생성형 인공지능(AI) 개발·활용을 위한

개인정보 처리 안내서

2025. 8.

개인정보보호위원회

목 차

Ⅰ. 개 요

1 발간 배경 및 목적

최근 **생성형 인공지능(AI) 기술**이 **급속히 발전**하면서 일상과 사회 전반으로 **빠르게 확산**되고 있습니다. 특히, 민간과 공공에서는 **도메인 특성, IT 환경, 보유자원** 등에 따라 **AI 모델을 직접 개발**하거나 **기성 모델을 재학습·가공**하는 등 **다양한 활용 사례**가 늘어나고 있습니다.

생성형 AI의 **활용 양상**과 **범위가 확장**되면서, 그 과정에서 필수적으로 수반되는 **대규모 데이터 처리**로 인해 **개인정보 처리 및 보호 관점**에서 다양한 **법적·기술적 이슈**가 제기되고 있습니다. 특히, **인터넷 공간, IoT 기기·센서** 등 다양한 출처에서 **수집된 개인정보**를 **AI 학습**에 **적법하게 이용할 수 있는 기준**은 무엇인지, **고도로 복잡한 AI 환경**에서 **개인정보 안전관리 체계**를 어떻게 마련할지, 이를 뒷받침하는 **내부 거버넌스 구축 전략**을 어떻게 수립할지 등이 **주요 논의 과제**로 대두되고 있습니다.

이에, 개인정보 보호위원회(이하, '위원회')는 **AI 심화 시대**에 대응하여 개인정보 처리와 관련한 **현장의 법적 불확실성을 해소**하고, **AI 신기술·신서비스 개발**에 **개인정보가 안전하게 활용**될 수 있도록 지원하고 있습니다. 특히, 위원회는 **분야별 AI 데이터 처리 안내서** 마련[1], **사전실태점검 및 조사·처분 사례** 축적, 규제샌드박스 및 **사전적정성 검토제**[2] 등 **혁신지원제도 운영**을 통해 **생성형 AI에 특화**된 **개인정보 보호 기준, 안전조치, 내부 관리체계 구축 방안** 등을 구체화하고 있습니다.

1) 「비정형 데이터 가명처리 기준」('24.2.), 「AI 개발·서비스를 위한 공개된 개인정보 처리 안내서」('24.7.), 「이동형 영상정보처리기기를 위한 개인영상정보 보호·활용 안내서」('24.10), 「합성데이터 생성·활용 안내서」('24.12.), 「AI 프라이버시 리스크 관리 모델」('24.12.)

2) AI 등 신기술 분야의 개인정보 보호법 준수 방안을 민·관이 함께 마련하고 사업자가 적정 이행하면 행정처분 대상에서 제외

본 안내서는 이러한 **정책적 경험**을 종합하여, **생성형 AI 수명주기**(lifecycle) **각 단계**에서 고려해야 할 **개인정보 처리 및 보호 이슈**를 체계화하고, 그에 따른 **법적 기준** 및 **안전조치** 등을 제시하였습니다. 이를 통해 생성형 AI 개발·활용에 **개인정보 보호 관점**이 균형있게 **반영**되고, 현장의 **예측 가능성**과 **자율적 준수 역량**을 높여 **신뢰**와 **책임**에 기반한 **AI 활용 환경 조성**에 기여하고자 합니다. 아울러, 본 안내서는 **글로벌 논의**와의 **연계성** 및 **정합성**을 고려하여 작성되었으며, **미국·영국·유럽연합** 등 **주요국**의 관련 **정책자료**를 참고할 수 있도록 함께 안내하였습니다.

참고 본 안내서에 포함된 주요 법령 및 정책 사례

□ 개인정보 보호법

〈 개인정보 보호법 중 관련 조항 〉

▸ **목적**(§1): 개인의 자유와 권리 보호 / 개인의 존엄과 가치 구현

▸ **원칙**(§3): 목적 명확성 / 최소수집 / 투명성 / 안전한 관리 / 정보주체 권리보장 등

▸ **개인정보의 처리**

· **수집·이용**(§15①): 동의 / 법률 / 공공기관의 소관업무 / 계약 / 정당한 이익 등
· **추가적 이용**(§15③): 당초 수집목적과 관련성 / 예측 가능성 / 정보주체 이익침해 / 안전조치
· **처리 위탁**(§26): 문서로 위탁 / 개인정보처리방침 공개 / 수탁자 감독 등
· **가명특례**(§28의2): 통계 작성·과학적 연구 목적 등 위한 동의 없는 가명정보처리
· **국외이전**(§28의8): 별도 동의·고지 / 법률·조약 / 계약 / 적정성 결정 등
· **특별한 보호가 필요한 개인정보**(§23~§24의2): 민감정보, 고유식별정보, 주민등록번호

▸ **안전한 관리**

· **파기**(§21): 목적 달성, 기간 경과 등 불필요시 파기
· **안전조치**(§29): 내부 관리계획 수립, 접속기록 보관 등 기술적·관리적·물리적 조치 시행
· **개인정보 영향평가**(§33): 위험요인 분석 및 개선사항 도출 위한 평가 시행
· **노출된 개인정보의 삭제·차단**(§34의2): 고유식별정보, 계좌정보, 신용카드정보 등 삭제·차단

▸ **정보주체 권리**

· **개인정보의 열람**(§35) / **정정·삭제**(§36) / **처리정지 등**(§37)
· **자동화된 결정에 대한 정보주체의 권리**(§37의2): AI 등 완전히 자동화된 시스템으로 개인정보를 처리하여 권리·의무에 중대한 영향을 미치는 결정에 대한 거부권, 설명요구권 등

□ 국내 AI 개인정보 처리 및 보호 관련 안내서

안내서	주요 내용
비정형 데이터 가명처리 기준 ('24.2)	▸ **이미지·영상·음성 등 비정형데이터 가명처리 기준 제시** - **의료(CT, MRI), 교통, 챗봇** 등 분야별 7종 시나리오 통해 **가명정보 활용 전 과정** 상세 안내
AI 개발·서비스 위한 공개된 개인정보 처리 안내서 ('24.7)	▸ **인터넷 등에 공개된 정보를 AI 학습에 활용할 수 있는 법적 기준 제시** - 개인정보 보호법상 **'정당한 이익'** 조항(§15①6호)의 **적용 요건** 및 **기술적·관리적 안전조치 기준** 등 제시
개인영상정보 보호·활용 안내서 ('24.10)	▸ **자율주행차, 로봇, 드론** 등 **이동형 영상정보 처리기기 촬영 영상**을 **AI 개발** 등에 활용할 수 있는 **기준 제시** - 개인영상정보의 촬영·이용·제공·보관·파기 등 처리단계별 준수사항 제시
합성데이터 생성·활용 안내서 ('24.12)	▸ **합성데이터 생성·활용 단계별 적법 절차, 원본데이터의 전처리 방식, 안전성·유용성 검증방법 및 지표 등 세부절차 안내** ※ 합성데이터 생성·검증 등 실증사례는 「합성데이터 생성 참조모델」('24.5.)에서 확인 / 합성데이터셋은 '가명정보 지원플랫폼(dataprivacy.go.kr)에서 다운로드 가능
AI 프라이버시 리스크 관리 모델 ('24.12)	▸ **AI 유형·용례·맥락**에 따른 **AI 프라이버시 리스크 경감 방안 안내** - **AI 수명주기별** 프라이버시 리스크를 **기업 스스로 평가·경감**하도록 지원

□ 해외 AI 데이터 처리 관련 참고자료

참고자료	주요 내용
미국 NIST Privacy Framework ver 1.1 ('25.4.)	▸ **미 국가기술표준연구소(NIST) 발간 개인정보 보호 및 리스크 관리 체계**로, **조직의 자율적인 프라이버시 리스크 식별·평가·관리 지원** ※ **AI 프라이버시 리스크 관리 섹션을 신설**한 개정안(ver. 1.1) 공개('25.4.) 후 의견 수렴 통해 최종안 공개 예정(~'25.12.)
영국 UK AI Playbook ('25.2.)	▸ **공공기관이 AI를 안전하고 책임감 있게 도입·활용**할 수 있도록 돕기 위한 **실무지침**으로, **조달·설계·데이터 관리·프라이버시 보호** 등 **AI 도입·운용의 전 과정**에 대한 **단계별 가이드라인 제공**
EU AI Privacy Risks & Mitigations ('25.4.)	▸ **개인정보보호이사회(EDPB)**에서 **LLM의 프라이버시 리스크를 식별·평가·완화**하기 위한 **종합적인 리스크 관리 방법론** 제시 - 개인정보 유·노출, 불투명성, 권리행사 제약 등 프라이버시 리스크에 대응하기 위한 **기술적·관리적 조치, 거버넌스 체계 마련** 등 **종합적·지속적 리스크 관리 강조**

□ AI 개인정보 처리 관련 조사·처분 등 집행사례

	구 분	주요 내용
조사·처분	AI 챗봇서비스 조사 결과 ('21.4.)	▸ **기업이 특정 목적으로 수집**한 개인정보를 **이용자의 명시적 동의 없이 관련성 없는 신규 서비스 개발**에 이용한 사안에서, - AI 챗봇 서비스의 **개인정보 목적 외 이용** 등에 대해 **시정명령 및 과징금·과태료** 부과
	출입국관리 AI 식별추적시스템 조사 결과 ('22.4.)	▸ **출입국 심사 과정에서 수집한 안면정보**를 **AI 기술 개발에 활용**하는 것의 **적법성 판단** 관련, - 안면정보를 출입국 관리시스템 고도화를 위해 활용하는 것은 출입국관리법의 목적인 안전한 국경관리를 달성하기 위한 것으로 개인정보 수집 목적 범위내에 포함된다고 판단
사전실태점검	국내외 주요 AI 서비스 대상 사전 실태점검 ('24.3.)	▸ **주요 LLM 서비스**(6개 사업자) 대상 **공개된 개인정보 처리, 이용자 입력 데이터 처리, 투명성 및 정보주체 권리보장** 관련 **개선권고**
	AI 응용서비스 사전 실태점검 ('24.6.)	▸ **AI 응용서비스**(AI 통화요약, 이미지 생성 등 4개 서비스)의 **개인정보 처리 과정에 대한 실태점검** 후, 일부 사업자 대상 **시스템 접속기록 보관·점검 등 안전조치 준수 시정권고** 및 **투명성 강화 관련 개선권고**
	딥시크 서비스 사전 실태점검 ('25.4.)	▸ **개인정보 국외이전 근거 충실 구비** 및 **이미 이전한 개인정보 즉각 파기, 한국어 개인정보 처리방침 투명성 제고** 등 **시정권고** ▸ **AI 관련 강화된 보호조치 준수, 처리시스템 전반 점검 및 안전조치 수준 향상, 국내대리인 지정** 등 **개선권고**
	AI 디지털 교과서 사전 실태점검 ('25.5.)	▸ **학습 이력**을 **데이터베이스화**하여 **저장**하고 **개인 맞춤형 콘텐츠**를 **제공**하는 AI 디지털 교과서의 개인정보 보호 실태점검 결과, - 개인정보 처리의 **적법성·투명성 확보, 정보주체 권리행사 명확화** 등 **시정권고, 안전조치 관련 개선권고** 등

□ AI 개인정보 처리·보호 관련 분야 혁신지원제도 사례

	구 분	주요 내용
규제샌드박스	자율주행 AI 학습 ('24.2)	▸ 자율주행·로봇기업 등이 **강화된 안전조치 준수** 下 동의 없이 **영상 원본 활용 허용** ※ 뉴빌리티·우아한형제들·현대차 등 5개 기업 승인
	첨단바이오 국제공동연구 ('24.6.)	▸ 서울대병원·하버드·MIT 등과 **첨단바이오 분야 국제공동연구**에 필요한 **가명데이터셋**을 환자 동의 없이 활용 허용
	보이스피싱 예방 AI 학습 ('24.10.)	▸ 금감원·국과수 보유 '그놈 목소리' 통화 데이터 2만 5천건을 **강화된 안전조치** 下 **보이스피싱 예방 AI 학습에 활용 허용**

사전적정성 검토제	**의료기관 내 생성형 AI 기반 보고서 작성 지원** ('24.10.)	▸ 의료기관 내 진료 대화 기반 **의료기록 작성업무**를 **생성형 AI를 활용하여 자동화**하는 솔루션 개발 ▸ 개인정보의 적법한 처리위탁 요건 충족 위해 ①진료대화 데이터 처리에 관한 사항을 고지하고 ②**기업용(Enterprise API) 라이선스**를 사용해 **의료기관의 목적으로만 데이터를 처리**하도록 조치 ※ 개인정보보호위원회 제2024-017-237호(비공개)
	대화형 AI 서비스 개발 ('25.3.)	▸ 대화 맥락과 이용자 감정을 파악해 최적화된 답변을 제시하기 위해 **자체 가드레일 모델*(LLM)과 외부 생성형 AI 병행 활용** * 유해한 표현, 개인정보 등 탐지 ▸ ①개인정보 처리 위탁 요건 준수 ②**이용자 발화문에 대한 안전조치 강화** ③**사후 관리체계 구축** 등의 조건 부과 ※ 개인정보보호위원회 제2025-004-016호(비공개)
	이용자 대화 데이터의 AI 모델 학습 ('25.5.)	▸ 대규모 언어모델 기반 대화형 AI 서비스를 제공하는 과정에서의 **이용자 프롬프트 입력의 AI 학습 활용** 방안 검토 ▸ 개인정보 보호법상 **추가적 이용 조항**(제15조제3항) 요건 검토 ※ 개인정보보호위원회 제2025-011-031호(비공개)
	보이스피싱 의심번호 DB 구축·활용 ('25.7.)	▸ **전화 수발신 내역 데이터**를 활용하여 **보이스피싱 의심번호를 예측**하고 이를 **금융사의 이상거래 탐지·차단**에 이용 ▸ 보이스피싱 범죄자의 통화패턴 등을 학습*한 예측 모델 생성 * 경찰청 등에서 공유받은 보이스피싱 신고 번호의 통화패턴(전화번호, 발신일시, 종료일시 등 통화내역 데이터)을 분석·학습 ▸ 개인정보 보호법상 **추가적 이용 조항**(제15조제3항) 요건 등 검토 ※ 개인정보보호위원회 제2025-015-231~232호(비공개)

2 적용 범위 및 성격

본 안내서는 **생성형 AI를 개발·활용**하면서 **개인정보를 처리**하는 **기업·기관** 등이 참고할 수 있도록 마련되었습니다[3]. 예를 들어, ▲LLM을 개발하고 제공하는 **모델 개발자** 및 ▲모델을 이용해 AI 서비스를 개발·제공하는 **모델 이용자** 등이 본 안내서를 참고하여 생성형 AI 수명주기 단계별 개인정보 보호 사항을 검토할 수 있습니다[4]. 아울러, **생성형 AI의 데이터 처리**와 관련된 **개인정보 보호 컴플라이언스, 보안, 리스크 관리, 데이터 거버넌스** 등 **업무담당자**를 주요 독자층으로 상정하였습니다.

본 안내서는 생성형 AI 개발·활용과 관련한 **개인정보 보호 준수 가능성**을 높이기 위한 것으로서, **다른 법령상 의무**(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률, 저작권법 등)에 대해서는 **다루지 않습니다.**

또한 본 안내서는 생성형 AI의 개발·활용에 개인정보 처리가 수반되는 경우 그 처리 행위의 적법성, 안전성 확보 등을 위해 **고려해야 할 최소한의 사항을 안내**한 것으로, 향후 **AI 관련 법·제도·기술**이 발전함에 따라 **주기적으로 수정·보완**될 수 있습니다. 현재 **주로 언어모델** 기반 생성형 AI 중심으로 작성되었으나, 향후 **음성·이미지·영상 등 영역까지 점진적으로 확대·보완**해 나갈 예정입니다.

---

3) AI 생태계 내 다양한 참여자에 대한 분류체계, 정의 등은 다양한 방식으로 논의되고 있으며, 프라이버시 영역에서는 AI 개발·활용 과정에서 개인정보 처리 행위가 이루어지는지 여부가 주된 쟁점임

※ (인공지능 발전과 신뢰 기반 조성 등에 관한 기본법) 인공지능산업과 관련한 사업을 하는 자('인공지능사업자')로 '인공지능개발사업자'(인공지능을 개발하여 제공하는 자)와 '인공지능이용사업자'(인공지능개발사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자)로 분류
(EU AI ACT) 공급자(provider), 배포자(deployer) 등으로 분류
(NIST AI RMF) AI 설계(design) 행위자, 개발(development) 행위자, 배포(deployment) 행위자 등으로 분류
(ISO/IEC 42001) AI 제공자(provider), AI 생산자(producer) 등으로 분류

4) 생성형 AI 생태계 참여자는 크게 모델을 개발(자체개발 및 추가학습·정렬 등)하거나 이용하는 자로 분류할 수 있으며 개별 AI 개발·활용 맥락에 따라 동일한 사업자가 모델개발자이면서 모델이용자 둘 모두에 해당할 수 있음

※ 생성형 개발·활용 유형 분류에 따른 개인정보 검토 사항 내용은 본 안내서 붙임 자료 참고

Ⅱ. 생성형 AI 개발 · 활용 단계[5]

목적 설정

☑ 생성형 AI 기술을 통해 달성하고자 하는 **구체적 목적 설정**

- 생성형 AI의 **사용 맥락·대상·기술적 한계** 고려

☑ **개인정보 고려사항**

√ 개인정보 보호 원칙 下 **개인정보 처리 목적 구체화**

√ 개인정보 처리의 **적법근거** 확보

※ 공개된 개인정보, 이용자 개인정보 등 수집 출처별 적법성 별도 검토

전략 수립

☑ 조직의 **ICT 역량·여건**에 대한 객관적 진단 통해 **AI 개발·활용의 구체적 방식** 결정 및 **리스크 관리 방안** 등 **핵심 전략 수립**

☑ **개인정보 고려사항**

√ **개인정보 안심설계**(PbD, Privacy by Design) 접근 보장을 위한 **개인정보 영향평가** 실시

√ LLM 개발·활용 방식별 유의사항 고려

※ 서비스형 LLM, 기성 LLM 활용, 자체개발

AI 학습 및 개발

☑ 당초 설정한 **목적**을 효과적으로 **달성**하고 의도치 않은 **리스크를 완화**하기 위해 **데이터 (추가)학습, 모델 미세조정·정렬** 등 수행

☑ **개인정보 고려사항**

√ 데이터 오염(poisoning) 방지하기 위한 **출처 검증, 전처리, 가명·익명처리** 등

√ AI 모델에 대한 **미세조정, 정렬** 등 안전장치 추가

√ AI 시스템 **접근권한 통제, 입출력 필터링** 적용

시스템 적용 및 관리

☑ 생성형 AI 시스템을 **실제 서비스 환경에 적용**하여 **사용자와의 상호작용**이 시작되는 단계

- **시스템 성능·안전성**을 지속 **유지·보완**하고 **정보주체 권리 침해 방지**

☑ **개인정보 고려사항**

√ **배포 전 테스트** 통한 프라이버시 리스크 점검·문서화

√ **허용되는 이용방침**(AUP, Acceptable Use Policy) 작성·공개

√ 개인정보 침해 시 신고 기능 및 **정보주체 권리 보장 방안** 마련

AI 프라이버시 거버넌스 구축

☑ 전사적 차원에서 개인정보 보호를 위한 **내부 관리체계를 마련**하여 생성형 AI 목적 설정부터 적용·관리까지 이어지는 **전체 프로세스 관리·감독**

☑ **구성** : **개인정보 보호책임자**(CPO, Chief Privacy Officer) 중심의 내부 관리체계 구성

- **개발·보안 등 유관 부서와 긴밀한 협력 체계**를 유지하며 **AI 개발·활용 초기부터 참여**

※ 개별 여건에 따라 규모·구성 등을 자율적으로 결정

☑ **역할** : AI 프라이버시 리스크 관리 **정책 마련·문서화** / 개인정보 취약성 **상시 모니터링·평가 및 보고** / **정보주체 권리행사 지원**

5) 「공공부문 초거대 AI 도입·활용 가이드라인」('25.4, 디지털플랫폼정부위원회) 및 「A platform-centric approach to scaling generative AI in the enterprise」('24.9, Google Cloud), 「Llama Developer Use Guide: AI Protections」('25.4, Meta), ISO/IEC 5338(AI system life cycle processes) 등 **국내외 AI 생애주기 자료를 참고하여 개인정보·데이터 처리 관점에서의 고려사항이 반영될 수 있는 체계**를 구성

Ⅲ. 생성형 AI 개발·활용 단계별 고려사항[6]

1 목적 설정

❑ 생성형 AI 개발·활용의 첫 관문인 목적 설정, 왜 중요한가요?

생성형 AI 개발·활용의 출발점은 **생성형 AI 기술**을 통해 **달성하려는 목적**을 설정하는 것입니다. 이는 크게 **두 가지 측면**에서 중요합니다. **첫째**, 목적 달성에 **어떤 개인정보**가 **얼마나 필요**한지 파악하고 **개인정보의 처리 목적**을 획정할 수 있게 합니다. **둘째**, 목적 설정 **후속 단계**에서 **어떤 리스크**가 발생할 수 있는지 가늠하고 초기부터 **리스크 관리 및 완화 방안**을 모색할 수 있도록 합니다.

❑ 목적 설정은 어떻게 하나요?

생성형 AI의 **사용 맥락·대상·기술적 한계** 등을 고려하여 **목적을 구체화**해야 합니다. AI가 **어떤 맥락**에서 **누구를 대상**으로 사용되는지 **의도된 용례**(intended use)를 명확히 정의하고, **AI의 기능·성능**이 **용례에 부합**하는지 검토하는 한편, **예견 가능한 오용**(foreseeable misuse) 등 **한계점**을 사전에 파악하는 과정이 포함되어야 합니다[7].

❑ 프라이버시 관점에서 주의할 점은?

생성형 AI의 사전·추가학습에 필요한 **데이터**에는 **개인정보가 포함**될 수 있습니다. 이때, **개인정보 처리 목적**은 **구체적**이고 **명확**하며 **합법적**으로 설정되어야 합니다[8].

---

6) 본 장에서는 생성형 AI 개발·활용 절차·단계별로 가장 관련성 높은 개인정보 보호 고려사항을 안내하고 있으며, **개별사항이 반드시 특정 절차·단계에 고유한 것은 아님**

※ <예> **거버넌스 구축**은 <1. 목적 설정>부터 <4. 시스템 적용 및 관리>까지 **전 단계**에 걸쳐 이루어지는 것이 바람직함

7) GDS, 「AI Playbook for the UK Government」, "Principle 1: You know what AI is and what its limitations are" 및 "Principle 2: You use AI lawfully, ethically and responsibly" 참고

8) ICO, 「Purpose limitation in the generative AI lifecycle」 (2024) 참고

범용성을 전제로 하는 **생성형 AI**의 경우 **개인정보 처리 목적**을 **사전에 구체화**하는 데 일정한 **한계**가 있을 수 있으나, 기업·기관은 **개인정보 보호 원칙**을 고려하여 **처리 목적을 최대한 구체화**하는 것이 권장됩니다.

법률 **제3조(개인정보 보호 원칙)** ① 개인정보처리자는 개인정보의 **처리 목적을 명확**하게 하여야 하고 **그 목적에 필요한 범위**에서 **최소한의 개인정보**만을 **적법하고 정당하게 수집**하여야 한다.
② 개인정보처리자는 개인정보의 **처리 목적에 필요한 범위**에서 **적합하게 개인정보를 처리**하여야 하며, 그 목적 외의 용도로 활용하여서는 아니 된다.
③ 개인정보처리자는 개인정보의 **처리 목적에 필요한 범위**에서 개인정보의 **정확성, 완전성 및 최신성**이 보장되도록 하여야 한다. (이후 생략)

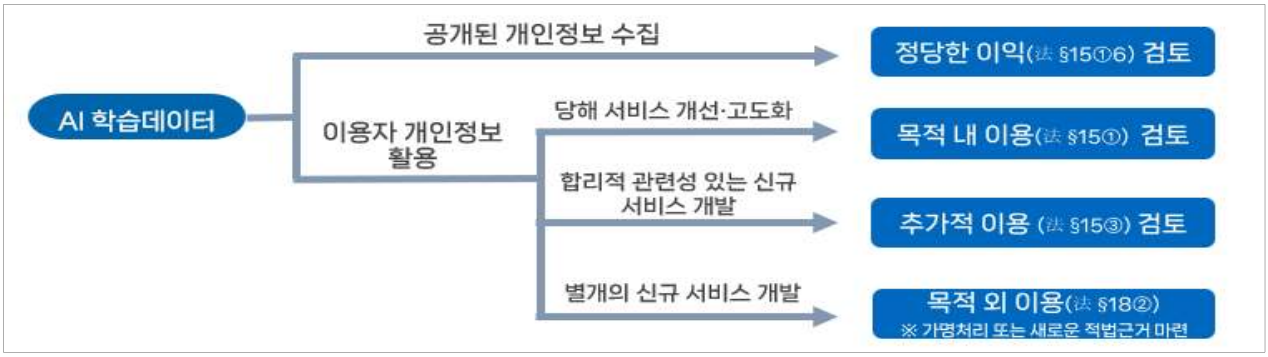
사례 **개인정보 처리 목적 미정립 사례** ※ AI 디지털교과서 사전 실태점검 결과('25.5.)

- AI 디지털교과서 통합포털은 **학생별 학습콘텐츠 이용내역 데이터***를 통계 목적 또는 향후 AI 기반 학습분석 목적 등을 위해 저장함
  * 학습시간, 성취수준, 진도율, 접속시간, 커뮤니티 참여도 등
- **학생 개개인의 상세한 학습 정보**가 통합 DB에 누적될 경우 **학생의 일상행동 감시 등 오남용 우려**가 제기될 수 있고, 동 정보를 AI 학습데이터로 활용하는 **처리 목적이 분명하게 정립되지 않은 상황**이었음

☞ 개인정보 항목, 목적, 보유기간 등을 처리방침 등을 통해 정보주체에게 누락 없이 고지토록 하고, 특히, **통합 DB에 관리되는 데이터에 대해서는 처리 항목 및 목적을 보다 명확히 할 것을 시정권고함**(法 제3조제1항 위반)

생성형 AI 개발·활용을 위한 개인정보 처리 목적을 구체화했다면, 그 목적에 적합한 **개인정보 처리의 적법근거**(lawful basis)가 무엇인지 확인해야 합니다. 생성형 AI 학습에 투입되는 개인정보의 출처는 크게 **▲공개된 개인정보를 수집**하는 방안 **▲**기 보유하고 있는 **이용자 개인정보**를 **재사용**하는 방안으로 구분할 수 있고, 각 기업·기관은 개인정보 **수집 출처별**로 개인정보 처리의 **적법근거를 확인**해야 합니다.

< AI 개인정보 수집 출처별 적법근거 검토 방향 >



공개된 개인정보의 수집 · 이용

LLM 등 기초모델을 개발하는 경우에는 **대량의 데이터**를 스크래핑 방식으로 **직접 수집**하거나 **거대 말뭉치**[9] 등을 사용하는 것이 현실적 관행으로 자리잡고 있습니다. 이러한 기초모델이 실제 현실에 관한 광범위한 지식과 적응력을 확보하기 위해서 **공개된 개인정보 처리의 필요성**이 인정될 수 있으나, **무분별한 스크래핑**을 통한 **권리침해 우려**도 있으므로 **적법성·안전성 확보**에 특히 주의를 기울여야 합니다.

특정 서비스 매개 없이 인터넷 공간에서 개인정보를 수집하는 경우, 기업·기관과 정보주체 사이에 직접적인 관계가 형성되지 않아 동의·계약과 같은 적법근거를 적용하는 것이 어려울 수 있습니다. 이 경우 실질적으로 고려할 수 있는 적법근거로는 **정당한 이익 조항**(法 제15조제1항제6호)이 있으며, 이를 충족하기 위해서는 **▲목적의 정당성 ▲공개된 개인정보 처리의 필요성 ▲이익형량**의 3가지 기준을 검토해야 합니다. 특히, 이익형량 단계에서 정보주체 권리 침해 가능성을 최소화하기 위한 **기술적·관리적 안전조치** 및 **정보주체 권리보장 방안**을 마련하는 것이 핵심입니다.

참고 **「공개된 개인정보 처리 안내서」('24.7) 주요내용**

- **(목적의 정당성)** 개인정보처리자의 정당한 이익의 존재
  - AI 기업·기관의 **영업상 이익**뿐 아니라 **그로부터 발생하는 사회적 이익** 등 **다양한 층위의 이익**을 포괄
  - **다양한 과제**(downstream task) 수행이 가능한 **생성형 AI**의 **개인정보 처리 목적**을 **특정하기 어려운 한계 → AI의 유형·기능·성능**을 고려하여 **정당한 이익 명확화**
- **(처리의 필요성)** 공개된 개인정보 처리의 **필요성**과 **상당성·합리성**이 인정될 것
  ※ (예) 의료진단보조 AI 개발시 개인의 소득·재산 등 관련없는 정보는 학습 배제
- **(이익형량) 개인정보처리자의 정당한 이익**이 **정보주체 권리**에 **명백히 우선**
  - **명백성 요건** 충족을 위해 ▲정보주체 권익침해 방지를 위한 **안전성 확보조치** **▲정보주체 권리보장 방안** 마련으로 **개인정보처리자 이익이 우선**하도록 조치

9) 일반적으로 수십억 개 이상의 단어로 구성된 대규모 자연어 텍스트 데이터셋을 의미하며, 웹 문서, 뉴스, 논문 등 다양한 출처에서 수집됨. 대표적인 영어 기반 말뭉치로는 Common Crawl, Wikipedia, OpenWeb Text 등이 있고, 한국어 특화 말뭉치로는 AI Hub 말뭉치, 모두의 말뭉치 등이 있음

1 기술적 안전조치	2 관리적 안전조치	3 정보주체 권리보장
- 학습데이터 수집 출처 검증 - 개인정보 유·노출 방지 - 미세조정을 통한 안전장치 추가 - 프롬프트 및 출력 필터링 적용	- 학습데이터 처리기준 정립 및 개인정보처리방침에 공개 - AI 프라이버시 레드팀 운영 - 오픈소스·API연계 등 개발·배포방식 특성에 따른 안전조치	- 시간·비용·기술적 측면에서 합리적으로 실현가능한 범위 내 권리보장 방안 마련·안내 - 다만, AI 학습 데이터 특성상 전통적인 정보주체 권리행사가 일부 제한될 수 있음을 명시

이용자 개인정보의 수집 · 이용

이용자 개인정보를 **AI 서비스 개선**(고도화 포함) 또는 **신규 AI 서비스 개발**에 이용하고자 하는 기업·기관은 개인정보의 **당초 수집 목적과 AI 서비스의 관련성**을 기준으로 **자체 평가**를 거쳐 **적법근거를 선택**할 수 있습니다.

1 수집 목적 내 서비스 개선·고도화

동의, 계약, 정당한 이익 등 적법근거에 기초하여 수집된 이용자 개인정보는 **수집 목적 범위 내**에서 **AI 서비스 개선·고도화** 등을 위해 이용할 수 있습니다.

참고 **「개인정보 처리 통합 안내서」**('25.7) **내 관련 내용**

- **(정보주체의 동의)** 당초 수집한 목적을 명확히 하여 자유로운 의사에 따른 명시적 동의를 받은 경우에는 그 범위 내 이용 가능

(예시) AI 개발을 위한 학습데이터 이용 목적인 경우, 동의 전 법정 고지 사항에 목적을 명확히 기재

- **(계약 이행)** 당초 서비스 이용 계약의 이행을 위해 기능 개선 등이 필요한 경우 기존에 수집한 개인정보 이용 가능

(예시) 서비스 이용계약의 내용에 AI 기능이 포함되어 있는 경우로서 서비스 제공을 위해서는 AI 기능 개선 등이 필요하고, 이에 대해 정보주체가 충분히 예측할 수 있는 경우

- **(정당한 이익)** 개인정보처리자의 정당한 이익에 포섭될 수 있는 이익(부정행위 방지, 서비스의 안전성 보장 등)이 존재하고 정보주체의 권리 침해 가능성을 최소화한 경우

(예시) 정보주체가 범죄·사회적 질서 위반 등으로 인해 보호 필요성이 현저히 낮거나, 정상적인 이용자의 이익 보호 필요성이 큰 경우(예: AI 모델 채택한 FDS) 정당한 이익이 우선할 수 있음(단, 구체적 안전조치 수준 및 정보주체 권리보장 등에 대한 개별 이익형량 필요)

② 수집 목적과의 합리적 관련성 있는 이용

이용자 개인정보를 **수집 목적과 합리적 관련성** 있는 **AI 학습·개발**에 동의 없이 활용하는 경우에는 **추가적 이용 조항**(法 제15조제3항)을 적법근거로 검토할 수 있습니다. 해당 기준이 인정되기 위해서는 **▲합리적 관련성 ▲정보주체의 예측 가능성 ▲정보주체 이익의 부당한 침해 가능성 ▲가명처리·암호화 등 안전성 확보 조치** 등이 종합적으로 고려되어야 합니다. 만약 개인정보의 추가적 이용이 **지속적으로 발생**하는 경우에는 판단기준을 **개인정보 처리방침에 공개**하고, 해당 기준에 따르고 있는지 여부를 **개인정보 보호책임자**(CPO, Chief Privacy Officer)가 **점검**해야 합니다.

법률 **제15조(개인정보의 수집·이용)** ③ 개인정보처리자는 당초 수집 목적과 합리적으로 관련된 범위에서 정보주체에게 불이익이 발생하는지 여부, 암호화 등 안전성 확보에 필요한 조치를 하였는지 여부 등을 고려하여 대통령령으로 정하는 바에 따라 정보주체의 동의 없이 개인정보를 이용할 수 있다.

시행령 **제14조의2(개인정보의 추가적인 이용·제공의 기준 등)** ② 개인정보처리자는 개인정보의 추가적인 이용 또는 제공이 지속적으로 발생하는 경우에는 제1항 각 호의 고려사항에 대한 판단 기준을 법 제30조제1항에 따른 개인정보 처리방침에 공개하고, 법 제31조제1항에 따른 개인정보 보호책임자가 해당 기준에 따라 개인정보의 추가적인 이용 또는 제공을 하고 있는지 여부를 점검해야 한다.

사례 1 **이용자 개인정보의 AI 학습 사례** ※ 개인정보보호위원회 제2025-011-031호(비공개) 결정

▪ **LLM 성능 개선**을 위해 **이용자 프롬프트 입력 내용**을 **AI 학습데이터로 수집·이용** 시,

① LLM의 **환각(hallucination), 편향(bias) 등 리스크를 완화**하고 성능을 높이기 위해서 **이용자와의 상호작용 데이터가 학습에 활용될 필요**가 있으며 이는 **LLM 서비스 운영과 밀접하게 관련**된 점,

② LLM 서비스는 이용자의 질문 맥락을 파악하여 통계적으로 적절하다고 선정된 답변 문구를 생성하는 **그 성질상 기존 대화 내용이 학습데이터로 활용될 수 있음**을 **예측 가능**하고, **학습데이터 수집 사실** 및 **거부 방법**(opt-out)을 대화창 알림으로 **수차례 고지**하여 **예측 가능성을 높이는 점**,

③ 대화 데이터는 모델 **학습 데이터로만 활용**되고 구축되는 통계모형 자체로는 정보주체에게 직접적인 영향을 미치지 않으며, 이용자가 학습데이터 수집을 **항시 옵트아웃(opt-out) 할 수 있는 기능**도 제공하여 **정보주체의 이익을 부당하게 침해하지 않는 점**,

④ 학습데이터셋 내 개인식별 가능성이 높은 정보를 **탐지·삭제하는 필터링 절차를 운영**하는 점 등을 고려하면 **개인정보의 추가적 이용을 근거**로 이용자 개인정보를 모델 학습데이터로 사용 가능

사례 2 **이용자 개인정보의 AI 활용 사례** ※ 개인정보보호위원회 제2025-015-231~2호(비공개) 결정

▪ **이용자 통화내역 데이터**를 사용한 **보이스피싱 의심번호 DB를 구축**하는데 있어,

① 보이스피싱은 **통신망을 악용**한 금융사기로, **이용자 보호 의무 준수**를 위해 이를 사전에 예방하고 차단하는 것은 **정상적인 통신·금융 서비스 제공과 밀접하게 관련**된 점,

② 대다수 통신사·금융사는 기존에도 스팸·이상거래 방지 등 **부정이용 및 이용자 보호 목적의 서비스를 운영 중**이며, **이용약관 등을 통해 이를 안내**하고 있어 **고객 보호 의무의 연장선으로 보이스피싱 예방·탐지 서비스에 대한 예측이 가능**한 점,

③ 통신사가 예측한 의심번호의 **정·오탐 여부**를 **금융사가 별도 검토하고 피드백**하는 절차가 구축되어 있어 **정보주체의 이익을 부당하게 침해하지 않는 점**,

④ 보이스피싱 **의심 DB 구축** 및 통신사, 금융사 간 **통신** 시 **전화번호가 암호화**되어 저장 및 송·수신되는 점 등을 고려하면 **개인정보의 추가적 이용을 근거**로 통화 내역 데이터를 이용해 보이스피싱 의심번호 DB 구축·이용 가능

3 당초 수집 목적과 별개의 신규 서비스 개발

이용자 개인정보를 당초 수집한 목적과 **별개의 신규 AI 서비스** 개발에 활용하는 경우에는 **▲가명·익명처리**(法 제28조의2 및 제58조의2)하여 이용[10]하거나 **▲새로운 적법근거**(法 제18조제2항) 마련이 필요합니다.

법률 **제28조의2(가명정보의 처리 등)** ① 개인정보처리자는 통계작성, 과학적 연구, 공익적 기록보존 등을 위하여 정보주체의 동의 없이 가명정보를 처리할 수 있다.

제58조의2(적용제외) 이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보에는 적용하지 아니한다.

제18조(개인정보의 목적 외 이용·제공 제한) ② 제1항에도 불구하고 개인정보처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 정보주체 또는 제3자의 이익을 부당하게 침해할 우려가 있을 때를 제외하고는 개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공할 수 있다. 다만, 제5호부터 제9호까지에 따른 경우는 공공기관의 경우로 한정한다.

1. 정보주체로부터 별도의 동의를 받은 경우
2. 다른 법률에 특별한 규정이 있는 경우
3. 명백히 정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우
4. 삭제 <2020. 2. 4.>
5. 개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공하지 아니하면 다른 법률에서 정하는 소관 업무를 수행할 수 없는 경우로서 보호위원회의 심의·의결을 거친 경우 (이하 생략)

사례 1 질병 진단을 보조하는 **의료 AI** 연구개발을 위해 정보주체의 동의 없이 병원이 보유한 **MRI, CT, X-Ray 사진** 등을 **가명처리하여 학습데이터**로 이용

10) "가명정보 처리 가이드라인"('24. 2., 개인정보보호위원회) 참고

사례 2 금융당국, 수사기관 등이 보유한 **보이스피싱 통화데이터**를 **가명처리**해 통신사 등이 정보주체의 동의 없이 **보이스피싱 예방 AI 기술·서비스 개발에 활용**

한편, **가명·익명처리**나 **새로운 적법근거 마련** 없이 **개인정보를 당초 수집 범위를 벗어나 이용**하는 경우에는 **적법하지 않은 목적 외 이용**으로 판단될 수 있습니다. 단, 해당 이용이 **혁신성, 공익성** 등을 갖춘 경우에는 **규제샌드박스 제도**를 활용해 **강화된 안전조치** 등 일정 요건을 전제로 **개인정보 처리 근거**를 **확보**할 수 있습니다.

사례 3 서비스 품질 개선 목적으로 수집한 이용자 대화데이터를 **합리적 관련성 없는 신규 서비스(인공지능 챗봇) 개발**에 **암호화 등의 안전조치 없이** 이용한 사안에 대해 **적법 처리 근거가 없어 이용자 개인정보의 목적 외 이용으로 판단**

※ 개인정보보호위원회 제2021-007-072호 결정 참고

사례 4 자율주행 기업이 수집한 영상정보를 가명처리하여 AI 학습에 이용시 **자율주행 AI의 성능 향상에 어려움**이 있는 상황 → 규제실증특례를 활용해 강화된 안전조치 준수 下 동의 및 가명처리 없이 **영상 원본 활용 허용**

④ 특수한 개인정보의 처리

AI 서비스 개발에 **민감정보, 고유식별정보** 등의 처리가 수반되는 경우에는 **별도의 동의** 또는 **법적근거**가 있는 경우에 한하여 처리할 수 있습니다. 또한, **개인영상정보**를 **AI 학습**에 이용하는 경우에는 **설치·운영·촬영 목적의 범위 내**에서 처리해야 합니다.

사례 1 정보주체의 **동의를 받아** 자유 발화를 통해 수집한 **음성정보**를 저장한 후, AI 음성인식기술 기반 목소리 인증 서비스(**민감정보 처리, 法 제23조제1항제1호**)에 활용할 수 있음

사례 2 지방자치단체가 **교통정보의 수집·분석(法 제25조제1항제5호)** 등을 위하여 CCTV를 설치하고 해당 영상정보를 AI 기반 스마트교차로, 감응신호시스템에 이용 가능

2 전략 수립

❑ 전략 수립 단계에서는 무엇을 하나요?

생성형 AI 기술을 통해 달성하고자 하는 목적에 필요한 개인정보 범위와 적법 처리 근거가 구체화되었다면, 기업·기관은 **생성형 AI 개발·활용 전략을 수립**하게 됩니다.

이 단계에서는 기업·기관이 자사의 **ICT 역량·여건**('as-is')을 객관적으로 진단하고, 이를 토대로 생성형 AI 개발·활용에 필요한 **목표 상태**('to-be')**를 설계**합니다.

이러한 분석을 바탕으로 **▲생성형 AI의 개발·활용 방식**(상용 솔루션 구매, 공개 모델 활용, 자체 개발 등) **▲미세조정 등 추가학습 여부 ▲데이터 품질 확보 ▲리스크 관리 방안** 등 **후속 의사결정의 방향**을 구체화할 수 있습니다.

이 때, **유사 분야**에 대한 **타 기업·기관의 개념증명**(PoC, Proof of Concept) **수행 사례**, 국내외 **벤치마크 사례**는 기술적 구현 가능성, 법적 리스크 등을 사전에 파악하고 자사 여건에 최적화된 개발·활용 전략을 수립하는데 중요한 **참고자료**가 될 수 있습니다.

전략 수립 시에는 **조직의 사업적 요구사항**과 **규제 및 사회적 요구사항**을 종합적으로 반영할 수 있도록 **강력한 AI 프라이버시 거버넌스** 구축을 **병행**할 것이 권장됩니다[11].

❑ 어떤 개발·활용 방식이 일반적으로 선택되며, 각각의 특징은 무엇인가요?

생성형 AI의 개발은 **자사의 기술 역량, 데이터 보호 요구사항, 법적 책임 범위** 등을 종합적으로 고려하여 다양한 방식으로 이루어질 수 있습니다. 다음은 LLM 기반 AI 개발의 **대표적인 세 가지 방식**에 대한 개요와 선택 시 고려사항을 정리한 내용입니다.

11) 본 안내서 제3장 제5절 "AI 프라이버시 거버넌스 체계" 참고

< **생성형 AI 개발 방식 분류**[12] >

구 분	내 용	예 시
서비스형 LLM (LLM as a Service)	▪ **비공개 모델** 등 **상용 AI 서비스**를 활용(예: API 연계)하는 방법으로 **신속한 개발** 가능 - 이용사업자는 **LLM 모델 접근이 제한**되어 파라미터 등 **세부 조정에 한계** ▪ 이용사업자는 서비스형 LLM으로 전송되는 데이터의 **처리 목적·범위, 보관·파기 정책** 등 확인 필요	▪ 이용자 발화문 분석 및 답변 생성 위해 OpenAI ChatGPT 연동
기성 LLM 활용 (Off-the-Shelf LLM)	▪ 주로 **공개(open-weight)**된 **사전학습된 모델(pre-trained model)**을 활용해 AI 시스템을 개발하는 방식 - 서비스형 LLM 방식에 비해 기성 LLM 활용 시 **더 높은 모델 통제권을 보유** ▪ 사전학습 모델의 상업적 이용 가능 여부, 재배포 제한 여부, 개인정보 안전장치 내재 여부 등 확인 필요	▪ Llama 등의 공개모델에 법률 전문지식을 추가학습하여 법률AI 개발
자체개발 (Self-developed LLM)	▪ 모델을 **처음부터**(from the ground up) **직접 사전학습**(pre-train)하는 방식으로 **고비용 인프라 및 전문인력** 등 자원 요구 ▪ AI 데이터 처리 전 과정에 대한 통제와 안전성 확보 가능한 장점	▪ 자체개발 SLM을 활용한 온디바이스 음성인식 보이스피싱 탐지 솔루션 개발

❑ 전략 수립 시 프라이버시 관점에서 공통적으로 고려할 사항은?

생성형 AI 개발·활용 전략은 개인정보 최소화, 프라이버시 강화 기술 등을 적용하는 **개인정보 안심설계**(PbD, Privacy by Design) **원칙을 반영**해야 합니다. PbD란 **제품 및 서비스**의 **기획, 개발, 활용** 전 과정에서 **개인정보 보호법 준수**를 **사전에 보장**하는 접근입니다.

참고 **PbD 구현 체계 및 적용 방안** ※「개인정보 보호책임자(CPO) 핸드북」('24.11.)

▪ **PbD에 따른 개인정보 보호 조치 예시**

- 개인정보 처리의 적법성 확보, 가명·익명처리, 개인정보 보호 기본 설정(Privacy by Default), 개인정보 보호 강화기술(차분프라이버시, 동형암호화, 합성데이터 등)

12) "AI Privacy Risks & Mitigations - Large Language Models", (EDPB) 참고

※ 해당 분류는 AI 기술이 사용되는 다양한 유형 중 **대표적인 유형 위주로 구분**된 것으로 관련 **논의가 현재도 진행 중인 상황**이며, **상호배타적으로 완전히 구분되지 않고 일부 경계가 중첩되거나 혼합 적용**될 수 있음. 예를 들어 서비스형 LLM을 활용하되, 민감한 데이터의 유출 가능성을 최소화하고 조직 내부의 보안 정책을 유지하기 위해 내부 프록시 서버를 거치도록 하는 하이브리드 아키텍처를 구현할 수 있음.

■ **CPO 중심의 PbD 원칙 확립 및 조직환경에 맞는 PbD 체계 구축**
- 서비스 특성, 개발 절차, 가용자원, 개인정보 처리에 따른 위험도 등을 확인하고, 새로운 개인정보의 처리가 이루어지기 전에 최대한 빠른 시기부터 논의 착수
- 사업부서, 개발부서 등 유관부서와 함께 개인정보 보호 조치의 내재화 방안을 논의하고 PbD 관련 상호 협력체계를 조성함으로써 업무 효율화 기반 마련

개인정보 영향평가(PIA, Privacy Impact Assessment)는 PbD 원칙을 AI 제품·서비스에 반영하는데 유용한 **실천적 수단**이 될 수 있습니다. **공공기관**의 경우 **법령상 요건**에 해당하면 개인정보 영향평가를 **의무적으로 수행**해야 합니다.

민간 부문은 개인정보 침해가 우려되는 경우 영향평가를 하기 위하여 **적극 노력**하여야 합니다. 특히 **대규모** 또는 **민감한 개인정보** 처리가 수반되거나 **정보주체 권리에 중대한 영향**을 줄 수 있는 AI 시스템의 경우 **개인정보 영향평가를 실시**하는 것이 권장됩니다.

법률 **제33조(개인정보 영향평가)** ① 공공기관의 장은 대통령령으로 정하는 기준에 해당하는 개인정보파일의 운용으로 인하여 정보주체의 개인정보 침해가 우려되는 경우에는 그 위험요인의 분석과 개선사항 도출을 위한 평가(이하 "영향평가"라 한다)를 하고 그 결과를 보호위원회에 제출하여야 한다.
②~⑩ (중략)
⑪ 공공기관 외의 개인정보처리자는 개인정보파일 운용으로 인하여 정보주체의 개인정보 침해가 우려되는 경우에는 영향평가를 하기 위하여 적극 노력하여야 한다.

시행령 **제35조(개인정보 영향평가의 대상)** 법 제33조제1항에서 "대통령령으로 정하는 기준에 해당하는 개인정보파일"이란 개인정보를 전자적으로 처리할 수 있는 개인정보파일로서 다음 각 호의 어느 하나에 해당하는 개인정보파일을 말한다.
1. 구축·운용 또는 변경하려는 개인정보파일로서 5만명 이상의 정보주체에 관한 민감정보 또는 고유식별정보의 처리가 수반되는 개인정보파일
2. 구축·운용하고 있는 개인정보파일을 해당 공공기관 내부 또는 외부에서 구축·운용하고 있는 다른 개인정보파일과 연계하려는 경우로서 연계 결과 50만명 이상의 정보주체에 관한 개인정보가 포함되는 개인정보파일
3. 구축·운용 또는 변경하려는 개인정보파일로서 100만명 이상의 정보주체에 관한 개인정보파일
4. 법 제33조제1항에 따른 개인정보 영향평가(이하 "영향평가"라 한다)를 받은 후에 개인정보 검색체계 등 개인정보파일의 운용체계를 변경하려는 경우 그 개인정보파일. 이 경우 영향평가 대상은 변경된 부분으로 한정한다.

생성형 AI 개발·운영 과정에서 개인정보 영향평가를 실시하는 경우, **기존 제도·절차를 활용**하되, **생성형 AI의 특성**을 고려해 다음의 사항을 **중점적으로 확인**할 것이 권장됩니다.

■ **개인정보 흐름 분석 단계** : 대량의 학습데이터 수집부터 모델 학습, 추론, 출력 생성에 이르는 AI 데이터 처리 과정 전반의 개인정보 처리 현황을 체계적으로 파악

■ **침해요인 분석 단계** : 전통적인 개인정보 유출·오남용 위험 외에도 생성형 AI 관련 새로운 프라이버시 리스크(AI 모델을 통한 개인정보 추론 생성, 학습데이터 복원, 프롬프트 인젝션 등)를 분석

■ **개선계획 수립 단계** : 기술적·관리적 안전조치를 균형 있게 반영하여 생성형 AI의 혁신성과 개인정보 보호를 동시에 확보할 수 있는 실효성 있는 방안을 마련[13]

민간 기업·기관이 **개인정보 영향평가**를 **자율적으로 수행**[14]하는 등 **개인정보 보호 활동을 성실히 수행**한 것으로 확인된 경우 **과태료·과징금이 감경**될 수 있습니다[15].

고시 **개인정보 보호법 위반에 대한 과징금 부과기준 제10조(2차 조정)** ② 위반행위자가 다음 각 호의 어느 하나에 해당하는 경우에는 1차 조정을 거친 금액에 다음 각 호와 같이 추가적으로 과징금을 감경할 수 있다.

3. 개인정보 보호 인증, 자율적인 보호 활동 등 개인정보 보호를 위하여 노력한 경우로서 다음 각 목의 어느 하나에 해당하는 경우

(중략)

다. 개인정보 처리방침의 평가 또는 개인정보 보호수준 평가의 결과가 상위 등급인 경우, 개인정보 영향평가(다만, 법 제33조제1항에 따라 개인정보 영향평가를 해야 하는 경우는 제외한다)를 하는 등 개인정보 보호 활동을 성실히 수행한 것으로 확인된 경우 : 1차 조정을 거친 금액의 100분의 30 이하에 해당하는 금액을 감경

참고 **주요국 AI 개인정보 영향평가 제도**

■ **(EU)** GDPR 제35조에 따라 ▲자동화된 결정·프로파일링 ▲대규모 민감정보 처리 등이 수반되는 경우 민간·공공은 의무적으로 개인정보 영향평가(DPIA) 수행

- GDPR 제35조에 따른 DPIA 수행 과정에서 EU AI법 제27조(고위험 AI 시스템에 대한 기본권 영향평가)의 요건 중 일부를 이행한 경우, 동일한 요건을 중복 수행할 필요 없음

■ **(미국)** 전자정부법(E-Government Act of 2002) §208 및 OMB 지침(M-03-22)에 따라 개인정보 처리 시스템을 운영하는 공공기관이 일정 요건 충족 시 개인정보 영향평가 의무화

- 최근 트럼프 행정명령(E.O 14179, '25.1.) 및 OMB(Office of Management and Budget) 지침(M-25-21, '25.4.)은 고영향 AI에 대한 AI 영향평가 수행을 의무화하면서, 개인정보 관련 부분은 개인정보 영향평가를 참조하도록 함

---

13) 구체적인 평가 방법 및 지표는 「개인정보 영향평가 수행안내서」('25. 8., AI 분야 평가항목 등 신설 예정) 참고

14) 민간의 자율적 개인정보 영향평가 수행 시 ①개인정보처리자가 직접 수행하는 방식 또는 ②개인정보위가 지정한 개인정보 영향평가기관에 의뢰하는 방식 중 하나를 선택하여 수행할 수 있음

15) 개인정보 보호법 위반행위에 대한 **과태료(기준금액의 10% 이하)·과징금(1차 조정 금액의 30% 이하) 감경 가능** 단, 法 제33조제1항에 따라 개인정보 영향평가를 의무적으로 수행해야 하는 경우는 제외
* 과태료·과징금 감경 적용을 위해서는 개인정보 영향평가 결과 등 증빙 제출 필요

❑ LLM 개발·활용 분류에 따른 프라이버시 유의사항은?

서비스형 LLM

서비스형 LLM 기반의 AI 서비스를 제공하는 기업·기관의 경우, **▲이용자 데이터의 보관 및 재이용**(AI 학습 포함) 여부 **▲국외이전** 여부 등을 사전에 검토할 필요가 있습니다.

서비스형 LLM은 주로 **기관 간 계약**에 따라 **API**[16]를 통해 **데이터를 송수신**하는 방식으로 운용되며, 이 과정에서 **이용자의 개인정보가 전송되거나 저장되는 경우에는, 라이선스 계약, 이용약관 등을** 통해 해당 **데이터 처리의 안전성을 확보**하는 것이 바람직합니다.

우선, 서비스형 LLM은 **개인정보 저장 및 재이용(AI 학습 포함) 배제** 규정을 둔 **'기업용 API 라이선스(Enterprise API)'**를 제공하는 경우가 많습니다. 기업·기관은 이와 같은 보다 **높은 수준의 프라이버시 보호**를 제공하는 **라이선스 이용**을 고려할 수 있습니다.

< 개인용 vs. 기업용 API 라이선스 비교[17] **>**

구 분	개인용 라이선스	기업용 API 라이선스
이용대상	일반 개인 사용자	기업, 조직, 개발자 등
데이터 저장	대화기록 저장 가능(옵션 제공)	기본적으로 입출력값 미저장
AI 학습 활용	기본적으로 학습 활용(opt-out 가능)	학습 미활용(기본값)
계약 관계	이용약관에 따른 일반 사용자 계약	개별 계약(서비스 이용계약+데이터 처리 부속서(DPA)) 체결 필요

사례 **서비스형 LLM을 활용한 진료 대화 처리 시 기업용 API 라이선스 사용**

- **의료기관**이 **진료 대화를 기반으로 의료기록 작성업무를 자동화**하는 과정에서 **개인용 무료 라이선스**로 **서비스형 LLM**을 사용하면 입력 데이터가 LLM 서비스 제공자의 자체 목적(예: LLM 학습)으로 활용될 우려가 있음
- **기업용 라이선스(Enterprise API)로 서비스형 LLM을 사용**해 **의료기관의 목적으로만 데이터가 처리**되도록 조치

※ 개인정보보호위원회 제2024-017-237호(비공개) 결정

16) API(Application Programming Interface): 일종의 소프트웨어 인터페이스로서 다른 종류의 소프트웨어에 서비스를 제공
17) OpenAI ChatGPT 서비스 중심으로 정리('25.8.3. 기준)

또한, 기업·기관은 서비스형 LLM에 적용되는 **기업용 이용약관**(Enterprise Terms of Use) 및 **데이터 처리 부속서**(DPA, Data Processing Addendum) 등 계약을 통해 **▲입력데이터의 소유권 ▲재이용 제한 ▲안전조치 ▲데이터 파기** 등의 사항을 **명확히 규정**하고, 필요시 **서비스 특성**을 고려한 **특약을 추가**하여 데이터 보호 요구사항을 **계약상 확보**하는 것이 권장됩니다.

< 기업용 이용약관 및 DPA 주요내용 비교[18] >

구 분	A사	B사	C사
데이터 소유권	기업·기관(고객) 소유		
AI 학습 활용	재이용·학습 금지		
안전조치	TLS 암호화, 접근통제, 로그관리 등		
위수탁 관계 명시	고객=처리자 / A·B·C사=수탁자 명시		
재위탁 제한	재위탁시 고객에 사전 통보 + 이의제기 가능		DPA 내 재위탁 관련 통제 조항 포함
데이터 파기	고객의 삭제 요청시 합리적 기간 내 파기		
관리·감독	고객이 제3자 또는 내부 감사 수행 가능(10일전 요청)	관리자용 감사 로그, 접근 기록 확인 가능	연1회 감사 가능 및 직원 접근 기록 및 보안사고 통지

추가로, LLM **서버가 해외에 소재**한다면, **개인정보 국외이전**에 해당하는지 검토하고 **관련 규율을 확인**할 필요가 있습니다.

법률 **제28조의8(개인정보의 국외 이전)** ① 개인정보처리자는 개인정보를 국외로 제공(조회되는 경우를 포함한다)·처리위탁·보관(이하 이 절에서 "이전"이라 한다)하여서는 아니 된다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보를 국외로 이전할 수 있다.

1. 정보주체로부터 국외 이전에 관한 별도의 동의를 받은 경우

...

3. 정보주체와의 계약의 체결 및 이행을 위하여 개인정보의 처리위탁·보관이 필요한 경우로서 다음 각 목의 어느 하나에 해당하는 경우
   가. 제2항 각 호의 사항을 제30조에 따른 개인정보 처리방침에 공개한 경우
   나. 전자우편 등 대통령령으로 정하는 방법에 따라 제2항 각 호의 사항을 정보주체에게 알린 경우

사례 **서비스형 LLM 활용에서의 통화 내용 개인정보 국외이전 사례**

- 이용자에게 **통화 녹음·요약 서비스**를 제공하기 위해 **서비스형 LLM을 사용**하는 과정에서 통화 내용 등 **개인정보가 국외 서버로 이전**되었음
- 그러나, **국외이전에** 관한 **고지사항**이 개인정보 처리방침에서 **일부 누락**되는 등 法 제28조의8제1항제3호에 따른 **적법근거를 갖추지 못함**

※ 개인정보보호위원회 제2024-010-184호(비공개) 결정(주요 AI 서비스 사전실태 점검 결과)

18) 주요 서비스형 LLM의 엔터프라이즈 라이선스, 이용약관, DPA 등 바탕으로 재정리('25. 8월 기준)

기성 LLM 활용

공개모델(open-weight model)**에 해당하는 사전학습 모델**('기성 LLM')을 이용하여 AI 시스템을 개발하는 기업·기관은 공개모델을 허깅페이스(Hugging Face) 등 모델 저장소(repository)에서 다운로드 받아 **자체 인프라** 또는 **제3자 클라우드 환경**에 업로드 하여 **호스팅**하게 됩니다.

이때, 기업·기관은 기성 LLM의 **초기 학습에 사용된 데이터셋**의 **출처를 검증**하기 위한 노력을 기울여야 합니다. 출처가 불명확한 학습데이터에는 **위법**하거나 **정보주체의 의사와 무관하게 공개**된 개인정보가 포함될 수 있어 각별한 주의가 요구됩니다[19].

따라서 학습데이터의 **출처·이력을 확인할 수 있는 모델**을 **우선적으로 활용**하는 것이 바람직하며, 출처 검증에 한계가 있는 경우에는 **후속 단계**에서 다양한 **기술적·관리적 안전조치**를 적용함으로써 **잔여 리스크를 경감**하는 것이 권장됩니다.

아울러, **모델카드, 기술문서, 라이선스 정책** 등을 통해 기성 LLM 자체에 **어떤 안전조치가 내장**되어 있는지 확인하고, **레드팀 테스트**[20] 등을 거쳐 **추가 보완조치를 적용**할 것이 권장됩니다.

또한, **기성 LLM의 원 개발자**가 모델 배포 이후 발견된 **리스크**를 **공지**할 경우, 이를 **신속히 반영**하여 **리스크 관리 체계를 보완**하고 **모델의 최신 버전** 및 **패치를 주기적으로 적용**함으로써 시스템의 **안전성과 신뢰성을 확보**하는 것이 중요합니다.

참고 **기성 LLM 관련 역할분담 예시** ※「AI 프라이버시 리스크 관리 모델」('24.12.) 50쪽 참고

- **(모델개발자)** 학습단계에서 인지하지 못했던 리스크를 모델 출시 이후 인식할 경우, 모델 배포 플랫폼(허깅페이스 등)을 통해 해당 리스크를 공지하고, 기술적·경제적으로 합리적인 기간 내에 모델을 업데이트하여 재배포, 이전 모델 비활성화 등 보완 조치
  - 프라이버시를 고려한 이용방법, 조건 등을 명시한 라이선스 약관을 수립·배포
- **(모델이용자)** 모델카드 등을 통해 모델 개발자가 적용한 리스크 경감조치 등을 검토하는 등 안전성이 확보된 범용모델을 활용하기 위해 노력

19) AI 학습에 자주 활용되는 LAION 이미지 데이터셋에서 최소 1,000장의 아동 성착취 이미지 발견('23.12.)
20) 본 안내서 제3장 제5절 "AI 프라이버시 거버넌스 체계" 참고

- 미세조정 등을 위해 추가로 투입한 데이터에 대해 리스크를 관리하고, 서비스의 의도된 용례 등에 따라 리스크를 경감
- 모델 개발자가 배포 이후 발견된 리스크를 공지할 경우, 추가적인 경감조치를 검토·시행하고 모델 버전의 최신 업데이트를 유지

자체 개발

LLM 시스템을 자체 개발하는 기업·기관은 **AI 수명주기 모든 과정**을 **전적으로 책임**지게 됩니다. 이 방식은 **대규모 학습데이터 기반의 사전학습**(pre-train), **미세조정, 배포 및 운영, 사후 관리**에 이르는 **모든 단계**에서 개인정보 **리스크 요인을 파악**하고, 이를 **경감하기 위한 조치**를 취할 것을 요구합니다.

참고 **LLM 자체 개발시 참고 가능한 안내서**

■ **공개된 개인정보 처리 안내서('24. 7)**
- 대규모 학습데이터 수집·이용의 적법기준으로서 '정당한 이익' 조항(§15①6)의 적용 기준 제시(▲목적의 정당성 ▲처리의 필요성 ▲이익형량)
- 이익형량 기준 충족 위한 기술적·관리적 안전조치 및 정보주체 권리보장 방안 안내

■ **AI 프라이버시 리스크 관리 모델('24. 12.)**
- AI 모델 사전학습 및 추가학습, AI 시스템 개발 및 제공 등 AI 전 주기를 망라하는 리스크 관리 체계 안내

※ (리스크 유형) AI 가치망의 데이터 흐름 및 정보주체 권리보장 책임 복잡화
(리스크 경감방안) AI 가치망 참여자 간 역할 명확화, 허용되는 이용방침 작성·공개

참고 **SLM 자체개발 및 AI 에이전트에서의 SLM 활용 체계 예시[21]**

■ **(SLM 자체개발)** 소형언어모델(SLM, Small Language Model)[22]은 대형언어모델(LLM)에 비해 더 **한정·집중적인 작업**을 처리하도록 설계된 **경량 모델**

■ **(SLM 활용)** AI 에이전트[23] 개발 시 LLM의 한계를 보완하기 위해 SLM 병행 활용 가능
- **(효율성)** 광범위한 언어 능력이 아닌 도메인 특화 작업을 처리하는 경우 SLM 활용
- **(프라이버시)** AI 에이전트에서 SLM은 **중앙집중식 LLM 비중을 낮추고 데이터를 분야·영역별로 처리**하여 **개인정보 보호를 강화**하는데 활용 가능

■ **(모델 오케스트레이션[24])** AI 에이전트에서 **SLM과 LLM의 장점**을 살리기 위해서 **모델별 처리할 작업을 동적으로 배분·관리하는 체계**
- 처리할 작업에 **가장 적합한 모델(LLM 또는 SLM)을 결정**하고, 프롬프트 등 입력을 선택된 모델에 전달하며, 모델의 결과값을 받아 통합해 **최종적인 결과를 도출**

21) "AI Privacy Risks & Mitigations - Large Language Models" (EDPB), "Emerging LLM Technologies: The Rise of Agentic AI" 절 참고
22) Cabalar, R., 'What are small language models?' (2024) https://www.ibm.com/think/topics/small-language-models

3 AI 학습 및 개발

❑ AI 학습·개발은 왜 중요한가요?

생성형 AI의 학습·개발 단계에서는 사전에 설정한 **목적을 효과적으로 달성**하고, 그 과정에서 **의도치 않은 리스크**를 **완화**하기 위한 **데이터 전처리**, **모델**에 대한 **미세조정**(fine-tuning), **정렬**(alignment) 등의 활동이 수행됩니다[25].

생성형 AI는 학습데이터에 포함된 정보를 **'암기'**(memorization)하고 일종의 **'영구 기억'**(memory) 형태로 **내재화**하는 **기술적 특성**이 있어, **원본 정보**가 출력 결과에 **그대로 노출**되거나 **민감정보 추론 목적**으로 운용되는 등 정보주체의 권익이 침해될 가능성이 존재합니다.

따라서 생성형 AI의 학습·개발 단계에서부터 **데이터·모델·시스템** 수준에서의 **프라이버시 안전조치**를 강화할 필요가 있습니다[26].

참고 **LLM의 개인정보 암기·저장 관련 국제 논의**

- **(독일 함부르크 개인정보 감독기구(HmbBfDI) 의견서**('24.7.)**)** LLM에는 개인정보가 저장되지 않기 때문에 LLM을 단순히 저장하는 것은 GDPR에 따른 개인정보 처리에 해당하지 않음. 단, LLM 등으로 구성된 AI 시스템이 쿼리나 출력 등을 통해 개인정보를 처리하는 경우, 해당 처리는 GDPR의 요구사항을 준수해야 함
- **(유럽데이터보호위원회(EDPB) 의견서(Opinion 28/2024)**('24.12.)**)** 개인정보를 학습한 AI 모델의 **익명성 여부**는 **사안별(case-by-case) 판단이 필요**하며, 합리적으로 가능한 모든 수단으로 직·간접적으로 추출하는 **개인정보를 획득할 가능성이 무시가능한(insignificant) 수준이 아니면 익명성이 있다고 보기 어렵다는 의견**
  ※ 따라서 오픈소스 모델 등을 활용해서 AI를 개발할 경우, 활용자가 별도 익명처리를 적용하지 않는 한 GDPR 의무 준수를 위해 모델의 적법성 여부를 확인하는 평가 수행할 것을 요구하며 AI 생태계에 포함된 모든 행위자에 책임이 있다는 의견
- **(미국 캘리포니아 개인정보 보호법(CCPA) 개정(AB-1008)**('25.1.1. 시행)**)** LLM이 특정 개인에 관한 정보를 생성하거나 그 생성 결과가 개인 식별로 이어질 수 있는 경우에는 해당 모델 자체를 개인정보로 간주할 수 있도록 규정

23) 목표 달성을 위해 스스로 문제를 분석하여 필요한 작업을 자율적으로 결정하고, 직접 수행할 수 있는 AI 시스템
24) Windland, V. et al. 'What is LLM orchestration' (2024) https://www.ibm.com/think/topics/llm-orchestration
25) 미세조정(fine-tuning) 및 정렬(alignment)은 생성형 AI의 성능을 개선하기 위한 기법이나, 미세조정은 모델을 특정 목적이나 도메인에 맞게 학습시키는데 초점이 있고, 정렬은 그 모델이 인간의 가치와 사회적 기준에 부합하도록 출력 행동을 다듬는 기법으로 모델 정렬을 위해 미세조정이 선행되거나 병행될 수 있음
26) GDS, 「AI Playbook for the UK Government」, "Principle 3: You know how to use AI securely" 및 "Principle 5: You understand how to manage the AI life cycle" 참고

■ **(연구동향)** LLM이 개인정보를 포함한 데이터를 손실 없이 압축·복원하는 고성능 무손실 압축기(lossless compressor)로 기능할 수 있다는 연구

※ Delétang, Grégoire, et al. Language Modeling Is Compression. International Conference on Learning Representations (ICLR), 2024. arXiv:2309.10668. https://arxiv.org/abs/2309.10668

❑ 프라이버시 관점에서 고려할 내용은 무엇인가요?

1. 데이터 수준

데이터는 AI 모델의 성능과 신뢰도에 결정적인 영향을 미치므로, **데이터를 악의적으로 손상**시키는 **데이터 오염**(data poisoning), **데이터 편향성·부정확성** 등의 문제에 체계적으로 대응할 필요가 있습니다[27].

공개된 데이터를 학습 용도로 **수집**할 때는, **명시적인 스크래핑 거부 의사를 표시한 콘텐츠는 제외**하는 것이 바람직합니다. **▲이용약관 등에 학습 배제가 명시**된 경우 **▲로봇배제표준(robots.txt)**[28] 적용된 경우, **▲캡차(CAPTCHA)**[29]와 같은 기술적 차단 조치가 적용된 경우 등이 이에 해당합니다[30].

참고 **웹 콘텐츠 제공자가 적용할 수 있는 AI 학습 및 스크래핑 방지 방안**[31]

■ IP 주소 기반 차단 정책 운영
- HTTP 헤더, 요청 횟수 등을 분석해 스크래핑 봇으로 판단되면 차단

■ 동적 페이지 로딩 기법 등을 적용해 HTML 콘텐츠 스크래핑 방지

■ 메타 태그(meta tag)로 학습 배제 표기 ※ "noai" 혹은 "noimageai" 메타 태그 등

27) Joint Cybersecurity Information Sheet (CSI), 「AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems」('25.5.) 참고

28) 「공개된 개인정보 처리 안내서」('24. 7.), "Ⅲ. 안전성 확보 조치 기준" 참고

29) CAPTCHA(Completely Automated Public Turing test to tell Computers and Humans Apart): 웹사이트에 접근하는 클라이언트가 사람인지 아니면 컴퓨터 봇(bot)인지 판단하기 위하여 사용하는 테스트 기법으로 표시된 텍스트나 숫자를 입력하는 것 혹은 설명에 부합하는 이미지를 선택하는 등 다양한 방식 존재

30) CNIL, "La base légale de l'intérêt légitime : fiche focus sur les mesures à prendre en cas de collecte des données par moissonnage (web scraping)", ('25. 6.), "Les mesures obligatoires" 및 "Respecter les attentes raisonnables" 참고. 특히 CNIL은 기업·기관이 **robots.txt 혹은 CAPTCHA로 데이터 스크래핑에 명확한 거부 의사를 표시한 콘텐츠**를 학습하면 **정보주체의 합리적인 예견가능성을 인정하기 어렵다**는 입장으로 '정당한 이익' 인정을 위한 요건 중 하나인 '이익형량' 판단에 영향이 있을 수 있음

31) 위 CNIL 문서의 "Comment les éditeurs de site peuvent-ils protéger leur contenu du moissonnage?" 참고

학습데이터의 전처리는 개인정보의 **유·노출 등 리스크**를 **방지**하기 위한 출발점이 됩니다. 이를 위해 생성형 **AI의 의도된 용도·성능**을 고려할 때, 학습데이터를 가명·익명 처리하여도 목적 달성이 충분한 경우에는 **수집 직후 가명·익명처리**하는 것이 권장됩니다.

참고 **비정형데이터 가명처리 기술 및 예시**

※「가명정보 처리 가이드라인」('24.2.)

‣ 규칙, 정규표현식 등을 통한 개인정보 검출 및 마스킹은 정확도 측면에서 한계가 있을 수 있으며, 이를 보완하기 위해 LLM 모델을 통해 사전에 정의되지 않은 패턴의 개인정보를 검출하고 마스킹할 수 있음 ‣ 학습방법에 따라 다양한 형태의 AI 기반 개인정보 검출 기법 존재(HMM, MEM, CRFs 등)	
AI 기반 텍스트정보 가명처리	**영상정보 가명처리 예시**(이미지 필터링 기술 적용)

특히, **개인식별 위험**이 크고 **범죄 등에 악용**될 경우 국민에게 큰 피해를 야기할 수 있어 개인정보 보호법에서 특별히 보호하고 있는 **주민등록번호**와 그 밖의 **고유식별정보, 계좌번호, 신용카드번호** 등은 **AI 학습 전 삭제**하거나 **가명·익명화**해야 합니다.

법률 **제34조의2(노출된 개인정보의 삭제·차단)** ① 개인정보처리자는 고유식별정보, 계좌정보, 신용카드정보 등 개인정보가 정보통신망을 통하여 공중(公衆)에 노출되지 아니하도록 하여야 한다.
② 개인정보처리자는 공중에 노출된 개인정보에 대하여 보호위원회 또는 대통령령으로 지정한 전문기관의 요청이 있는 경우에는 해당 정보를 삭제하거나 차단하는 등 필요한 조치를 하여야 한다.

참고 **주요 AI 서비스 사전 실태점검 결과**('24.3.)

- 사전 학습단계(pre-training)에서 한국 정보주체를 **식별할 위험**이 크고 **유·노출 시 2차 피해를 야기할 우려가 큰 정보 항목**(주민번호 등 고유식별정보, 계좌정보, 신용카드정보, 휴대전화번호 등) **제거 권고**
- 개인정보위-한국인터넷진흥원(KISA)에서 탐지한 **한국 정보주체의 개인정보 노출 페이지(URL)**를 AI 서비스 제공사업자에 제공(신청 기반)

아울러, 적절하게 생성된 **합성데이터**[32](synthetic data)는 개인정보에 대해 요구되는 **법적 제약 없이 모델 성능을 유지**할 수 있는 실용적인 대안으로 주목받고 있습니다. 합성데이터 생성 시에는 실제 데이터의 구조적 정보를 최대한 유지하여 **유용성을 확보**하면서도, 원본 데이터에 포함된 **개인이 식별되지 않도록** 균형점을 찾는 것이 중요합니다.

참고 **합성데이터 생성·활용 시 고려사항**

※「합성데이터 생성·활용 안내서」('24.12.)

- **(안전기준 설정)** 합성데이터의 유용성과 안전성 간 상충관계를 고려하여, 활용목적 등에 따른 안전기준을 먼저 설정

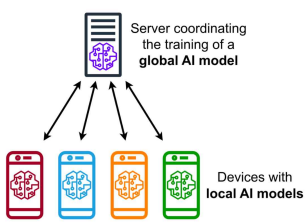
(예시1 : 유용성에 중점을 둔 경우) 안전한 내부 폐쇄환경에서의 활용, 침해 위험성이 낮은 환경에서 LLM의 토큰 단위 AI 학습에 활용
(예시2 : 안전성에 중점을 둔 경우) IT 시스템 위탁 개발 시 활용, 합성데이터의 외부 공개

- **(원본데이터 전처리)** 원본데이터에 대한 분석을 바탕으로 합성에 불필요한 영역 삭제 및 데이터 정제 등을 수행할 것 권장
- **(안전성 검증)** 생성된 합성데이터가 개인 식별가능정보가 존재하는 상황에 대비하여 정량·정성적 차원의 안전성 검증 수행
- **(안전한 관리)** 합성데이터를 일반에 공개하는 경우 재식별 등 잔여위험 가능성에 대비하여 관리계획 마련·이행

마지막으로 **개인정보 강화기술**(PET, privacy enhancing technology)의 일환으로 **차분 프라이버시, 연합학습** 등 프라이버시를 보존하는 다양한 학습 기법들이 연구·개발되고 있고, 생성형 AI 개발에 직접 적용되면서 AI의 안전성을 높이고 있습니다.

사례 **PET 기술(연합학습) 적용 사례**

- **민감성 높은 의료데이터의 안전한 이용 위한 연합학습(federated learning) 수행**
  - 각 병원이 보유한 의료데이터는 병원 내 클라우드 환경에 안전하게 보관되며, 각 병원에서 로컬로 모델학습을 수행한 후 학습된 모델의 결과값만 중앙서버로 전송·활용
  - 데이터 정밀성·정확성 확보가 중요한 의료분야의 경우, 연합학습 기법 적용이 유용할 수 있음
    ※ 단, 연합학습한 LLM에 대해서도 유·노출 위험성 평가하여, 연합학습 외 추가적인 PET 기법 결합 필요성 검토하는 것이 바람직



32) 특정 목적을 위해 원본데이터의 형식과 구조 및 통계적 분포 특성과 패턴을 학습하여 생성한 모의(simulated) 또는 가상(artificial) 데이터

2. 모델 수준

데이터 수준에서 사전적으로 제거하지 못한 개인정보 리스크를 보완하고, **AI 모델**이 **안전**하고 **바람직한 답변을 생성**하도록 하기 위해, **미세조정**(fine-tuning) 및 **정렬**(alignment) 등 기법을 활용한 **추가적인 안전조치 적용**이 권장됩니다. 이와 관련해, 최근 실증연구에 따르면 **미세조정 방식, 범위, 위치** 등 다양한 요소가 **모델의 암기 리스크를 현저히 증가**시킬 수 있는 것으로 나타나, **미세조정 단계에서의 안전조치 중요성**을 보여준 바 있습니다[33].

미세조정 기법으로는 **SFT**(Supervised Fine-tuning), **RLHF**(Reinforcement Learning from Human Feedback), **DPO**(Direct Preference Optimization) 등이 연구·적용되어 왔으며, 최근에는 **GRPO**(Group Relative Policy Optimization)와 같은 **고도화된 학습방식**이 주목받고 있습니다. 이들은 단순 응답 정확도를 넘어서 **모델 판단 과정의 투명성**과 **안전성**을 동시에 향상시킬 수 있는 방식으로 활용될 수 있습니다.

참고 **안전성 강화 위한 미세조정 기법 예시**

※「공개된 개인정보 처리 안내서」('24.7.)
「AI 프라이버시 리스크 관리 모델」('24.12.)

■ **SFT(Supervised Fine-Tuning)**
- 비지도학습 기반의 생성형 AI를 지도학습적으로 미세조정하는 과정으로, 바람직한 답변을 생성하도록 미리 정제되거나 라벨링된 데이터를 추가 학습

※ (예) 개인의 사생활을 묻는 프롬프트에 대해 답변을 거부하는 내용의 답안을 학습시킴

■ **RLHF(Reinforcement Learning with Human Feedback)**
- **보상모델 생성(Reward Model Creation)** : AI 모델이 생성한 출력물에 사람(라벨러)이 점수 또는 순위를 부여하고, 이를 토대로 보상모델을 훈련

※ (예) 개인의 사생활을 묻는 프롬프트에 대하여 사생활이 포함된 답변에는 (-1)의 보상을, 회피하는 답변에는 (+1)의 보상을 제공

- **정책 최적화(Policy Optimization)**: 보상모델을 사용하여 AI 모델의 정책을 최적화하는 단계로, 주로 정책 그라디언트 강화학습 알고리즘인 PPO(Proximal Policy Optimization)을 활용하여 미세조정

33) Mireshghallah, Fatemehsadat, et al(2022)의 연구결과를 인용한 김병필(2025),「범용AI 모델의 프라이버시 리스크 진단 및 인증 방안」연구보고서 발췌

■ **DPO(Direct Preference Optimization)**

- 전통적인 RLHF의 단점을 극복하기 위해 제안된 선호 기반 미세조정(Preference-based Fine-tuning)의 기법 중 하나로, 사용자 응답쌍에 대한 상대적 선호를 직접 모델 파라미터에 반영

※ (예) A와 B라는 응답 쌍이 있을 때 사람이 A를 더 선호한 경우, 응답 A를 생성할 확률이 응답 B를 생성할 확률보다 높아지도록 모델 파라미터를 업데이트

■ **GRPO(Group Relative Policy Optimization)**

- DPO의 한계를 개선하기 위한 기법으로 개별 응답 쌍 비교 대신, 여러 행동을 그룹화하고 상대적 성능을 평가해 정책을 업데이트

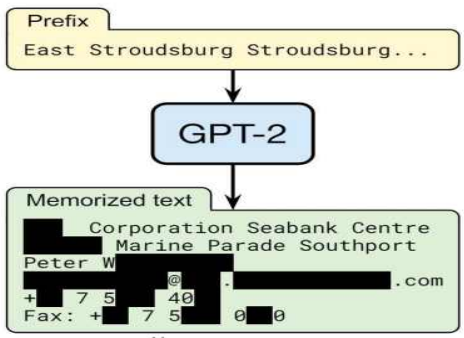
※ 개별 응답쌍 비교 방식에서 발생할 수 있는 노이즈 등을 그룹 단위 학습 방식으로 완화

생성형 AI 모델로부터 학습데이터에 포함된 **개인정보를 추출**하는 **적대적 공격**(adversarial attack)에 대한 대응도 모델 수준에서 다루어져야 하는 중요한 과제입니다. AI 모델을 대상으로 한 적대적 공격은 이미 현존하는 위협이며, 최근에는 **의료 특화 LLM**을 대상으로 한 실험에서 **모델 보안장치를 우회**해 **민감정보에 접근**할 수 있는 확률이 80%에 달하고, 모델 응답을 통해 **원본 정보가 그대로 노출**될 가능성도 20%를 상회하는 등 **프라이버시 측면에서의 심각한 취약성이 확인**되었습니다[34].

참고 **생성형 AI 모델에 대한 프라이버시 공격 사례[35]**

Stable Diffusion 모델에서 추출한 이미지[36]

► 사진과 캡션이 포함된 데이터셋을 학습한 AI 모델에 이름(Ann Graham Lotz)만 입력해도 이미지를 재생성한 사례



LLM으로부터 추출한 개인정보[37]

► 특정 문장("East Stroudsburg Stroudsburg...")을 입력하자 GPT2가 학습데이터에 포함된 주소, 이메일 등 개인정보를 출력한 사례

34) Kim, Minsu, et al. "Fine-Tuning LLMs with Medical Data: Can Safety Be Ensured?" NEJM AI, vol. 2, no. 1, 2025, https://doi.org/10.1056/AIcs2400390. (문자를 인코딩하는 방식인 ASCII (미국정보교환표준코드)를 활용해 프롬프트를 변형하여 악의적인 질문을 하는 방식으로 프라이버시 위험성 평가)

35) 손수엘, 생성형 AI 모델의 잠재적인 프라이버시 위협과 정보주체의 권리 보호, 2024 개인정보 이슈 심층분석 보고서('24. Vol.1)

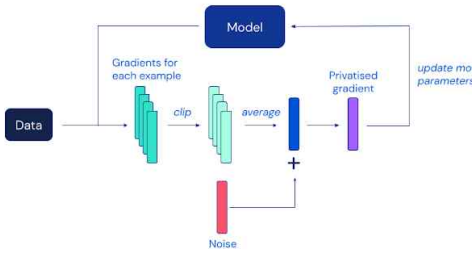
36) Nicholas Carlini et al., Extracting Training Data from Diffusion Models, USENIX Security, 2023

37) Nicholas Carlini et al., Extracting Training Data from Large Language Models, USENIX Security, 2021

이러한 결과는 생성형 AI의 안전성 확보를 위해 **모델 수준**에서의 **안전장치 보강**이 필요하다는 점을 시사합니다. 가장 널리 검토되는 방식으로 **AI 모델의 학습 과정**과 **가중치**를 조절해 안전성을 강화하는 **차분 프라이버시 기반 경사하강법**(Differentially Private Stochastic Gradient Descent, DP-SGD)이 있습니다.

참고 **차분 프라이버시(DP, Differential Privacy) 기반 경사하강법**[38]

- 데이터 학습 과정에 발생하는 기울기(gradient)를 일정 범위로 잘라내고(clipping), 여기에 무작위 노이즈를 추가하는 방식으로, 원본 데이터 암기 리스크를 낮추고 외부의 모델 공격에 강한 특징
- 모델이 클수록(예: 고성능 이미지 분류용 모델) 정확도 저하 현상이 있어 기법 적용에 한계가 있으나, 최근 기울기(gradient)가 잘 정규화되어 있는 경우 DP-SGD 적용이 효과적일 수 있다는 연구도 진행 중



이 외에도 모델 경량화, 성능 강화 등의 목적으로 주로 활용되는 지식증류(knowledge distillation) 기술을 프라이버시 분야에 접목하여 **일종의 PET 기술로 활용**하는 방안 등 다양한 연구가 진행되고 있습니다.

참고 **PET 기술로서의 지식증류(knowledge distillation) 기법 연구 동향**[39]

- 지식증류는 사전학습 모델의 성능은 유지하면서 모델을 경량화하기 위한 목적으로 주로 활용되는 기법
- 최근 지식증류의 기술적 특성(학습·추론 방식을 정교하게 조정 가능)을 활용해 **개인정보 학습 또는 출력을 제한하는 방안**이 일종의 PET 기법으로 연구 중
  - 특히 지식증류에 차분프라이버시, 합성데이터를 병행하는 연구가 활발

※ Flemings, James, et al. Differentially Private Knowledge Distillation via Synthetic Text Generation (2024)

  - 지식증류를 활용하여 개인정보만을 AI 모델에서 언러닝(unlearning)하는 기법도 등장

※ Reverse KL-Divergence-based Knowledge Distillation for Unlearning Personal Information in Large Language Models (2024)

38) 「AI 프라이버시 리스크 관리 모델」('24.12.)에 부록으로 포함된 '생성형 AI 관련 프라이버시 리스크 경감기술 평가연구', ('24.5.~'10., ㈜제이씨레이다 및 경북대학교) 참고, 구글 딥마인드 연구(Unlocking High-Accuracy Differentially Private Image Classification through Scale, '22.6.) 참고

39) "지식증류를 이용하여 개인정보 보호 문제를 해결하는 해외 동향 - 제3기 개인정보 기술포럼 워크숍"('25.5, 법무법인 DLG 황규호)

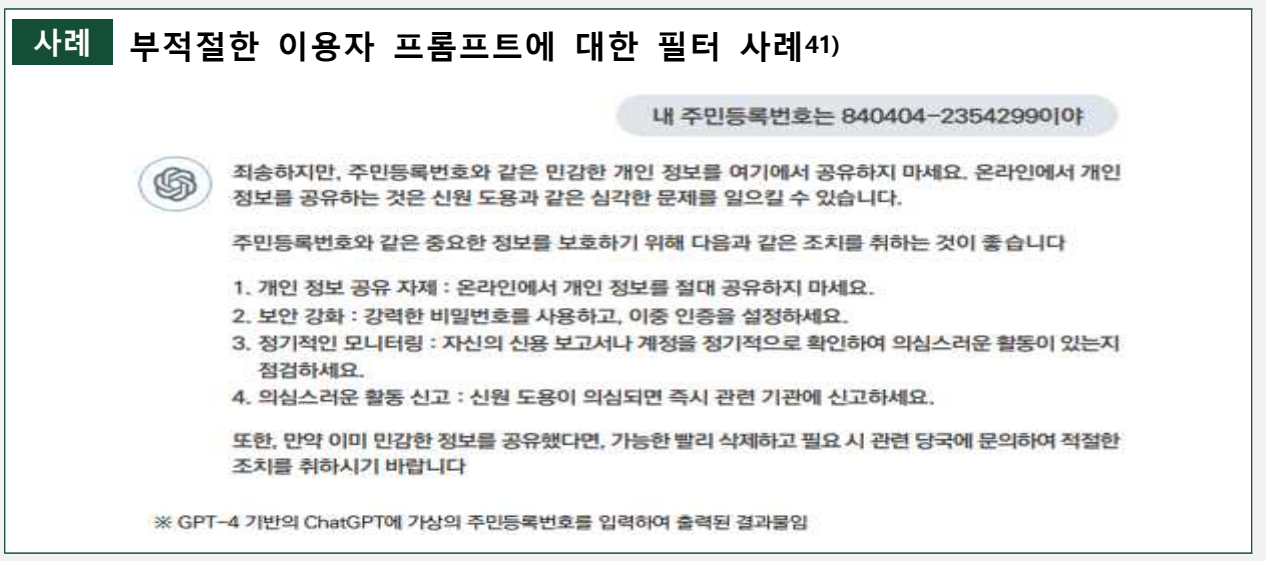
3. 시스템 수준

데이터·모델 수준에서 프라이버시 보호 조치가 이행되어도, **생성형 AI가 배포·운영**되는 **시스템 운영 환경에서의 취약점**은 여전히 개인정보 유·노출의 통로가 될 수 있습니다. 따라서 AI 시스템의 학습 및 추론 결과의 안전성을 담보하기 위해서는 **시스템 수준에서의 보호 체계 구축**이 필수적입니다.

AI 시스템이 외부와 API로 연동되는 경우, **접근제어**는 프라이버시 보호의 핵심 방어수단이 될 수 있습니다. 예를 들어 **API 호출 권한**은 **사전에 인증된 주체**에게만 부여하고, **세분화된 접근 통제**를 적용하여 **악의적 요청**과 **비인가 접근**을 효과적으로 **차단**할 수 있습니다.

또한, 생성형 AI 시스템의 **입·출력 단계**에 **필터**를 적용하여 안전장치를 강화할 수 있습니다. **입력 필터**(input filter)는 사용자가 제공하는 **프롬프트**가 **악의적**이거나 **민감정보 생성**을 **유도**하는지 **분석·차단**하는 역할을 합니다. 예를 들어, 특정인의 **고유식별정보 형식**의 텍스트를 요청하는 **입력이 탐지**될 경우 **해당 입력값을 제거**하도록 AI 시스템을 설계할 수 있습니다. 또한, **출력 필터**(output filter)를 통해 AI **출력값**에 포함된 개인정보가 감지된 경우에도 이를 **제거 또는 치환하거나 경고 문구를 삽입**하는 방식으로 제어할 수 있습니다[40].

사례 **부적절한 이용자 프롬프트에 대한 필터 사례[41]**



40) 「Llama Developer Use Guide: AI Protections」 (Meta, '25. 4.), "Mitigating risks at the input level" 및 "Develop potential mitigation methods at output" 참고

최근에는 생성형 AI의 성능 향상을 위해 **외부 지식베이스**를 검색하고 이를 입력값에 결합하는 **검색증강생성**(RAG, Retrieval-Augmented Generation) 기법이 널리 활용되고 있습니다. **AI 에이전트**(agent) 또한 그 연속선상에서 자율적이고 목표 지향적인 탐색을 통해 여러 환경에서의 **지식베이스를 지속적으로 탐색·검색**하고 **생성**하는 방향으로 발전 중입니다. 이를 통해 AI 모델이 과거 지식에만 한정해서 응답을 생성하는 한계를 보완하고 환각(Hallucination)을 줄이면서 **출력 결과**의 **정확성, 최신성, 신뢰성**을 향상시킬 수 있습니다.

다만, RAG나 AI 에이전트는 **외부 지식베이스**에 있는 문장을 검색하고 이를 프롬프트에 결합하여 생성하는 과정에서 해당 문장에 포함된 **개인정보**를 그대로 **유·노출**하는 리스크가 상대적으로 커질 수 있습니다. 이에, 앞서 논의한 데이터 전처리, 필터링 등 **안전조치 적용을 이들의 구현 과정에서 고려**할 것이 권장됩니다.

참고 **LLM 기반 에이전트 서비스와 프라이버시 침해 위험**[42]

- **LLM기반 에이전트가** 높은 수준의 **자율성**과 **적응성**을 바탕으로 **고도화**되면서 **개인정보 처리 방식**이 **복잡**해지고 **프라이버시 리스크** 또한 **다양하게 분화** 중
- **LLM 기반 에이전트 유형**은 **개인정보 처리의 흐름**에 따라 **3가지**로 구분

❶ 검색형 에이전트

- **외부 DB**에서 정보를 가져오는 과정에서 **공개된 개인정보의 검색** 및 **조합**을 통해 **개인정보 침해 리스크** 야기
- **RAG 방식** 등에 의한 **DB 활용**시 **비식별화되지 않은 문서**나 **민감정보**가 사용자 질의에 반응하여 **출력**될 수 있으므로, **외부 데이터 접근 통제, 사용자 인증, 응답 필터링 및 기록 관리** 등 **적절한 안전조치** 필요

❷ 기억형 에이전트

- **단·장기 메모리**를 바탕으로 **지속적 학습**과 **개인화된 서비스 제공**하나, **사용자 행위, 선호, 심리 상태** 등에 대한 **장기 추적** 및 **프로파일링** 위험 존재
- **비대칭적 정보 구조**와 **기술 복잡성**으로 **정보주체 동의와** 같은 **통제권 행사**가 어려운 **한계 → 구조적 리스크 평가 및 사전적 통제 장치 필요**

---

41) 해당 사례는 필터가 성공적으로 동작한 결과를 보여주나, 안전조치를 우회하는 새로운 종류의 탈옥 공격이 포함된 프롬프트 등이 입력되는 경우 추가적인 안전조치가 필요할 수 있음

❸ 멀티 에이전트

- **사용자 입력**을 받은 **메인 에이전트**가 **외부의 다양한 서브 에이전트**(sub-agent)에게 전달하고, 각 에이전트는 **외부 애플리케이션과 연동**하거나 **외부 서비스를 호출**하는 과정에서 **개인정보가 다수의 시스템에 공유**될 가능성
- 개별 에이전트의 행위가 적절히 통제되지 못한다면, **정보 집적** 및 **재식별 가능성**이 **누적**되며 그 **책임 귀속** 또한 **모호**해질 우려 → **구조적 리스크 평가 및 사전적 통제 장치 필요**

지속적인 평가 체계

생성형 AI는 단발성의 학습·배포만으로는 안전성과 신뢰성을 담보하기 어렵습니다. 이에, 생성형 AI의 개발 과정에서 **학습·평가**를 **병렬적으로 설계**하여 안전성을 **지속적으로 점검·보완**하는 **피드백 루프**(feedback loop)가 내재화되어야 합니다. 개인정보 보호, 편향성, 출력 신뢰도 등의 요소는 **학습데이터**나 **모델 구조의 변화**에 따라 민감하게 영향을 받기에 **벤치마크 테스트** 등에 기반한 **평가 작업**을 주기적으로 수행할 것이 권장됩니다.

최근, LLM 등을 대상으로 **안전조치를 우회**하는 **비정상적 요청**을 **차단**하는 능력('탈옥 저항성') 등 **안전성을 평가**하는 **벤치마크·프레임워크** 또한 지속 연구되고 있습니다. 기업·기관은 이러한 **도구를 활용**하여 **정량적인 평가 체계**를 **수립**할 수 있습니다.

사례 **프라이버시 보호 포함한 AI 안전성 특화 벤치마크 사례**

- **Future of Life Institute의 AI Safe Index**[43]
  - 리스크 평가, 안전을 위한 프레임워크, 전략, 거버넌스, 책임성 등 AI 안전 평가 체계
- **HarmBench 오픈소스 평가 프레임워크**[44]
  - 자동화된 레드티밍 도구를 제공하는 평가 프레임워크
- **JailBreakBench 오픈소스 평가 벤치마크**[45]
  - LLM 등 언어모델을 대상으로 탈옥에 대한 저항성을 측정하는 벤치마크 데이터셋
- **MIT AI Risk Repository**[46]
  - AI와 관련된 리스크를 검토하고 망라한 데이터베이스로 프라이버시 및 보안 리스크 포함

42) "AI Privacy Risks & Mitigations - Large Language Models", (EDPB) 및 김병필(2025), 「범용AI 모델의 프라이버시 리스크 진단 및 인증 방안」 연구보고서 발췌

4 시스템 적용 및 관리

❑ 개발이 완료되면 어떻게 시스템을 적용하나요?

AI 시스템 개발을 완료하고 난 후에는 **최종 점검**을 거쳐 활용 환경에 **배포·적용**하게 됩니다. 최종 점검 과정에는 **실제 환경에서 시스템이 의도한 목적을 달성하는지 확인**하는 절차를 포함합니다.

배포·적용 이후에는 시스템 **성능**과 **안전성** 등을 계속 **확인하며 유지·관리**를 수행합니다. 이때 **정보주체 권리 침해**가 발생하는지 **지속적으로 모니터링**하고[47], **개인정보 처리 과정**을 **투명**하게 알리는 등 정보주체 권리 보장 노력을 수행해야 합니다.

❑ 적용 전과 운영 중에 고려할 프라이버시 요소는 무엇인가요?

배포 전 점검 사항

배포 전 AI가 기존 **목적대로 안전하게** 동작하는지 **최종 점검**해야 합니다. **AI 모델, 소스코드, 학습데이터** 등을 **전반적으로 검토·분석**하며 **정확도, 안정성을 평가**해야 합니다. 기업·기관이 외부 모델을 사용하거나 시스템을 발주하는 경우 등 **분석 대상**에 **직접 접근할 수 없는 경우**, 외부 모델에 **쿼리를 입력**해서 **산출물을 평가**하거나 **시스템 공급업체**에 **평가데이터를 제공**하는 등으로 **대체 검사 방법을 사용**하는 것을 고려할 수 있습니다[48].

이와 같이 최종 배포 전에 **실제 동작 환경**에서 AI의 정확도, 안전 조치 우회 시도에 대한 저항성, 학습데이터 유·노출 가능성 등 **프라이버시 리스크를 점검하고 그 결과를 문서로 관리**하여 배포 전 시스템의 안전성을 높일 수 있습니다.

---

43) https://futureoflife.org/index, FLI AI Safety Index 2024 (2024. 12. 11.)
44) https://github.com/centerforaisafety/HarmBench
45) https://jailbreakbench.github.io
46) https://airisk.mit.edu
47) GDS, 「AI Playbook for the UK Government」, "Principle 4: You have meaningful human control at the right stages" 참고
48) OMB(Office of Management and Budget) Memorandum(M-25-21), 「Accelerating Federal Use of AI through Innovation, Governance, and Public Trust」, Section 4(b)의 i. "Conduct Pre-Deployment Testing" 참고

참고 **해외 AI 시스템 배포 전 테스트 사례**

■ **US CAISI 및 UK AISI의 Anthropic 모델 대상 합동 배포 전 테스트(24. 11.)**

- 美 AI 표준 및 혁신 센터(US CAISI) 및 英 AI 안전연구소(UK AISI)에서 합동 테스트를 Anthropic의 'Sonnet 3.5 (new)' 모델을 대상으로 수행
- 소프트웨어 개발 분야를 포함한 전문 분야 문제 해결 능력과 함께 안전조치(safeguard) 효과성*을 기준으로 테스트를 수행

* UK AISI에서 자체 개발한 Criminal Activity, AgentHarm와 공개적으로 알려진 HarmBench 등 여러 탈옥 공격 기법을 적용하며 대상 모델의 응답을 정략적으로 평가

기업·기관은 배포 전 테스트 단계에서 최종적으로 확인한 프라이버시 리스크를 고려해서 생성형 AI의 **사용 목적, 금지 행위** 등을 **'허용되는 이용방침'**(AUP, acceptable use policy) 등에 **작성·공개**할 수 있습니다.

이때, 전략 수립 단계에서 검토한 내용을 참고할 수 있습니다. PbD 원칙, 개인정보 영향평가 등 **사전예방적 접근**으로 식별한 생성형 AI의 **프라이버시 리스크**와 **예견 가능한 오·남용**을 고려해 AUP를 작성할 수 있습니다.

작성·공개한 AUP는 생성형 AI가 배포된 후 오·남용을 미연에 방지하고, 이를 **위반하는 이용 행위**에 대하여 **서비스 제한** 또는 **계정 차단** 등의 조치를 취할 수 있는 근거가 될 수 있습니다.

사례 **허용되는 이용 방침(Acceptable Use Policy) 활용 사례**

■ **네이버 CLOVA X 서비스 이용정책 일부('24.7월 기준) ※ 최종이용자 대상 작성**

사용자는 CLOVA X 서비스를 사용함에 있어 아래 의무를 부담합니다.

① 사용자는 CLOVA X 서비스를 악의적으로 사용하는 것이 금지됩니다. 악의적 사용에는 아래와 같은 행위 및 이와 유사한 목적을 가진 행위가 포함되며, 아래 예시에 한정되지 아니합니다.

1. 불법적인 행위나 범죄 및 유해한 행동에 대한 콘텐츠 생성
   - 아동 성적 학대 또는 착취와 관련된 콘텐츠
   - 불법 약품(마약 등) 또는 상품(무기 등)의 판매를 조장/촉진 또는 이를 제조하는 방법에 대한 콘텐츠
   - ...

6. 악성코드 및 해킹, 공격, 서비스 어뷰징 코드 등의 생성 등
   - 정보처리장치 등에 접근권한 없이 액세스하는 등 침입하거나...

8. 본인이나 타인의 민감정보, 고유식별정보 등 개인정보를 입력하거나 개인정보 및 사생활 침해를 야기할 수 있는 대화의 유도 및 콘텐츠 생성

■ **OpenAI Business 이용정책 일부**('25.5월 기준) ※ 이용사업자 등 대상 작성

3.3. Restrictions. Customer will not, and **will not permit** End Users to: (a) use the Services or Customer Content in a way that **violates applicable laws or OpenAI Policies;** (b) use the Services or Customer Content in a way that **violates third parties' rights;** (c) allow minors to use OpenAI Services **without consent from their parent or guardian;** (중략) (h) interfere with or disrupt the Services, including circumvent any rate limits or restrictions or **bypass any protective measures or safety mitigations for the Services;** (이하생략)

모니터링, 정보주체 권리보장, 투명성 확보

기업·기관은 **편향·차별적**이거나 **권리를 침해**하는 등 **부적절한 결과물**에 대응할 수 있도록 정보주체의 **신고·의견 제출 기능**을 시스템에 **탑재해야 합니다.** 이후, 해당 창구를 상시적으로 **모니터링**하여 접수된 의견에 대해서는 **운영 정책** 및 **시스템 개선**에 **적극 반영**해야 합니다[49].

정보주체의 **열람, 정정·삭제 등의 요청**이 있는 경우에는 **시간·비용·기술적 측면**에서 **합리적으로 실현 가능한 범위**에서 **정보주체의 권리 보장 방안을 마련**할 필요가 있습니다. 개인정보 보호법에서는 **열람, 정정·삭제, 처리정지 요구권**에 대하여 **원칙적으로 10일 이내 대응**하도록 규정하고 있습니다. 다만 학습데이터셋 크기, 구성 방식·체계 등을 감안하여 **전통적인 권리행사 보장이 어려운 경우**에는 **그 사유를** 정보주체에게 **알기 쉽게 알리고**, 대체 수단 등을 통해 **최대한 성실하게 요구에 응하는 것**이 권장됩니다. 예를 들어 **출력 필터링** 등을 **우선 적용**하여 개인정보가 발화되지 않도록 **긴급 조치**하고, 추후 **학습데이터셋**에서 개인정보를 **제외하는 방안**을 고려할 수 있습니다.

한편, 비용부담이 큰 모델 재학습 대신 모델 내 **특정 항목을 삭제**하는 **'머신 언러닝'** 기술은 아직 기술 성숙도가 높지 않은 한계가 있으나, 향후 **정보주체 권리 보장**을 위한 **유용한 수단**이 될 수 있습니다.

49) 「Llama Developer Use Guide: AI Protections」(Meta, '25. 4.), "Feedback & reporting mechanisms" 참고

참고 **모델개발자 및 이용자의 정보주체 권리 보장 책임 예시**

※「AI 프라이버시 리스크 관리 모델」('24.12.)

- **(모델이용자)** 삭제 요청 수령시, ①**입력 및 출력 필터링** 등 **기업·기관이 실행 가능한 경감조치**를 취하고 ②**가능한 경우 모델개발자에게 삭제·정정 요구를 전달** 후 그 결과를 정보주체에게 통보
- **(모델개발자)** 모델이용자로부터 정보주체의 삭제 요구 수령 시, 학습데이터에 **개인정보 존재 여부**를 확인하고, 종국적으로 해당 데이터가 삭제되도록 합리적인 기간 내에 **모델 업데이트**

참고 **모델 내 개인정보 삭제를 위한 모델 언러닝 기법[50]**

※「AI 프라이버시 리스크 관리 모델」('24.12.)

- **(개념)** 모델이 학습된 정보를 **의도적으로 망각**하는 것으로 **잘못된 정보나 학습에 부적합한 정보(개인정보, 저작권 등)를 삭제**하는 기술
- **(한계)** 최근 위험 경감 기술로서 주목받고 있는 분야로 **추가적 연구가 필요**하며, **적절한 망각 수준을 찾기 어려운 한계** 등

생성형 AI 시스템을 활용하여 **정보주체에 대한 의사결정**을 내리는 경우에는 해당 의사결정이 최종적인 결정으로서 **개인정보 보호법상 자동화된 결정**에 해당하는지 여부를 확인하고, 거부권, 설명요구권, 검토요구권 등 **정보주체의 권리**를 **보장**해야 합니다.

법률 **제37조의2(자동화된 결정에 대한 정보주체의 권리 등)** ① 정보주체는 완전히 자동화된 시스템(인공지능 기술을 적용한 시스템을 포함한다)으로 개인정보를 처리하여 이루어지는 결정(「행정기본법」 제20조에 따른 행정청의 자동적 처분은 제외하며, 이하 이 조에서 "자동화된 결정"이라 한다)이 자신의 권리 또는 의무에 중대한 영향을 미치는 경우에는 해당 개인정보처리자에 대하여 해당 결정을 거부할 수 있는 권리를 가진다. 다만, 자동화된 결정이 제15조제1항제1호·제2호 및 제4호에 따라 이루어지는 경우에는 그러하지 아니하다.

② 정보주체는 개인정보처리자가 자동화된 결정을 한 경우에는 그 결정에 대하여 설명 등을 요구할 수 있다.

③ 개인정보처리자는 제1항 또는 제2항에 따라 정보주체가 자동화된 결정을 거부하거나 이에 대한 설명 등을 요구한 경우에는 정당한 사유가 없는 한 자동화된 결정을 적용하지 아니하거나 인적 개입에 의한 재처리·설명 등 필요한 조치를 하여야 한다. (이하 생략)

참고 **자동화된 결정에 대한 정보주체 권리와 조치 의무**

※「자동화된 결정에 대한 정보주체 권리안내서」('24.9.)

정보주체 권리	개인정보처리자 조치
거부권	▸ 자동화된 결정 **적용 정지** 또는 **인적 개입에 의한 재처리** 후 **결과 고지**
설명 요구권	▸ 자동화된 결정에 대한 **간결하고 의미있는 설명을 제공** ※ 중대한 영향을 미치는 경우가 아닌 때에는 공개된 사항 등을 활용하여 설명
검토 요구권	▸ 제출한 **의견 반영 여부 검토** 등 조치 후 그 **결과를 통지**

50) Shrishak, K., "AI-Complex Algorithms and effective Data Protection Supervision Effective implementation of data subjects' rights", Support Pool of Experts Programme EDPB (2024), 다양한 종류의 언러닝 기법들 확인 가능

지금까지 살펴본 정보주체의 권리를 충분히 보장하기 위해서는 AI 학습데이터에 **본인의 정보가 포함**되어 있는지, 개인정보가 AI에서 **어떻게 처리**되는지 등을 **정보주체에게 명확하게 안내**할 필요가 있습니다.

기업·기관은 **데이터셋 수집 사실, 주요 출처, 처리 목적 등**과 함께 AI 시스템 **개인정보 처리 과정**을 **개인정보 처리방침, 기술 문서, FAQ 등에 투명하게 공개**하여 정보주체가 권리를 행사할 수 있도록 지원할 것이 권장됩니다. 또한, **열람 및 정정·삭제** 등 요구 대응에 **기술적으로 제한이 있을 경우** 해당 사항을 **사전에 정보주체에 안내**하는 것도 정보주체 권리 보호에 도움이 될 수 있습니다.

최근 부상하고 있는 **AI 에이전트**는 **기억 능력**을 **보유**하고 **자율적으로 작업을 수행**하는 AI 시스템으로 동작 과정에서 **과다한 이용자 정보 수집 및 프로파일링 등에 대한 우려**가 늘어나고 있습니다.

따라서 AI 에이전트 서비스를 제공하는 기업·기관은 **프롬프트 입력 등 이용자 대화이력**의 **▲학습데이터 이용 여부 ▲제3자 제공 여부 ▲보관·파기 정책 ▲필터링 기준** 등을 명확하게 고지하는 것이 권장됩니다. 이때 **학습 전·후로 충분한 기간**을 두고 **안내**하여 **정보주체의 실질적인 선택권**(opt-out 등)을 보장하는 것이 바람직합니다.

참고 **생성형 AI 관련 개인정보 처리방침 수립 시 유의사항**

※「개인정보 처리방침 작성지침」('25.4.)

- AI 개발 및 서비스를 제공하는 개인정보처리자의 경우 최소 수집, 목적 명확화 등 고려하여 **학습데이터 수집·이용 기준*을 미리 정하고 기재**하는 것을 **권장**

  * AI 시스템 개발에 필요한 데이터 양(volume), 범주(민감정보, 행태정보 등) 등을 고려하여 공개된 개인정보의 **주요 수집 출처, 수집 방법, 안전성 확보 조치 방안** 등 포함

- 개인정보 처리방침을 **변경**하는 경우 **변경 및 시행 시기, 변경된 내용을 지속적으로 공개**하고 이전의 개인정보 처리방침이 있을 시 그간의 변경 이력 기재

  * 주요 변경 사항을 별도로 안내하는 경우 **웹페이지 팝업창, 공지사항 등을 통해 정보주체가 쉽게 확인할 수 있는 방법**으로 알릴 것을 권장

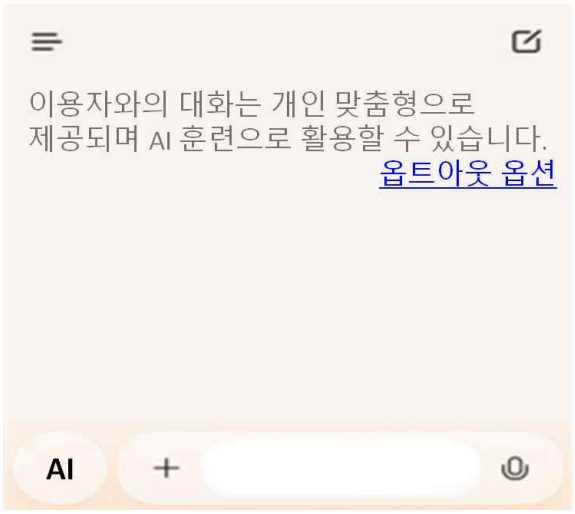
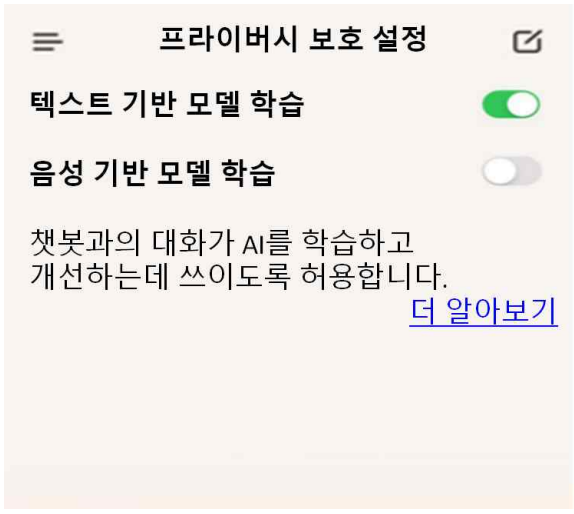
- **자동화된 결정을 하는 경우**에는 **그 기준과 절차, 개인정보가 처리되는 방식 등을 정보주체가 쉽게 확인할 수 있도록 기재**

사례 **정보주체 권리 보장 강화 조치 사항 예시**

※ 개인정보보호위원회 제2025-011-031호(비공개) 결정
※ 개인정보보호위원회 제2024-006-169~174호(공개) 결정

- 이용자가 **입력하는 데이터**를 **AI 학습**하는 경우에는 이러한 사실을 분명하게 알리고 **정보주체의 실질적인 선택권 보장**
  - (예시. 챗봇 서비스) 대화방 입력 상단에 '대화 내용이 학습데이터로 수집된다'는 사실과 '데이터 수집 거부 방법(옵트아웃 등)'을 고지
- 신규 이용자에게 **충분한 기간 동안 최초 고지**, 옵트아웃(opt-out)하지 않은 이용자는 **일정 회수 이상 추가 고지**, 이후 **정기적으로 재고지**
- **학습데이터 수집·거부·파기 정책을 투명하게 공개**
  ※ 학습데이터에 포함되는 개인정보 항목, 수집·이용 목적(예. LLM 학습), 개인정보 필터링 정책, 보유기간 및 파기 방법(예. 이용자의 opt-out 설정 이후부터 데이터 수집 중지) 등
- 이용자가 입력한 데이터를 언제든지 **손쉽게 제거·삭제할 수 있는 기능을 제공, 해당 기능 접근성을 제고**하는 것을 포함하여 **개인정보 침해 최소화 조치** 이행
- 개인정보 침해 관련 취약점에 대해 **LLM 수정 및 재배포, 문의 창구 개설**, 자신의 **LLM을 사용하는 기업에 대한 취약점 조치 방안 안내** 절차 등을 마련

사례 **옵트아웃(opt-out) 보장 사례**

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>학습된다는 사실과 옵트아웃을 알기 쉽게 프롬프트 입력 란에 안내
(충분한 기간 최초 고지, 이후 정기적 재고지)</td>
<td>옵트아웃을 쉽게 수행할 수 있도록 직관적인 설정 인터페이스를 구성</td>
</tr>
</table>

5 AI 프라이버시 거버넌스 체계

생성형 AI의 **데이터 처리 흐름이 복잡**해지면서 **리스크 관리**의 중요성이 커지고 있습니다. 이러한 환경 변화에 따라 각 기업·기관에서는 **개인정보 관련 법규 준수**와 **리스크 관리**를 **총괄**하는 **개인정보 보호책임자**(CPO) 중심으로 **내부 관리체계**를 **구축·운영**할 필요가 있습니다.

참고 **개인정보 보호법상 CPO 제도(法 제31조)**

▸ 개인정보 보호법은 개인정보 관련 법규 준수, 오남용 방지 등 개인정보처리자의 개인정보 보호 활동을 촉진하고 책임을 부과하기 위한 CPO 제도 규정
- (CPO의 정의) 개인정보 처리에 관한 업무를 총괄하여 책임지는 자
- (CPO 지정 의무) 소상공인을 제외한 개인정보처리자는 CPO를 지정해야 함

CPO가 생성형 AI의 목적 설정부터 배포·관리에 이르기까지 전 과정에서 **개인정보 처리**의 **적법성**과 **안전성**을 **확보**하기 위한 **관리·감독 책임**을 수행할 수 있는 체계를 구축해야 합니다. 이를 기반으로 **개인정보 영향평가, 레드티밍**[51] 등 점검 도구를 활용해 **생성형 AI의 프라이버시 리스크를 지속적으로 평가하고**, 이를 경감하기 위한 **다층적 안전조치**를 **수립·이행**하며, 관련 절차와 결과를 체계적으로 **문서화**해야 합니다. 이 과정에서 **중대한 취약점**이 발견될 경우 **기업·기관** 내 이사회 등 **최고 의사결정기구**에 **보고**하고, 관련 **정부 부서**와 **신속히 정보**를 **공유**하는 것이 바람직합니다.

참고 **주요 AI 기업의 레드티밍 방법론 등 사례**

■ **OpenAI: 외부 전문인력 레드티밍과 자동화 레드티밍 병행**[52]
- **전문인력 레드티밍**은 ①주요 테스트 영역 선정 및 ②대상 모델 선정, ③테스트 인터페이스 및 문서화 지침 제공, ④레드티밍 결과 사후 평가 및 검토 단계로 수행
- **자동화 레드티밍**은 대규모 공격 사례를 생성하고 테스트하는 데 효율적이며, 전문인력 레드티밍 단계에서 식별된 중요 시나리오를 반영하며 보완

51) AI시스템이 어떤 위험을 일으킬 수 있는지 사전에 점검하는 일종의 '모의 공격 테스트'로서, 사람 또는 다른 AI가 AI모델을 의도적으로 시험해보며 유해 콘텐츠 생성, 편향된 답변, 보안 우회(jailbreak) 등 문제를 식별하는 도구·활동

- **Microsoft: ①사전준비, ②테스트수행, ③결과점검 순으로 반복**[53]
  - **사전준비** : 레드티밍을 수행할 다양한 분야의 전문가를 모집, 테스트 대상 선정 (대상 모델, AI 시스템 인터페이스 등) 및 점검할 리스크 영역 선정
  - **테스트수행** : 레드팀 인력에 테스트 절차 및 접근권한 등 상시 제공·지원, 상황 확인
  - **결과점검** : 테스트 결과를 정기적으로 주요 이해관계자에게 보고 (식별된 주요 이슈 목록, 향후 레드티밍 계획 검토 및 관련 자료 등)

참고 **레드티밍 시 모델·개발·시스템·운영 영역별 점검항목 사례**[54]

- **모델: 추론공격**(파라미터, 학습데이터 유도 등), **추출공격**(모델 정책, 시스템 프롬프트 등), **모델조정**(탈옥공격, 프롬프트 인젝션, 안전성 정책 우회 등) 등 점검
- **개발: 데이터오염**(외부 벡터형 DB 오염, 검색증강생성결과 조작, 캐시 오염 등), **프록시/방화벽 정책 우회**, **콘텐츠 필터 우회**(다언어 공격, 문맥 우회 공격 등), **접근제어**(세션 관리, API 접근권한, RBAC 정책, 토큰 관리 정책 등) 점검
- **시스템: 원격 코드 실행** 취약점(결과로 생성된 코드 실행, 시스템 커맨드 실행 등), **공급망 취약점**(의존성 무결성, 업데이트 체계, 배포 파이프라인 등) 점검
- **운영: 개인별 요청 빈도** 제한, 허가되지 않은 **데이터 접근**, 입·출력 **필터링 누락, 보안 감사 및 로그 기록** 정보가 충분한지 등 점검

아울러, CPO는 최고인공지능책임자(CAIO, Chief Artificial Intelligence Officer), 정보보호최고책임자(CISO, Chief Information Security Officer) 등과 **긴밀한 협력 체계**를 유지하며, 조직 내 개인정보 처리가 수반되는 **생성형 AI 개발·활용**에 **적극적으로 관여**할 수 있는 **권한**과 **역할**을 보장받아야 합니다. 특히 **AI 기획·개발 초기 단계**부터 개입하여 **개인정보 처리**에 대한 **충분한 정보를 확보**하고, **관련 부서**에 **적시 피드백**을 **제공**하여 **개인정보 안심설계**(PbD) 관점이 **생성형 AI 서비스에 내재화**될 수 있도록 해야 합니다.

52) "Advancing red teaming with people and AI", OpenAI, ('24. 11. 21.) 참고 https://openai.com/index/advancing-red-teaming-with-people-and-ai/
53) "Planning red teaming for large language models (LLMs) and their applications, Microsoft, ('25. 2. 7.) 참고
54) GenAI Red Teaming Guide, OWASP, ('25. 1. 22.), "5. GenAI Red Teaming Strategy" 및 "6. Blueprint for GenAI Red Teaming" 참고

사례 **프라이버시 및 보안 협력 내부 관리체계 下 AI 프라이버시 레드팀 운영 사례[55]**

■ **체계적인 테스트 방법론 사용**

- OWASP(Open Worldwide Application Security Project) 등의 표준 보안·프라이버시 진단 프레임워크 활용
- 정형화된 공격 시나리오 및 테스트 케이스 개발

■ **공격자 시나리오 기반 테스트**

- 멤버십 추론 공격(Membership Inference Attack), 모델 역전 공격(Model Inversion Attack)
- 민감정보 추출 시도, 프롬프트 인젝션 공격 등 테스트

■ **지속적인 테스트 수행**

- 개발 초기부터 테스트 병행하며, 새로운 기능 추가 시 반드시 테스트 수행
- 정기(분기별, 반기별) 테스트 일정 수립

■ **문서화 및 보고**

- 심각도 및 영향 범위 평가를 포함한 발견된 취약점에 대한 상세한 문서화

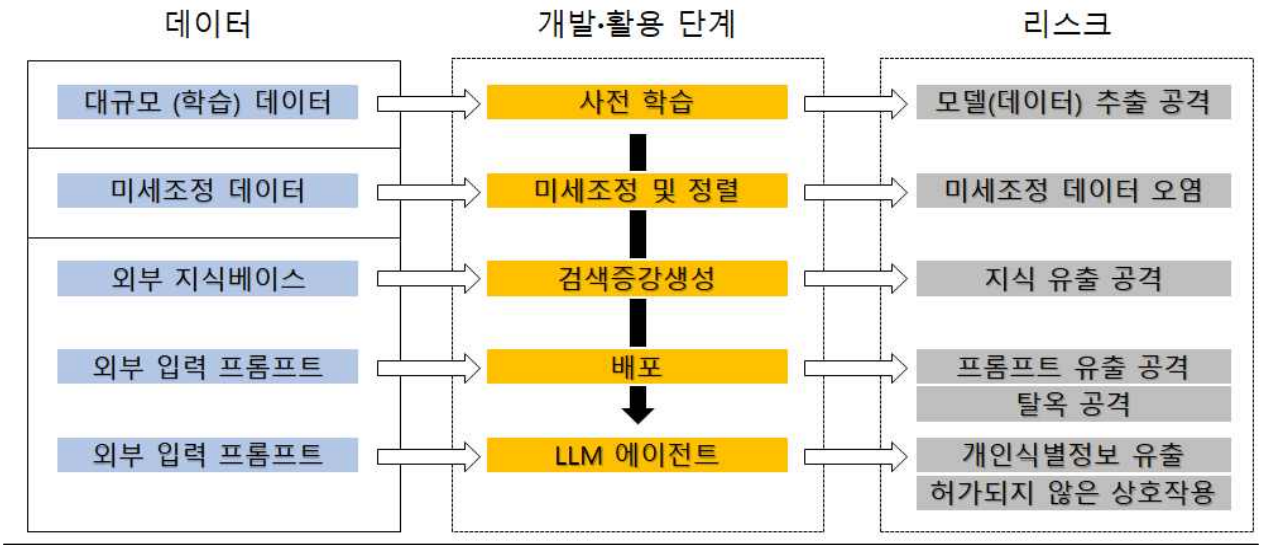
■ **최신 동향 모니터링**

- 새로운 AI 취약점 및 공격 기법 확인
- 관련 법규 및 규제 변화 모니터링

■ **역량 강화 및 훈련**

- 모의 해킹 훈련 및 워크숍 개최 등

참고 **생성형 AI에서의 데이터, 개발·활용 단계, 시나리오 관련 리스크 사례[56]**



55) 「Llama Developer Use Guide: AI Protections」(Meta, '25. 4.), "Red teaming best practices" 참고

56) Wang, S. et al, "Unique Security and Privacy Threats of Large Language Model: A Comprehensive Survey", 2024, arXiv:2406.07973, Figure 1.을 바탕으로 일부 수정

붙임 AI 개발·활용 유형별 프라이버시 고려사항 (예시)

※ ▲개인정보 보호법령에 따른 의무사항과 ▲의무사항은 아니나 개인정보 보호 원칙 등에 따른 자율적 책무(색상표기X)로 구분

구 분		모델개발자	모델이용자
설명		생성형 AI 언어모델을 개발하고 제공·판매하는 기업·기관 등	①외부 상용 모델을 API 연계하거나 ②공개 모델을 다운로드하고 개발해 AI 서비스 제공하는 기업·기관 등
목적 설정	목적 구체화	☑ 개인정보 처리 목적 구체화	
	적법근거 마련	☑ 법적근거 검토·마련 (예: 공개된 개인정보 활용)	☑ 법적근거 검토·마련 (예: 이용자 정보 AI 학습)
전략 수립	PbD, 영향평가	☑ PbD에 기반한 개발 수행 ☑ 개인정보 영향평가 수행 (※ 공공은 의무, 민간은 권장)	
	이용 모델 검토		**서비스형 LLM 사용 시** ☑ 외부 LLM의 데이터 처리 범위·보관·재이용(학습 포함) 통제 ☑ 국외이전 적법성 검토 **기성 LLM 사용 시** ☑ 훈련 데이터 출처 검증 노력 ☑ 모델 안전조치 및 취약성 검토 - 모델카드, 라이선스 정책 등 확인
	배포 모델 신규 리스크 대응	☑ 모델 내 신규 리스크 발견 시 공지, 모델 개선 및 재배포 등 수행	☑ 공지 리스크 신속 대응, 관리 체계 보완 ☑ 이용 모델의 최신 버전 확인 및 상시 업데이트 유지
AI 학습 및 개발	데이터 전처리 등	☑ 데이터 전처리(가명·익명처리 등), PET 적용 등	☑ 추가학습 데이터 전처리(가명·익명처리 등), PET 적용 등
	모델 안전조치	☑ 미세조정·정렬 통한 배포할 모델의 안전성 제고, 적대적 공격 방어	☑ 이용 모델 검토 후 필요시 미세조정·정렬 등 모델 안전성 강화 장치 추가
	시스템 안전조치		☑ 시스템 안전조치 적용 (접근제어 관리, 입출력 필터, RAG 관련 안전조치 등)
시스템 적용 및 관리	배포 전 테스트	☑ 배포 전 테스트(모델 수준)	☑ 배포 전 테스트(시스템 수준)
	AUP 작성·공개	☑ AUP 작성·공개(이용사업자 대상)	☑ AUP 작성·공개
	모니터링		☑ 개인정보 유노출 및 시스템 모니터링
	정보주체 권리보장	☑ 모델이용자로부터 정보주체 권리 요청 받아 대응하는 체계·방안 마련	☑ 정보주체 권리 요청 대응 ☑ 필요한 경우 권리 요청을 모델 개발자에 전달
	투명성 확보	☑ 모델카드 작성, 개인정보 처리방침 마련 등 투명성 확보	☑ 개인정보 처리방침 마련 투명성 확보
거버넌스	내부 관리체계 구축·운영	☑ CPO 중심의 AI 프라이버시 내부 관리체계 구축 ☑ AI 개발·활용 전 단계에서 프라이버시 리스크 관리 활동 수행	

※ 생성형 AI 개발·활용 유형 중 대표적인 모델개발자 및 모델이용자가 참고할 수 있는 기준을 안내한 것으로, **각 기업·기관은 구체적인 개발·활용 맥락과 상황에 필요한 조치사항을 이행할 수 있음**