 개발 : ▲모델 편향 제거 ▲모델 공격에 대한 방어 대책	1. 추진 배경 2. AI 데이터 처리방식 변화와 프라이버시 이슈 3. AI와 개인정보 보호 기본원칙 ※ ①헌법상 개인정보자기결정권, ②개인정보보호 원칙 기반 AI에 대한 기본원칙 제시 4. 인공지능 단계별 데이터 처리기준과 보호조치 ※ ①AI 모델·서비스 기획 ②데이터 수집 ③데이터 학습 ④AI 서비스 각 단계별 개인정보 처리, 활용, 관리 및 정보보호 등을 위한 정책적 필요사항

<table>
<tr><th>구분</th><th>챗GPT 등 생성형 AI 활용 보안 가이드라인
(국가정보원, ‘23. 6.)</th><th>금융분야 AI 보안 가이드라인
(금융보안원, ‘23. 4.)</th><th>신뢰할 수 있는 인공지능 개발 안내서 – 일반분야
(과기정통부, ’23.7)</th><th>인공지능 시대 안전한 개인정보 활용 정책 방향
(개인정보위, ’23.8)</th></tr>
<tr><td></td><td>사용 주의사항, AI 모델 생성 기반 공격 대처 방안
4. 생성형 인공지능 기반 정보화사업 구축 방안 및 보안 대책
※ ①구축 유형별 생성 AI 기술 도입시 고려사항, ② AI 모델 API 활용 및 민간 AI 모델 도입 방안, ③ 자체 데이터 세트 및 AI 모델 구축 단계별* 보안 위협 대응 방안 등으로 구성하여 보안 대책 제시
* (AI 모델 구축 단계) ▲ 데이터 수집/전처리 ▲모델 학습 ▲평가 및 테스트 ▲ 배포 및 서비스</td><td><table><tr><th>단계</th><th>주요 내용</th></tr><tr><td>학습 데이터 수집</td><td>AI 모델 학습에 필요한 데이터 수집 단계</td></tr><tr><td>학습 데이터 전처리</td><td>AI 모델에 적합한 형태로 변환하고 데이터 오염 여부 확인 단계</td></tr><tr><td>모델 설계·학습</td><td>AI 모델을 설계하고 학습하는 단계</td></tr><tr><td>모델 검증·테스트</td><td>AI 모델의 성능 확인 및 테스트 단계</td></tr></table>
※ AI 모델 대상 보안 공격 기법: 데이터 오염 공격, 모델 오염 공격, 모델 추출 공격, 모델 인버전 공격, 회피 공격</td><td>수립, ▲모델 명세 및 추론 결과에 대한 설명 제공
④ 시스템 구현: ▲시스템에서의 편향 제거 ▲안전모드 구현 및 문제 발생 알림 절차 수립 ▲사용자 이해도 제고
⑤ 운영 및 모니터링: ▲추적 가능성 및 변경이력 확보, ▲서비스 제공 범위 및 상호작용 대상 설명 제공
3. 부록</td><td>분석 및 정책 추진 방향 제시

5. 원칙 기반 규율 추진체계
※ AI 분야 정부·민간 간 소통·협력의 구심점으로서 개인정보 규제를 함께 설계하고 국제적인 공조 체계를 구축 등 추진
※ AI 프라이버시 전담팀 및 AI 프라이버시 민·관 정책협의회 구성·운영
6. 향후 계획</td></tr>
<tr><td>응용 분야</td><td>• 정부 서비스 및 기타 민간 분야</td><td>• 금융 서비스, 시스템 등 AI 적용이 가능한 금융 분야 전반</td><td>• AI 적용이 가능한 공공·민간 全분야</td><td>• AI 적용이 가능한 공공·민간 全분야</td></tr>
</table>

05 AI 개발자 대상 보안 프레임워크

01 프레임워크 개념과 필요성

현재 많은 분야에서 다양하게 프레임워크라는 말을 쓰고 있고 기술과 시대가 변하면서 그 의미도 조금씩 변하고 있으나, 큰 틀에서 프레임워크의 정의를 내리면 다음과 같다.

- 첫째, 만들고자 하는 구조물의 기본 골격으로 아키텍처와 마찬가지로 소프트웨어 공학 뿐만 아니라 건축, 비행기, 선박, 다리 같은 구조물을 만드는 모든 분야에서 발견할 수 있는 개념이라고 할 수 있다. 또한 실체가 있는 구조물뿐만 아니라 실체가 없는 정책, 전략에도 쓰인다.
- 둘째, 정부정책 라이프사이클 프레임워크(GPLC framework)에 따르면 정책수립에서 정책집행까지 정책과 관련된 모든 과정의 틀을 제공하는 것이라고 할 수 있다. 이처럼 프레임워크는 자신이 다루는 대상의 틀을 결정하는 것이라 할 수 있다.
- GoF(Gang of Four)의 디자인 패턴으로 유명한 랄프 존슨(Ralph Johnson) 교수[14]는 프레임워크를 공학적 측면으로 "소프트웨어의 구체적인 부분에 해당하는 설계와 구현을 재사용이 가능하게끔 일련의 협업화된 형태로 클래스들을 제공하는 것"이라고 정의하였다.
- 이에 AI 보안 프레임워크는, 다가올 AI 산업의 능동적인 대응 방안을 마련하기 위해 상호 관련성이 있는 기술과 지식을 일목요연하게 정리하여 신뢰할 수 있는 AI 기반 산업의 적용 및 활성화를 위한 전체적인 기반 구조를 만드는 것이라 할 수 있다.

프레임워크의 필요성

- 안전한 AI 개발을 위한 정보보안은 다양한 새로운 기술 요소로 구성되어 있어 이를 통합하고 관리하는 것은 매우 복잡한 작업이다. 이러한 복잡성을 효과적으로 관리하고 개발과 구현의 일관성을 유지하며 시스템에 대한 안전성과 상호 운용성을 높이기 위해 AI 시스템과 서비스에 대한 정보보안 프레임워크는 필요하다고 판단된다.
- 안전한 AI 개발을 위한 정보보안 프레임워크는 전략적·관리적·정책적·기술적 및 메커니즘 측면에서 AI 환경 내 발생할 수 있는 침해 문제를 해결하고 이를 통해 개발자는 사용자에게 더욱 안전하고 투명한 AI 서비스 환경을 제공할 수 있을 것으로 판단된다.

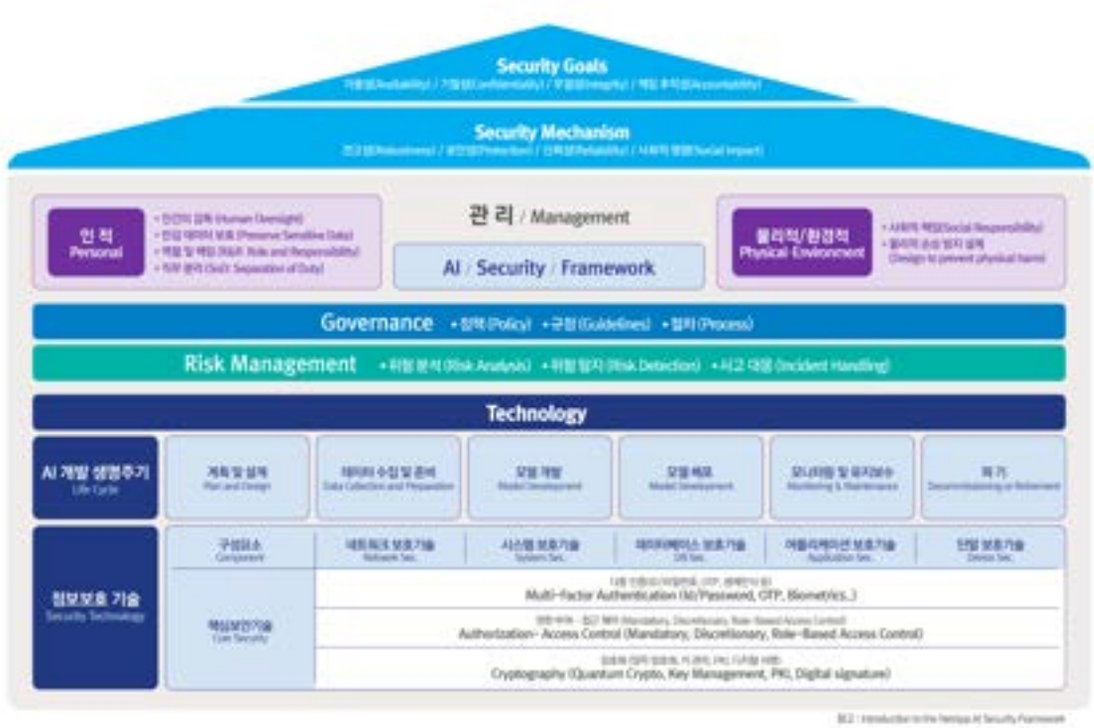
14 랄프 존슨(Ralph Johnson) 교수는 소프트웨어 공학 및 디자인 패턴 분야에서 매우 중요한 인물로, 특히 디자인 패턴의 고전인 GoF(Gang of Four)의 공동 저자 중 한 사람이다. GoF는 에릭 감마(Erich Gamma), 리처드 헬름(Richard Helm), 랄프 존슨(Ralph Johnson), 존 블리시디스(John Vlissides) 네 명의 저자가 공동으로 집필한 "Design Patterns: Elements of Reusable Object-Oriented Software" (1994)로, 객체 지향 소프트웨어 설계에서 재사용 가능한 23가지 패턴을 소개하여 현대 소프트웨어 공학에 큰 영향을 미쳤다.

02 AI 시스템 보안(Security) 목표

AI 시스템 보안(Security) 목표는 AI 기술을 안전하게 보호하고 신뢰할 수 있는 방식으로 운영되도록 보장하는 데 있다. 이 목표는 AI 시스템이 외부 공격, 데이터 유출, 시스템 오작동 등 다양한 보안 위협에 대응할 수 있도록 설계되고 유지되는 것을 의미한다.

- AI 시스템에서 보안의 주요 목표는 전통적인 정보보호의 3대 요소인 기밀성(Confidentiality), 무결성(Integrity), 가용성(Availability)을 기본으로 하고, AI 모델·시스템 등 개발 생애 주기에서의 책임성 확보 및 검증을 위한 Accountability(책임 추적성)를 추가하였다.

부록 그림 6 AI 개발자 대상 보안 프레임워크(Security Framework)



2.1 가용성

가용성(Availability)은 AI 보안 전략에서 중요한 요소 중 하나로, 서비스와 데이터가 항상 접근 가능하고 사용할 수 있는 상태를 유지하는 것을 목표로 한다. 이는 AI 시스템의 연속성과 신뢰성을 보장하기 위해 필수적이다. 가용성에 대한 주요 관점과 이를 유지하기 위한 방법들은 다음과 같다.

주요 관점	가용성 유지 방법
서비스 중단 방지	• AI 시스템이 항상 사용 가능하도록 보장함 • 장애 발생 시 신속한 복구 및 대응 계획을 수립함
데이터 접근성 보장	• 필요한 데이터가 언제나 접근 가능해야 함 • 데이터 손실을 방지하기 위한 백업 및 복구 계획을 마련함
시스템 성능 유지	• 높은 부하나 공격에도 시스템이 안정적으로 작동하도록 보장함 • 성능 저하를 방지하기 위한 자원 관리를 함

2.2 기밀성

기밀성(Confidentiality)은 AI 보안 전략에서 중요한 요소 중 하나로, 정보가 승인된 사람만 접근할 수 있도록 보호하는 것을 목표로 한다. 이는 데이터 유출과 같은 보안 사고를 방지하고, 민감한 정보가 보호되도록 하는 데 필수적이다. 기밀성에 대한 주요 관점과 이를 유지하기 위한 방법은 다음과 같다.

주요 관점	가용성 유지 방법
데이터 접근 제어	• 민감한 정보에 대한 접근을 승인된 사용자로 제한함 • 사용자 인증 및 권한 관리를 통해 불법 접근을 방지함
데이터 암호화	• 데이터를 저장하거나 전송할 때 암호화하여 보호함 • 데이터가 탈취되더라도 내용을 해독할 수 없도록 함
데이터 유출 방지	• 내부 및 외부로부터의 데이터 유출을 방지함 • 데이터 접근 및 전송 경로를 모니터링하여 유출 시도를 감지함

2.3 무결성

무결성(Integrity)은 AI 보안 전략의 중요한 요소 중 하나로, 데이터와 시스템이 허가되지 않은 변경 없이 정확하고 일관된 상태를 유지하는 것을 목표로 한다. 이는 AI 시스템이 신뢰할 수 있는 결과를 제공하고, 데이터의 변조나 손상을 방지하기 위해 필수적이다. 무결성에 대한 주요 관점과 이를 유지하기 위한 방법은 다음과 같다.

주요 관점	가용성 유지 방법
데이터 무결성 보장	• 데이터가 생성, 저장, 전송되는 동안 변조되지 않도록 보호함 • 데이터의 정확성과 일관성을 유지함
시스템 무결성 보장	• 시스템 구성 요소와 소프트웨어가 허가되지 않은 변경 없이 유지되도록 함 • 시스템이 예상대로 작동하도록 보장함
변경 추적 및 감사	• 데이터와 시스템의 변경내역을 기록하고 추적할 수 있도록 함 • 변경 사항을 모니터링하고, 비정상적인 변경이 감지되면 경고를 발송함

2.4 책임 추적성

책임 추적성(Accountability)은 AI 보안 전략에서 중요한 요소 중 하나로, AI 시스템의 행동과 결정에 대한 책임을 명확히 하고, 문제가 발생했을 때 원인을 추적할 수 있도록 하는 것을 목표로 한다. 이는 AI 시스템의 투명성을 높이고, 신뢰성을 보장하며, 윤리적이고 법적 기준을 준수하기 위해 필수적이다. 책임 추적성에 대한 주요 관점과 이를 유지하기 위한 방법은 다음과 같다.

주요 관점	책임 추적성 유지 방법
행동 및 결정 기록	• AI 시스템의 모든 행동과 결정을 기록하여 추적할 수 있도록 함 • 기록된 데이터를 통해 문제가 발생했을 때 원인을 분석하고 책임을 규명함
투명성	• AI 시스템의 동작 원리와 의사결정 과정을 이해할 수 있도록 설명함 • 시스템의 작동 방식과 데이터 사용에 대한 정보를 공개하여 신뢰를 구축함
책임 규명	• AI 시스템의 개발자, 운영자, 사용자 간의 역할과 책임을 명확히 함 • 문제가 발생했을 때 책임 소재를 명확히 규명하여 신속히 대응함

03 안전한 AI를 위한 요소(Mechanism)

1.1 견고성(Robustness)

- 견고성(Robustness)은 AI 보안 메커니즘에서 중요한 요소로, AI 시스템이 다양한 공격이나 예기치 않은 상황에서도 안정적이고 일관된 성능을 유지하도록 보장하는 것을 목표로 한다. 견고한 AI 시스템은 외부의 악의적 시도나 데이터 왜곡에 흔들리지 않고 신뢰할 수 있는 결과를 제공하여 사용자 보호와 시스템의 신뢰성을 높이는 역할을 한다.

주요 관점	구현 방법
내성 강화	• AI 모델이 다양한 공격 시나리오에 대응할 수 있도록 방어 메커니즘을 추가한다. • 악의적인 데이터 입력에 대한 내성을 강화하여 시스템의 신뢰성을 확보한다.
가용성 유지	• AI 시스템이 언제나 사용할 수 있도록 높은 가용성을 유지한다. • 시스템 다운타임을 최소화하여 사용자 경험을 개선한다.
지속적 검증 및 모니터링	• AI 시스템의 성능과 보안 취약점을 정기적으로 테스트하고 검증한다. • 실시간 모니터링을 통해 이상 징후를 조기에 탐지하고 대응한다.

1.2 중요정보 보호

- 중요정보(Critical Information) 보호는 AI 보안 메커니즘의 중요한 요소 중 하나로, AI 시스템이 중요 데이터를 보호하고, 승인되지 않은 접근이나 유출을 방지하는 것을 목표로 한다. 이는 사용자의 신뢰를 유지하고, 법적 및 규제 요구 사항을 준수하기 위해 필수적이다.

주요 관점	구현 방법
데이터 등급화 분류 및 관리	• 데이터 중요도에 따라 조직의 목적에 맞는 데이터 분류 기준을 수립한다. • 데이터 등급에 따라 접근 권한을 제한하고, 관리할 수 있는 사용자를 명확히 한다.
데이터 접근 제어	• 중요 데이터에 대한 접근을 엄격히 통제하고, 승인된 사용자만 접근할 수 있도록 한다. • 데이터 접근 권한을 최소한으로 제한하고, 접근 로그를 기록한다.
암호화	• 중요 데이터에 대한 접근 통제와 암호화를 적용한다. • 데이터 왜곡이나 변조 방지를 위해 안전한 데이터 저장 및 처리 방법을 도입한다.

1.3 신뢰성

- 신뢰성(Reliability)는 AI 보안 메커니즘에서 중요한 요소로, AI 시스템이 일관되고 예측 가능한 성능을 제공하며, 다양한 상황에서도 안정적으로 작동하는 것을 목표로 한다. 이는 AI 시스템이 신뢰할 수 있는 결과를 제공하고, 사용자가 의존할 수 있도록 하는 데 필수적이다.

주요 관점	구현 방법
안정성 (Stability)	• AI 시스템이 다양한 환경과 조건에서도 안정적으로 작동하도록 보장한다. • 시스템 장애나 오류가 발생하지 않도록 예방하고, 발생 시 신속히 복구한다.
무결성	• AI 시스템은 수집된 데이터나 개발된 모델이 변조되지 않도록 보장한다. • 무결성 검사를 정기적으로 수행하고, 데이터나 모델에 악의적인 변조가 있는지 모니터링한다.

1.4 사회적 영향

- 사회적 영향(Social Impact)은 AI 보안 메커니즘에서 중요한 요소로, AI 시스템이 사회에 미치는 긍정적 또는 부정적 영향을 고려하여, 긍정적 영향은 최대화하고 부정적 영향은 최소화하는 것을 목표로 한다. 이는 AI 기술이 공익적이고 책임감 있게 사용되도록 보장하고, 사회적 신뢰를 구축하기 위해 필수적이다.

주요 관점	관리 방법
사회 기반시설 보호	• 사회 기반시설에 대한 정보는 국가 및 공공의 안보와 직결되므로, 중요 데이터를 별도로 분류하고 최고 수준의 보호를 적용한다.
공공시스템 보호	• 중요한 공공시스템은 일반 네트워크와 물리적으로 분리하거나, 강력한 네트워크 분할(segmentation)을 통해 외부 접근으로부터 보호하고 다양한 기술적 조치를 통해 보안을 강화한다.
악용 방지	• AI 기술이 사회적 해악을 끼칠 목적으로 악용되지 않도록 예방하는 보안 메커니즘이 필요하다. 예를 들어, 자동화된 사이버 공격 등에 대한 방지책을 마련한다. • 시스템의 악용 가능성을 사전에 분석하고, 이러한 위험에 대한 대응책을 포함한 설계를 적용한다.

04 안전한 AI를 위한 위험 관리(Risk Management)

4.1 인적 측면(Personal Aspect)

4.1.1 사람의 감독

- 사람의 감독(Human Oversight)은 AI 관련 사업에서 중요한 요소로, AI 시스템이 의사결정 과정에서 인간의 감독과 개입을 받을 수 있도록 하는 것을 목표로 한다. 이는 AI 시스템의 신뢰성과 보안성을 보장하고, 잠재적인 위험을 최소화하기 위해 필수적이다.

주요 관점	구현 방법
인간의 개입	• AI 시스템의 결정에 대해 책임질 수 있는 인간 감독자가 존재해야 한다. • 인간 감독자는 AI 시스템의 결과와 그에 따른 행동에 대해 책임을 진다.
AI 모델의 학습 및 업데이트 감독	• AI 시스템이 새로운 데이터로 학습하거나 모델을 업데이트할 때, 사람이 그 데이터를 검토하여 악의적 데이터, 또는 보안 취약성을 식별하도록 해야 한다. • 학습 과정 중 생성된 결과나 변경 사항에 대해 정기적으로 감사(Audit)를 실시한다.

4.1.2 데이터 보호

- 데이터 보호는 AI 관련 사업에서 중요한 요소로, 데이터가 무단으로 수집되거나 사용되지 않도록 보호하는 것을 목표로 한다. 이는 사용자의 신뢰를 유지하고, 법적 및 규제 요구 사항을 준수하며, 데이터 기밀성과 무결성을 보장하기 위해 필수적이다.

주요 관점	구현 방법
데이터 암호화	• 전송/저장 중 암호화: 네트워크를 통해 전송되는 데이터를 암호화하고 데이터베이스와 스토리지 시스템에 저장된 데이터를 암호화하여 물리적 접근 및 데이터 유출 시에도 기밀성을 유지한다.
접근 제어	• 최소 권한 원칙(Principle of Least Privilege): 사용자가 최소한의 데이터에만 접근할 수 있도록 접근 권한을 제한한다. • 역할 기반 접근 제어(Role-Based Access Control, RBAC): 사용자의 직무나 역할에 따라 접근 권한을 부여한다.
사용자 인증 및 신원 확인	• 멀티팩터 인증(Multi-Factor Authentication, MFA): 비밀번호 외에 추가 인증 요소를 요구한다. • 생체 인증(Biometric Authentication): 얼굴 인식, 지문, 홍채 등 고급 인증 기술을 사용한다.

4.1.3 역할과 책임

- 역할과 책임(Role and Responsibility, R&R)은 AI 관련 사업에서 중요한 요소로, 각 개인의 역할과 책임을 명확히 정의하여 AI 시스템의 개발, 운영, 유지보수 과정에서 발생할 수 있는 혼란과 문제를 최소화하는 것을 목표로 한다. 이는 AI 시스템의 효율적이고 신뢰할 수 있는 운영을 보장하기 위해 필수적이다.

주요 관점	구현 방법
명확한 역할 정의 (Clear Role Definition)	• 각 개인의 역할과 책임을 명확히 정의하여, 중복되거나 누락되는 부분이 없도록 한다. • 역할에 대한 명확한 정의는 책임의 한계를 명확히 하고, 효율적인 업무 분담을 한다.
책임 분담 (Responsibility Allocation)	• AI 시스템의 다양한 기능과 과업에 대해 책임을 적절히 분담한다. • 책임 분담을 통해 각 팀원이 자신의 역할에 충실하고, 협력하여 목표를 달성한다.
지속적인 모니터링과 평가 (Continuous Monitoring and Evaluation)	• 각 역할과 책임이 제대로 이행되고 있는지 지속적으로 모니터링하고 평가한다. • 이를 통해 문제를 조기에 발견하고, 신속히 해결한다.

4.1.4 직무 분리

- 직무 분리(Segregation of Duties, SoD)는 AI 관련 사업에서 중요한 요소로, 특정한 직무와 책임을 여러 개인에게 분산시켜 비윤리적 행위나 실수를 방지하고, 내부 통제와 보안을 강화하는 것을 목표로 한다. 이는 AI 시스템의 신뢰성과 투명성을 보장하기 위해 필수적이다.

주요 관점	구현 방법
내부 통제 강화 (Enhanced Internal Control)	• 한 사람이 모든 권한을 가지는 것을 방지하여 내부 통제를 강화한다. • 비윤리적 행위나 실수로 인한 위험을 최소화한다.
책임성 증대 (Increased Accountability)	• 각 개인의 책임을 명확히 하여 책임성을 증대한다. • 직무와 권한이 분리됨으로써 서로를 감시하고 견제한다.
투명성 확보 (Ensured Transparency)	• 직무 분리를 통해 업무 과정의 투명성을 확보한다. • 권한 남용이나 비리의 가능성을 줄인다.

4.2 물리적/환경적 측면(Physical/Environmental Aspect)

4.2.1 사회적 책임

- 사회적 책임(Social Responsibility)은 AI 관련 사업에서 중요한 요소로, AI 시스템이 사회와 환경에 미치는 영향을 고려하고, 이를 통해 사회적 가치와 국가적 이익을 증진하는 것을 목표로 한다. 이는 AI 시스템의 개발, 운영, 유지보수 과정에서 발생할 수 있는 사회적 및 환경적 영향을 최소화하고, 지속 가능한 발전을 도모하기 위해 필수적이다.

주요 관점	구현 방법
사회적 영향 (Social Impact)	• AI 시스템이 사회에 미치는 긍정적 및 부정적 영향을 평가하고 관리한다. • 모든 사회 구성원이 공평하게 AI의 혜택을 누릴 수 있도록 한다.
환경적 지속 가능성 (Environmental Sustainability)	• AI 시스템의 개발과 운영이 환경에 미치는 영향을 최소화한다. • 지속 가능한 자원 사용과 환경 보호를 고려한다.
법적 및 윤리적 준수 (Legal and Ethical Compliance):	• AI 시스템이 관련 법규와 윤리적 기준을 준수하도록 한다. • 사용자와 사회의 신뢰를 유지하기 위해 투명하고 책임 있는 AI 개발을 추진한다.

4.2.2 물리적 피해 방지 설계

- 물리적 피해 방지 설계(Design to Prevent Physical Harm)는 AI 관련 사업에서 중요한 요소로, AI 시스템이 물리적 환경에서 작동할 때 사용자와 주변 환경에 신체적 피해를 주지 않도록 설계하는 것을 목표로 한다. 이는 AI 시스템의 안전성과 신뢰성을 보장하고, 법적 및 윤리적 책임을 다하기 위해 필수적이다.

주요 관점	구현 방법
안전 설계 (Safety by Design)	• AI 시스템이 물리적 환경에서 안전하게 작동하도록 설계한다. • 위험 요소를 사전에 식별하고 제거하여 안전성을 확보한다.
위험 평가 및 관리 (Risk Assessment and Management)	• AI 시스템의 물리적 작동이 초래할 수 있는 잠재적 위험을 평가하고 관리한다. • 위험 발생 시 신속히 대응할 수 있는 절차를 마련한다.
법적 및 규제 준수 (Legal and Regulatory Compliance)	• AI 시스템의 설계와 운영이 관련 법규와 규제 기준을 준수한다. • 안전 표준과 가이드라인을 따라 시스템을 개발하고 운영한다.

05 AI 시스템 보안을 위한 거버넌스(Governance)

AI 시스템 보안을 위한 조직 내 규정과 정책, 그리고 거버넌스 체계는 AI 모델이 조직에서 안전하게 운영되고, 보안 리스크를 관리하는 데 중요한 역할을 한다. 효과적인 AI 보안 거버넌스는 기술, 사람, 프로세스를 통합하여 AI 시스템의 무결성, 기밀성, 가용성을 유지하고, 법적 규제와 윤리적 기준을 준수하도록 설계되어야 한다. 이를 위해 조직은 명확한 정책과 절차를 수립하고, 보안 거버넌스를 통해 AI 시스템이 지속적으로 안전하게 운영될 수 있도록 관리해야 한다. AI 보안 거버넌스의 핵심 요소는 정책(Policy), 규정(Guidelines), 절차(Processes) 등을 포함한다.

5.1 정책(Policy)

- AI 보안 정책 수립: 조직은 AI 시스템을 안전하게 운영하기 위한 포괄적인 보안 정책을 수립해야 한다. 정책에는 AI 시스템의 설계, 개발, 배포 및 운영에 필요한 보안 기준과 절차를 명확히 정의해야 한다.
- 접근 권한 관리 정책: AI 시스템에 대한 접근 권한을 관리하기 위한 정책을 수립하여, 민감한 데이터와 시스템에 대한 무단 접근을 방지해야 한다. 역할 기반 접근 제어(RBAC)를 통해 AI 시스템에 접근할 수 있는 사용자와 권한을 제한해야 한다.
- 데이터 보호 정책: AI 시스템이 다루는 데이터의 기밀성과 무결성을 보장하기 위해 데이터 암호화, 데이터 저장소 보호에 대한 규정을 마련해야 한다.

5.2 규정(Guidelines)

- AI 모델 관리 규정: AI 모델의 학습, 배포 및 사용에 대한 구체적인 지침을 제공해야 한다. 이 규정은 AI 모델의 검증 및 테스트 절차, 안전한 데이터 사용, 모델 업데이트 및 폐기 절차를 포함한다.
- AI 시스템 모니터링 규정: AI 시스템이 보안 위협에 대응할 수 있도록 실시간 모니터링 및 로그 분석 규정을 마련해야 한다. 이를 통해 시스템 내 이상 활동이나 보안 위협을 조기에 감지하고 대응할 수 있어야 한다.
- 리스크 관리 규정: AI 시스템에 내재된 보안 리스크를 식별하고 관리하기 위한 규정을 수립해야 한다. 이는 잠재적인 보안 위협을 사전에 평가하고 완화하는 절차를 포함한다.

5.3 절차(Process)

- AI 보안 평가 절차: AI 시스템 개발 초기부터 보안 평가 절차를 구축하여, 보안 위험을 사전에 감지하고 해결해야 한다. 이 절차는 보안 취약점 분석, 침투 테스트 및 코드 리뷰를 포함할 수 있다.
- AI 시스템 업데이트 및 패치 절차: AI 시스템과 모델이 최신 보안 위협에 대응할 수 있도록 정기적인 업데이트 및 패치 절차를 수립해야 한다. 이 절차는 새로운 보안 위협에 대한 대응 방안을 포함하며, 보안 업데이트가 원활하게 이루어지도록 해야 한다.
- 비상 대응 계획: 보안 사고가 발생했을 때 즉각적으로 대응할 수 있는 비상 대응 계획을 마련해야 한다. 여기에는 사고 보고, 대응 팀 구성, 시스템 복구 및 데이터 손실 방지 절차가 포함된다.

06 AI 시스템 보안를 위한 위험관리(Risk Management)

위험관리(Risk Management)는 AI를 활용한 서비스에서 필수적인 요소로, AI 시스템이 직면할 수 있는 잠재적 위험을 식별하고, 이를 효과적으로 관리하고 대응하기 위한 체계적인 접근을 포함한다. 위험 관리는 위험 분석, 위험 감지, 사고 대응의 세 가지 주요 단계로 나눌 수 있다.

6.1 위험 분석

- 위험 분석(Risk Analysis)은 AI 시스템이 직면할 수 있는 잠재적 위험을 식별하고, 그 심각성과 발생 가능성을 평가하는 과정이다. 이 과정은 위험 관리의 첫 단계로, 체계적인 접근을 통해 위험 요소를 사전에 파악하고 대응 전략을 마련하는 것을 목표로 한다.
- 주요 활동
  - 위험 식별(Risk Identification): AI 시스템과 관련된 모든 잠재적 위험을 식별하고, 기술적·운영적·윤리적·법적 측면에서 발생할 수 있는 다양한 위험을 고려한다.
    ※ 예시: 데이터 손실, 시스템 오류, 보안 침해, 윤리적 문제 등
  - 위험 평가(Risk Assessment): 식별된 위험의 심각성과 발생 가능성을 평가하고, 각 위험 요소의 영향을 분석하여 우선순위를 정한다.
  - 위험 대응 계획(Risk Mitigation Plan) 수립: 평가된 위험에 대한 대응 계획을 수립하고, 위험을 줄이기 위한 예방 조치와 대응 전략을 마련한다.
    ※ 예시: 보안 강화, 데이터 백업, 긴급 대응 절차 마련 등

6.2 위험 감지

- 위험 감지(Risk Detection)는 AI 시스템의 운영 중 발생하는 이상 징후나 위험 요소를 실시간으로 감지하고, 이를 신속하게 보고하는 과정이다. 이는 위험 발생 시 신속한 대응을 위해 중요한 단계이다.
- 주요 활동
  - 모니터링 시스템 구축(Establish Monitoring Systems): AI 시스템의 성능과 안전성을 실시간으로 모니터링하는 시스템을 구축한다.
    ※ 예시: 네트워크 트래픽 모니터링, 시스템 로그 분석, 사용자 활동 추적 등
  - 이상 징후 탐지(Anomaly Detection) : 정상적인 패턴에서 벗어난 이상 징후를 자동으로 감지할 수 있는 알고리즘과 기술을 도입한다.
    ※ 예시: 머신러닝 기반 이상 탐지 알고리즘, 실시간 경고 시스템 등

- 자동화된 경고(Automated Alerts): 이상 징후나 위험 요소 감지 시 즉각적으로 관련 담당자에게 경고를 보내는 자동화된 시스템을 운영한다.

  ※ 예시: 이메일 알림, SMS 경고, 대시보드 알림 등

- 정기적 검토(Regular Reviews): 모니터링 결과와 경고 로그를 정기적으로 검토하여, 새로운 위험 요소나 패턴을 식별하고 대응 방안을 업데이트한다.

  ※ 예시: 주간/월간 보고서 작성, 경고 로그 분석 회의 등

6.3 사고 대응

- 사고 대응(Incident Handling)은 실제로 위험이 발생했을 때 이를 효과적으로 대응하고 해결하는 과정이다. 이는 신속한 대응과 문제 해결을 통해 피해를 최소화하고, 재발 방지를 위한 교훈을 도출하는 것을 목표로 한다.
- 주요 활동
  - 사고 대응 절차 수립(Establish Incident Response Procedures): 사고 발생 시 신속하고 체계적으로 대응할 수 있는 절차와 계획을 마련한다.

    ※ 예시: 사고 대응 매뉴얼 작성, 비상 연락망 구축, 역할 및 책임 정의 등
  - 사고 대응 팀 구성(Form Incident Response Team) : 사고 발생 시 대응할 전담 팀을 구성하고, 각 팀원의 역할과 책임을 명확히 한다.

    ※ 예시: 보안 담당자, 데이터 과학자, IT 지원팀 등으로 구성
  - 초기 대응 및 완화(Initial Response and Mitigation) : 사고 발생 시 초기 대응을 통해 피해를 최소화하고, 추가 피해를 방지하기 위한 조치를 취한다.

    ※ 예시: 시스템 격리, 데이터 복구, 보안 패치 적용 등
  - 사고 분석 및 보고(Incident Analysis and Reporting) : 사고의 원인을 분석하고, 사고 발생 과정과 대응 결과를 상세히 기록하여 보고한다.

    ※ 예시: 사고 원인 분석 보고서 작성, 대응 결과 리뷰 등
  - 사후 조치 및 재발 방지(Post-Incident Actions and Prevention) : 사고 종료 후 사후 조치를 취하고, 재발 방지를 위한 교훈을 도출하여 시스템과 절차를 개선한다.

    ※ 예시: 시스템 업데이트, 보안 정책 강화, 교육 프로그램 운영 등
- 예시 시나리오
  - AI 모델 오류 사고 대응
    - ▸위험 분석: AI 모델의 예측 오류 가능성을 평가하고, 모델 검증 및 테스트 계획 수립
    - ▸위험 감지: AI 모델의 실시간 성능 모니터링 시스템 도입. 예측 오류 발생 시 경고 시스템 운영
    - ▸사고 대응: AI 모델의 예측 오류 발생 시 즉시 모델 사용 중지 및 수정 작업. 오류 원인 분석 및 보고. 수정된 모델의 검증 및 테스트 후 재배포. 재발 방지를 위한 모델 검증 절차 강화

06 AI 서비스 제공자 대상 보안 프레임워크

01 AI 서비스 제공자 관점의 프레임워크 필요성

- AI 서비스를 위한 정보보안은 다양한 새로운 기술 요소로 구성되어 있어 이를 통합하고 관리하는 것은 매우 복잡한 작업이므로, 이러한 복잡성을 효과적으로 관리하고 시스템에 대한 안전성과 상호 운용성을 높이기 위해 AI 시스템과 서비스에 대한 정보보안 프레임워크는 필요하다.
- AI 서비스를 위한 정보보안 프레임워크는 전략적·관리적·정책적·기술적 및 메커니즘 측면에서 AI 환경 내 발생 할 수 있는 침해 문제를 해결하고 이를 통해 서비스 제공자는 AI 기술을 효과적으로 도입하여 이용자에게 더욱 안전한 AI 서비스 환경을 제공할 수 있다.
  - (데이터 보호 및 기밀성 유지) AI 서비스는 대량의 데이터를 처리하며, 이 중에는 중요 정보가 포함될 수 있다. AI 서비스 제공자는 정보보안 프레임워크를 통해 데이터의 기밀성을 유지하고 불법 접근이나 유출을 방지할 수 있다.
  - (AI 모델의 무결성 유지) AI 모델이나 서비스가 사이버 공격(예: 데이터 조작, 모델 해킹 등)에 노출되면 결과의 정확성과 무결성이 위협받을 수 있으므로, AI 서비스 제공자는 보안 프레임워크를 통해 AI 시스템이 악의적 변경이나 조작으로부터 보호되도록 예방할 수 있다.
  - (사이버 위협 대응 및 리스크 최소화) AI 서비스 제공자는 사이버 공격, 해킹, 랜섬웨어 등 다양한 위협에 노출될 수 있으므로, 보안 프레임워크를 통해 사전 대비와 함께, 발생 시 대응 프로토콜을 제공하여 피해를 최소화할 수 있다.
  - (서비스 연속성 및 가용성 보장) AI 서비스의 중단은 기업의 신뢰와 수익에 큰 타격을 줄 수 있으므로, 보안 프레임워크를 통해 백업, 복구 계획 등을 포함하여 AI 서비스의 안정적인 가용성을 보장할 수 있다.
  - (고객 신뢰 확보) 고객들은 자신의 데이터를 다루는 AI 기업이 신뢰할 수 있어야 서비스를 사용하므로, AI 서비스 제공자는 보안 프레임워크를 통해 고객의 데이터 보호 및 신뢰를 높이고, 비즈니스 지속성을 강화할 수 있다.

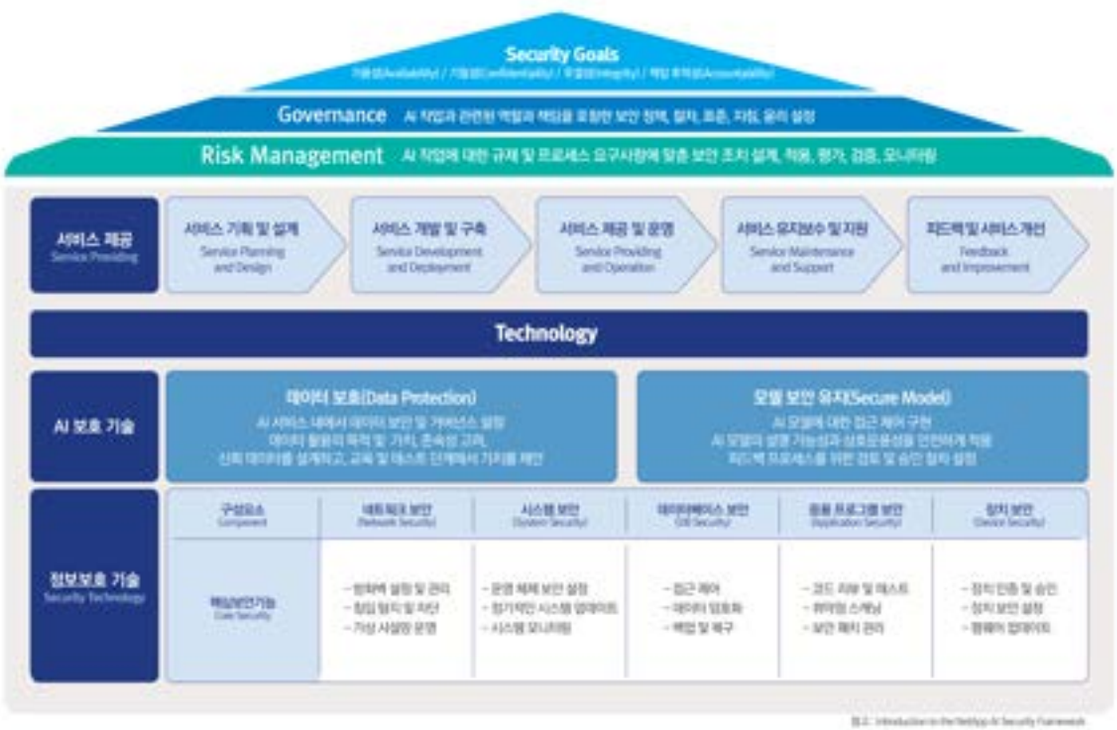
02 AI 서비스 제공자 관점의 보안 목표

2.1 보안 프레임워크(Security Framework) - 예방 단계

예방(Prevention) 단계에서의 보안 목표: 예방 단계의 Framework는 AI 시스템에 대한 잠재적 위협을 사전에 대비하고 보호하기 위한 다양한 전략과 메커니즘을 포함한다.

- **거버넌스(Governance)**: 보안 정책, 절차, 표준, 가이드라인 및 윤리를 수립하여 AI 워크로드와 관련된 역할과 책임을 명확히 한다.
- **위험 관리(Risk Management)**: AI 시스템과 서비스에 대한 보안 요구사항을 준수하기 위해 보안 조치를 설계, 적용, 평가, 검증한다.

부록 그림 7 AI 서비스 제공자 대상 보안 프레임워크(Security Framework) - 예방 단계



예방 단계에 대한 보안 기술(Security Technology) 적용 방안

데이터 보호

- 데이터 보호(Data Protection)는 AI 시스템에서 데이터의 기밀성, 무결성, 가용성을 유지하여 보안성을 확보하는 중요한 과정이다. 이는 데이터가 불법적으로 접근되거나 수정되지 않도록 하며, 시스템의 신뢰성을 높이는 역할을 한다.

구 분	내 용	추진방안
설정 및 설계 (Setup and Design)	데이터 보호를 위한 보안 정책과 시스템을 설정하고 설계	• 데이터 보안 정책 수립: 데이터를 보호하기 위한 정책과 절차를 수립한다. 이는 데이터 접근 권한, 저장, 전송, 삭제 등의 과정에서의 보안 요구사항을 포함한다. • 거버넌스 체계 구축: 데이터 관리와 보호를 위한 거버넌스 체계를 구축하여 책임과 역할을 명확히 한다. 이를 통해 데이터 보호가 조직 전반에 걸쳐 일관되게 적용될 수 있도록 한다. • 보안 아키텍처 설계: 데이터 보호를 위한 보안 아키텍처를 설계한다. 이는 데이터 암호화, 접근 통제, 로그 관리 등의 기술적 요구사항을 포함한다. • (예시) 데이터 암호화 정책 수립, 데이터 접근 통제 설계, 데이터 저장 및 백업 절차 설정.
단계 설정 (Stage Setup)	데이터 보호를 위한 구체적인 단계와 절차를 설정	• 신뢰할 수 있는 데이터 구축: 데이터 수집, 저장, 처리 과정에서 데이터의 신뢰성을 보장할 수 있는 절차를 마련한다. 이는 데이터의 정확성, 일관성, 무결성을 포함한다. • 훈련 및 테스트 단계 설계: AI 모델의 훈련 및 테스트 단계에서 데이터 보호를 위한 절차를 마련하여 데이터가 안전하게 사용될 수 있도록 한다. • 보안 검토 및 업데이트: 데이터 보호 절차가 최신 보안 표준과 법적 요구사항을 준수하도록 정기적으로 검토하고 업데이트한다. • (예시) 데이터 검증 절차 설정, 훈련 데이터셋의 무결성 검증, 테스트 데이터의 보호 및 관리.

● 모델 보안 유지

- 모델 보안 유지(Secure Model)는 AI 시스템에서 모델의 보안을 보장하기 위한 다양한 활동과 절차를 포함한다. 이를 통해 AI 모델이 안전하게 작동하고 신뢰성을 유지하도록 한다. 여기에는 다음과 같은 구체적인 활동이 포함된다:

구 분	내 용	추진방안
모델 접근 통계 (Implement Access Control on Model)	AI 모델에 대한 접근 권한을 통계하여 무단 접근을 방지하고 보안을 강화하는 과정	• 사용자 인증 및 권한 부여: AI 모델에 접근할 수 있는 사용자나 시스템을 인증하고, 적절한 권한을 부여한다. • 접근 로그 관리: AI 모델에 대한 접근 시도를 기록하고, 이상 활동을 모니터링하여 보안 위협을 식별한다. • 역할 기반 접근 통계(Role-Based Access Control): 사용자 역할에 따라 접근 권한을 제한하여 민감한 모델에 대한 무단 접근을 방지한다. • (예시) AI 모델에 대한 접근 시 다단계 인증을 요구하고, 접근 로그를 주기적으로 검토하여 비정상적인 접근을 탐지한다.
모델 설명 가능성 및 상호 운용성 적용 (Apply Model Explainability and Interoperability with Secure Manner)	AI 모델의 결정 과정을 투명하게 설명하고, 다른 시스템과의 상호 운용성을 보장하는 동시에 보안을 유지하는 과정	• 설명 가능 AI(Explainable AI): AI 모델의 예측이나 결정이 어떻게 도출되었는지 설명할 수 있는 기능을 추가하여 투명성을 높인다. • 상호 운용성 보장: AI 모델이 다른 시스템과 원활하게 연동될 수 있도록 설계하며, 데이터 교환 시 보안을 유지한다. • 보안 모니터링: 모델 설명 과정에서 민감한 정보가 노출되지 않도록 보안 모니터링을 실시한다. • (예시) AI 모델의 예측 결과를 설명하는 대시보드를 제공하고, 다른 시스템과의 데이터 교환 시 암호화를 적용하여 보안을 유지한다.
리뷰 및 승인 프로세스 설정 (Set Up Review and Approval Process for Feedback Process)	AI 모델의 변경 사항을 검토하고 승인하는 절차를 마련하여, 모델의 신뢰성과 보안을 보장하는 과정	• 변경 사항 검토: AI 모델의 업데이트나 수정 사항을 사전에 검토하고, 잠재적인 보안 위험을 평가한다. • 승인 절차 마련: 변경 사항을 적용하기 전에 승인 절차를 거쳐, 보안 및 품질 기준을 충족하도록 한다. • 피드백 반영: 사용자 및 시스템에서 제공된 피드백을 기반으로 모델을 개선하고, 보안성을 유지한다. • (예시) AI 모델 업데이트 시 전문가 리뷰를 통해 변경 사항을 검토하고, 필요시 보안 패치를 적용하여 보안성을 강화한다.

● **사이버 보안 기술 적용**

- AI Security Framework에서 사이버 보안(Cyber Security)은 AI 시스템 및 데이터의 보안을 강화하는 데 중점을 둔다. 이는 다양한 보안 기술과 절차를 통해 AI 시스템을 보호하고, 데이터 무결성과 기밀성을 유지하며, 시스템의 가용성을 보장하는 것을 목표로 한다.

  ▸ **구성요소(Component)별 목표 및 요구사항은 다음과 같다**

구분	보안 목표 및 요구사항
Application Security (응용 프로그램 보안)	[목표] AI 애플리케이션의 보안을 강화하여 취약점 및 공격으로부터 보호한다. • 코드 리뷰 및 테스트: 애플리케이션의 코드를 주기적으로 리뷰하고, 보안 취약점을 발견하여 수정한다. • 취약점 스캐닝: 애플리케이션 내의 보안 취약점을 탐지하고, 이를 해결하기 위한 조치를 취한다. • 보안 패치 관리: 애플리케이션에 대한 최신 보안 패치를 적용하여 알려진 취약점을 방지한다. • (예시) 정기적인 취약점 스캐닝 도구를 사용하여 AI 애플리케이션의 보안 상태를 점검하고, 발견된 취약점을 신속히 수정한다.
Network Security (네트워크 보안)	[목표] AI 시스템이 연결된 네트워크를 보호하여 데이터의 무결성과 기밀성을 유지한다. • 방화벽 설정 및 관리: 외부 공격으로부터 네트워크를 보호하기 위해 방화벽을 설정하고 관리한다. • 침입 탐지 시스템(IDS): 네트워크 트래픽을 모니터링하여 비정상적인 활동을 탐지하고 대응한다. • 가상 사설망(VPN): 네트워크를 통한 데이터 전송 시 암호화된 연결을 제공하여 데이터의 기밀성을 유지한다. • (예시) 네트워크에 IDS를 설치하여 실시간으로 비정상적인 트래픽을 감지하고, 즉시 대응할 수 있도록 한다.
System Security (시스템 보안)	[목표] AI 시스템의 운영 체제 및 관련 인프라를 보호하여 무단 접근과 공격을 방지한다. • 운영 체제 보안 설정: 시스템의 운영 체제에 대한 보안 설정을 강화하여 취약점을 줄인다. • 정기적인 시스템 업데이트: 운영 체제 및 소프트웨어에 대한 최신 업데이트를 적용하여 보안을 유지한다. • 시스템 모니터링: 시스템 로그를 주기적으로 모니터링하여 비정상적인 활동을 탐지하고 대응한다. • (예시) 시스템에 최신 보안 업데이트를 정기적으로 적용하고, 로그 모니터링 도구를 사용하여 시스템 활동을 실시간으로 감시한다.
DB Security (데이터베이스 보안)	[목표] AI 시스템에서 사용되는 데이터베이스를 보호하여 데이터 무결성과 기밀성을 유지한다. • 접근 통제: 데이터베이스에 대한 접근 권한을 관리하여 무단 접근을 방지한다. • 데이터 암호화: 저장된 데이터를 암호화하여 데이터 유출 시에도 기밀성을 유지한다. • 백업 및 복구 계획: 데이터베이스의 정기적인 백업을 수행하고, 데이터 손실 시 복구 계획을 마련한다. • (예시) 중요한 데이터베이스 필드를 암호화하고, 정기적으로 백업을 수행하여 데이터 손실에 대비한다.
Device Security (장치 보안)	[목표] AI 시스템에 연결된 장치를 보호하여 보안 위협으로부터 안전하게 유지한다. • 장치 인증 및 승인: AI 시스템에 연결된 모든 장치에 대한 인증 절차를 마련하고, 승인된 장치만 연결되도록 한다. • 장치 보안 설정: 각 장치에 대한 보안 설정을 강화하여 취약점을 최소화한다. • 펌웨어 업데이트: 장치의 펌웨어를 최신 상태로 유지하여 보안 취약점을 해결한다. • (예시) AI 시스템에 연결된 IoT 장치에 대해 정기적인 펌웨어 업데이트를 실시하고, 보안 인증 절차를 적용하여 무단 장치 연결을 방지한다.

▸ 핵심 보안기술(Core Security)별 목표 및 요구사항

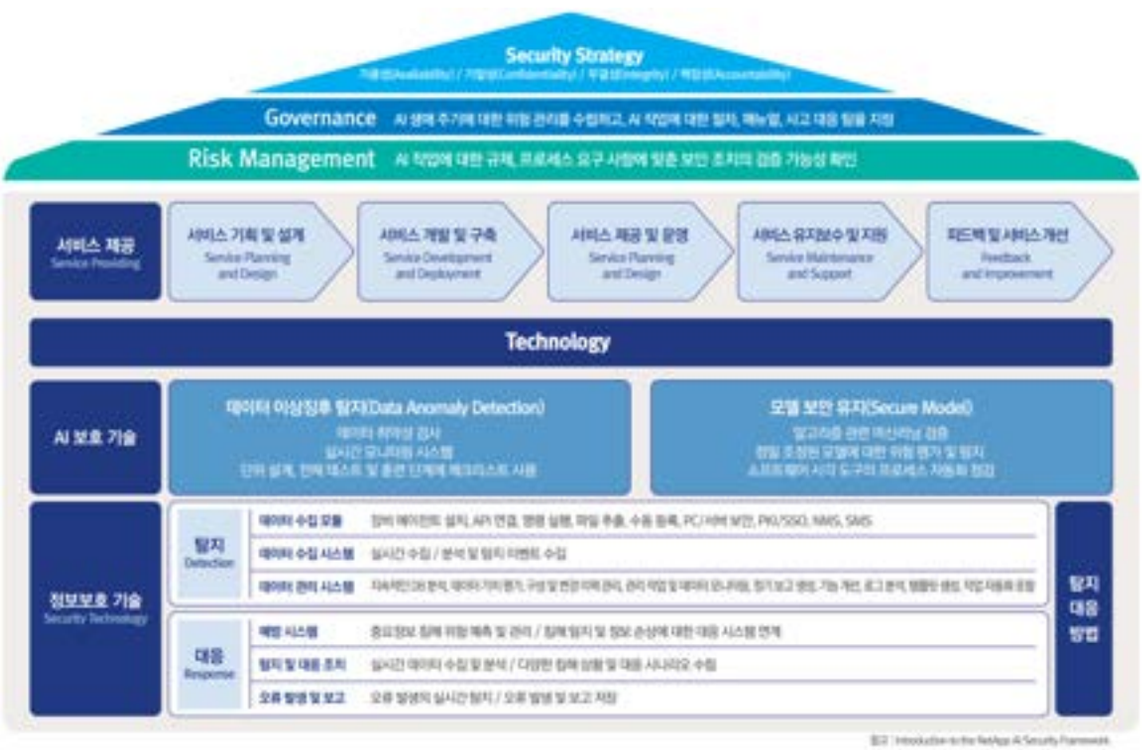
구분	보안 목표 및 요구사항
Multi-Factor Authentication (MFA)	[목표] Multi-Factor Authentication은 두 개 이상의 인증 요소를 사용하는 보안 체계이다. 이는 사용자나 시스템이 본인임을 확인하기 위해 여러 단계를 거치는 방식으로, 보안 강화를 목표로 한다. • 지식 기반 요소: 사용자만 알고 있는 정보(예: 비밀번호, PIN) • 소유 기반 요소: 사용자만 소유한 물리적 장치(예: 스마트폰, OTP 토큰) • 고유 기반 요소: 사용자의 생체 정보(예: 지문, 얼굴 인식, 음성 인식) • OTP(One-Time Password): 일회용 비밀번호를 생성하여 로그인 시 추가 인증 단계로 사용 • 생체인식: 지문 스캐너, 얼굴 인식 시스템을 통해 물리적 접근 및 데이터 접근 제어 • 보안 토큰: 하드웨어나 소프트웨어 토큰을 사용하여 두 번째 인증 단계 제공 • (예시) 은행 거래 시 스마트폰에 OTP를 생성하여 입력하거나, 회사 시스템 로그인 시 지문 인식을 사용한다.
Access Control (접근 통계)	[목표] 접근 통계는 시스템 내에서 사용자가 접근할 수 있는 자원과 권한을 관리하는 메커니즘이다. 이는 무단 접근을 방지하고 데이터 기밀성과 무결성을 유지하는 데 중요한 역할을 한다. • Mandatory Access Control(MAC): 시스템 관리자에 의해 설정된 정책에 따라 접근이 통제되며, 사용자는 이를 변경할 수 없다. 주로 높은 보안이 요구되는 군사나 정부 기관에서 사용한다. • Discretionary Access Control(DAC): 데이터 소유자가 누구에게 접근 권한을 부여할지 결정하는 방식으로, 유연성이 높다. • Role-Based Access Control(RBAC): 사용자의 역할에 따라 접근 권한을 부여하는 방식으로, 대규모 조직에서 효율적이다. • 파일 시스템 보안: 특정 사용자나 그룹에 파일 읽기/쓰기 권한을 설정한다. • 네트워크 보안: 네트워크 장치와 서버에 대한 접근 권한을 역할에 따라 설정한다. • 데이터베이스 보안: DBMS에서 사용자 역할에 따라 테이블이나 행 단위의 접근 권한을 관리한다. • (예시) 직원이 회사의 기밀 문서에 접근할 때, 해당 직원의 역할에 따라 접근 권한을 설정하여 불필요한 접근을 방지한다.
Cryptography (암호화)	[목표] 암호화는 데이터를 보호하기 위해 정보를 특정 알고리즘을 사용해 암호화하고, 인가된 사용자만이 이를 해독할 수 있게 하는 기술이다. • 양자 암호화(Quantum Cryptography): 양자 역학의 원리를 이용한 암호화 방식으로, 매우 높은 보안성을 제공한다. • 키 관리(Key Management): 암호화 키의 생성, 배포, 저장, 교체, 폐기를 관리하여 보안성을 유지한다. • PKI(Public Key Infrastructure): 공개 키 암호화 기술을 사용하여 안전한 통신을 보장하고, 디지털서명 등을 통해 데이터의 무결성을 검증한다. • 디지털서명(Digital Signature): 전자문서나 메시지의 출처를 검증하고, 변경 여부를 확인하기 위한 기술을 사용한다. • 데이터 전송 보안: SSL/TLS를 사용하여 웹 브라우저와 서버 간의 데이터 전송을 암호화한다. • 파일 암호화: 중요한 문서 파일을 암호화하여 무단 접근을 방지한다. • 전자상거래: 거래 정보와 결제 정보를 암호화하여 안전하게 처리한다. • (예시) 이메일 전송 시 PGP를 사용하여 메시지를 암호화하고, 디지털서명을 통해 발신자의 신원을 확인한다.

2.2 보안 프레임워크(Security Framework) - 탐지·대응 단계

탐지·대응 단계에서의 Security Framework 목표: 탐지·대응(Detection) 단계 Framework는 사업자 관점에서 AI 시스템의 보안 위협을 탐지하기 위한 전략과 메커니즘을 설명한다. 탐지(Detection) 단계는 AI 시스템에서 발생할 수 있는 보안 위협을 실시간으로 모니터링하고, 이를 신속하게 식별하여 대응할 수 있도록 하는 데 중점을 둔다.

- **거버넌스(Governance)**: AI 생명주기 전반에 대한 위험 관리를 수립하고, AI 워크로드에 대한 탐지 절차, 매뉴얼, 사고 대응 팀을 배정한다.
- **위험 관리(Risk Management)**: AI 시스템과 서비스에 대한 보안 요구사항을 준수하기 위해 보안 조치를 적용한다.

부록 그림 8 AI 서비스 제공자 대상 보안 프레임워크(Security Framework) - 탐지·예방 단계



탐지·대응 단계에 대한 보안 기술(Security Technology) 적용 방안

데이터 이상 징후 탐지

- 데이터 이상징후 탐지(Data Anomaly Detection)는 AI 시스템의 데이터를 정기적으로 점검하고, 실시간 모니터링 시스템을 통해 이상 징후를 감지하며, 체크리스트를 활용하여 테스트와 훈련 단계를 점검하는 과정이다. 이를 통해 데이터의 무결성, 기밀성, 가용성을 보장하고, AI 시스템의 안정성과 신뢰성을 유지할 수 있다.

구 분	내 용	추진방안
Data Vulnerability Check (데이터 취약성 점검)	AI 시스템의 데이터 취약성을 점검하여 보안 취약점을 발견하고 수정하는 과정	• 데이터 저장소, 데이터 전송 경로, 데이터 접근 권한 등을 점검하여 잠재적인 보안 취약점을 식별하고, 이를 해결하기 위한 보안 조치를 시행한다. • (예시) 데이터베이스의 취약성 스캔, 데이터 암호화 여부 점검, 접근 통계 정책 검토
Real-Time Monitoring System (실시간 모니터링 시스템)	실시간으로 AI 시스템의 데이터를 모니터링하여 이상 징후를 탐지하는 시스템	• 실시간 데이터 흐름을 분석하여 비정상적인 활동을 탐지하고, 이를 경고하거나 자동으로 대응하는 시스템을 구축한다. 이러한 모니터링 시스템은 AI 모델의 예측 결과와 실제 데이터 간의 불일치를 탐지할 수 있다. • (예시) 실시간 네트워크 트래픽 모니터링, 사용자 활동 모니터링, 시스템 로그 분석
Design to Unit, Entire Testing and Training Phased with Checklist (체크리스트를 활용한 단위, 전체 테스트 및 훈련 단계 설계)	AI 시스템의 각 구성 요소에 대해 테스트와 훈련 단계를 설계하고, 이를 체크리스트를 통해 검증하는 과정	• 시스템의 각 구성 요소별로 테스트와 훈련 계획을 수립하고, 이를 체크리스트 형태로 문서화하여 각 단계에서 수행해야 할 검증 항목들을 명확히 정의한다. 이를 통해 체계적인 검증과 훈련이 이루어지도록 한다. • (예시) 모델의 학습 데이터 검증, 모델 업데이트 후 테스트 계획 수립, 시스템 통합 테스트 계획 수립 및 체크리스트 작성

● **모델 보안 유지**

- 모델 보안 유지는 AI 모델이 안전하게 동작할 수 있도록 알고리즘 검증, 위험 평가 및 모델 조정, 소프트웨어 시각화 도구를 통한 프로세스 자동화 등의 과정을 포함한다. 이를 통해 AI 모델의 신뢰성을 높이고, 외부 위협으로부터 보호할 수 있다.

구 분	내 용	추진방안
Verified to Algorithm Related Machine Learning (알고리즘 관련 머신러닝 검증)	AI 모델이 사용하는 알고리즘이 안전하고 신뢰할 수 있도록 검증하는 과정	• 모델의 학습 알고리즘을 검토하고, 보안 표준을 충족하는지 확인한다. 알고리즘의 동작 방식과 결과를 분석하여 예상치 못한 오류나 보안 취약점이 없는지 점검한다. • (예시) 알고리즘 코드 리뷰, 알고리즘의 보안 표준 준수 여부 검토, 알고리즘의 동작 검증 테스트
Risk Evaluation and Detect to Fine-tuned Model (위험 평가 및 세밀 조정된 모델 탐지)	AI 모델의 잠재적인 위험을 평가하고, 이를 미세 조정하여 보안성을 강화하는 과정	• 모델의 예측 결과를 분석하여 잠재적인 위험 요소를 식별하고, 이를 기반으로 모델을 조정하여 보안성을 향상시킨다. 위험 평가에는 데이터 입력, 모델의 예측 과정, 예측 결과 등이 포함된다. • (예시) 모델의 예측 정확도 평가, 예측 결과의 이상 탐지, 모델 파라미터 조정 및 튜닝
Process Automation for Software Visual Tools (소프트웨어 시각화 도구를 위한 프로세스 자동화)	소프트웨어 시각화 도구를 사용하여 모델의 동작과 보안 상태를 시각적으로 모니터링하고, 이를 자동화하는 과정	• 모델의 동작 상태와 보안 상태를 실시간으로 모니터링할 수 있는 시각화 도구를 구축하고, 이를 통해 자동으로 보안 상태를 점검하고 보고하는 시스템을 운영한다. • (예시) 실시간 보안 대시보드 구축, 모델 동작 모니터링 시각화 도구 개발, 보안 상태 자동 보고 시스템

● 사이버 보안(Cyber Security) 기술 적용

- 탐지 관점에서 Cyber Security는 다양한 영역에서 보안 위협을 실시간으로 탐지하고 대응하는 것을 목표로 한다. AI Security Framework 탐지 및 대응 부문에서 Cyber Security는 애플리케이션, 네트워크, 시스템, 데이터베이스, 장치 등의 다양한 계층에서 보안을 강화하며, 다중 인증, 접근 통제, 암호화와 같은 최신 보안 기술을 적용한다. 이를 통해 AI 시스템이 안전하게 운영되고, 데이터 무결성 및 기밀성이 유지되며, 시스템 가용성이 보장될 수 있도록 한다. 또한, 실모니터링 및 침해 대응 절차를 통해 잠재적인 위협을 빠르게 식별하고 대응한다.

▸탐지·대응 부문별 목표 및 요구사항은 다음과 같다.

구분	목표 및 요구사항
Data Collection Modules (데이터 수집 모듈)	[목표] 다양한 데이터를 수집하기 위한 장치와 방법을 추진한다. • 장비 에이전트 설치, API 연결, 명령 실행, 추출 파일, 수동 등록, PC/서버 보안, PKI/SSO, NMS, SMS 등을 통해 데이터를 수집한다. • (예시) 네트워크 장비에 에이전트를 설치하여 데이터 수집, API를 통해 외부 시스템과 연결, 서버의 보안 로그를 수집한다.
Data Collection System (데이터 수집 시스템)	[목표] 실시간 데이터 수집 및 분석, 이벤트 감지를 담당한다. • 실시간으로 데이터를 수집하고, 이를 분석하여 보안 이벤트를 감지한다. • (예시) 실시간 로그 수집 시스템, 네트워크 트래픽 분석 시스템, 보안 이벤트 모니터링 시스템
Data Management System (데이터 관리 시스템)	[목표] 데이터베이스 분석, 데이터 평가, 설정 및 변경 이력 관리 등을 포함한 데이터 관리 기능을 제공한다. • 지속적인 DB 분석, 데이터 가치 평가, 설정 변경 이력 관리, 관리 작업 및 데이터 모니터링, 정기 보고서 작성, 기능 개선, 로그 분석, 템플릿 생성, 작업 자동화 등을 수행한다. • (예시) 데이터베이스 관리 시스템(DBMS), 로그 분석 도구, 자동화된 데이터 모니터링 및 보고 시스템

▸핵심 보안기술(Core Security)별 목표 및 요구사항

구분	목표 및 요구사항
Prevention System (예방 시스템)	[목표] 침해 위험 예측 및 관리, 실시간 침해 및 정보 손상 감지, 대응 시스템 연계 등을 담당한다. • 침해 위험을 예측하고 관리하며, 실시간으로 침해 및 정보 손상을 감지하고 대응 시스템과 연계한다. • (예시) 개인정보 보호 시스템, 실시간 침해 감지 시스템, 침해 대응 연계 시스템
Detection and Response Measures (탐지 및 대응 조치)	[목표] 실시간 데이터 수집 및 분석, 다양한 침해 상황 및 대응 시나리오 수립을 포함한다. • 실시간으로 데이터를 수집하고 분석하며, 다양한 침해 상황에 대한 시나리오를 수립하고 대응 조치를 마련한다. • (예시) 침해 대응 시나리오, 실시간 데이터 분석 시스템, 침해 대응 절차 수립
Error Occurrence and Reporting (오류 발생 및 보고)	[목표] 실시간 오류 감지, 오류 발생 및 보고를 담당한다. • 실시간으로 오류를 감지하고, 발생한 오류를 저장 및 보고한다. • (예시) 오류 감지 시스템, 오류 보고 시스템, 실시간 오류 모니터링 도구

집필진

- 상명대학교 유진호 교수
- 한신대학교 홍승필 교수
- 개인정보보호협회 정상호 부장
- 상명대학교 김민정 교수
- 과학기술정보통신부 정보보호기획과
- 한국인터넷진흥원(KISA)
  - 황보성 본부장, 이익섭 실장, 김성훈 팀장, 김관영 선임연구원

자문반

- 중앙대학교 이기혁 교수
- 연세대학교 권태경 교수
- 제이앤시큐리티 김경하 대표
- 카카오모빌리티 김정민 실장

인공지능(AI) 보안 안내서

인 쇄 2025년 12월

발 행 2025년 12월

발 행 처 한국인터넷진흥원(KISA, Korea Internet & Security Agency) 전라남도 나주시 진흥길 9
Tel: 1544-5118

편집·제작 아람에디트