FOSSLight Community Day 2026

생성형 AI 학습데이터 분쟁과 AI-BOM기반 투명성 컴플라이언스

책임변호사 조정원

April, 2026

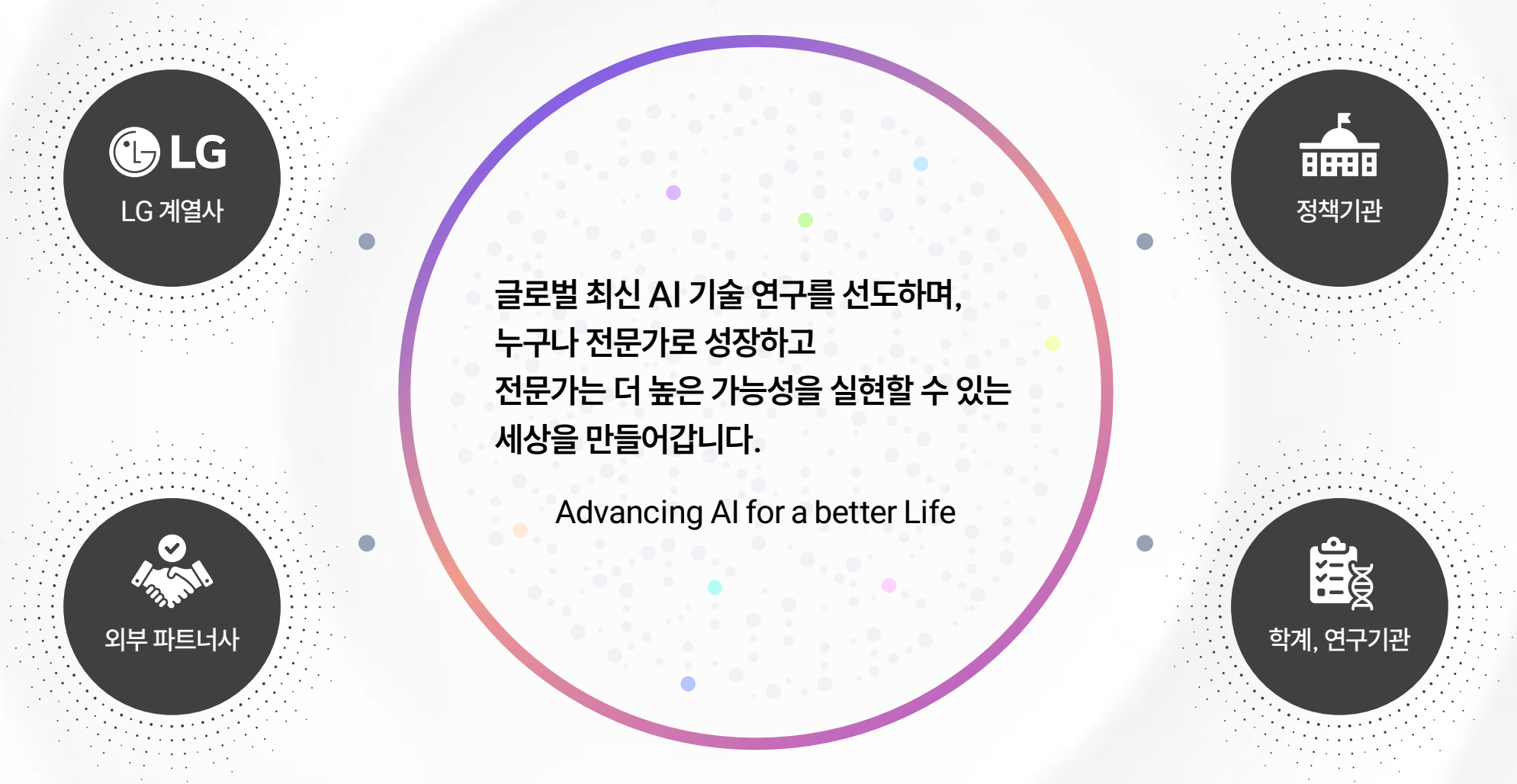
LG AI Research

Prologue

LG AI연구원은 2020년 12월에 설립된 LG그룹의 AI 싱크탱크로,
사업 난제 해결과 최신 AI 선행 연구, AI 윤리원칙 수립 및 이행 등을 통해 그룹 차원의 AI 역량을 강화하고 있습니다.

Mission and Ecosystem

글로벌 최신 AI 기술을 연구하고 그룹의 핵심 난제 해결을 주도하며, 우리가 상상하는 더 나은 미래가 현실이 될 수 있도록 LG그룹을 넘어 글로벌 선도 파트너사와 정책기관, 학계 및 연구기관과 함께 새로운 생태계를 만들어 나가고 있습니다.



Contents

Chapter

1

분쟁 사례로 보는 AI 오픈 데이터셋 관리 필요성

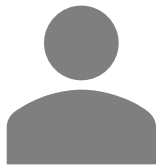
LG AI Research

Open Source v. Open Data

- 오픈소스와 오픈데이터는 형태, 이용목적, 컴플라이언스 준수 방법 등에서 차이를 가지고 있어, 다른 방법으로 접근해야 합니다.

Source Code	VS	AI Training Data
코드	형태	코드, 텍스트, 이미지, 영상, 오디오 등
소프트웨어 개발	이용목적	본래 저작물의 목적과 다른 목적
Open source compliance	컴플라이언스	표준화되지 않음
표준화되어 준수됨	고지의 의무?	표준화되지 않음

생성형AI Output의 저작권 침해 책임



⑤ AI User

Input

EXAONE

Output

① 데이터 권리자의 2차적저작물?
② 데이터셋 데이터베이스 생성자의 저작물?
③ AI 개발자의 저작물? 업무상저작물?
④ AI 서비스 제공자의 저작물? 업무상저작물?
⑤ AI User가 생성한 저작물?

Distribution

⑥ Output 이용자

광저우인터넷법원판결에 따라,
⑤ AI User의 권리가 침해될 수 있음.

서비스제공자

④ AI 서비스 제공자

베이징인터넷법원판결에 따라,
① 데이터 원저작자의 권리가 침해될 수 있음.

LG AI Research

③ AI 개발자

데이터셋 생성자

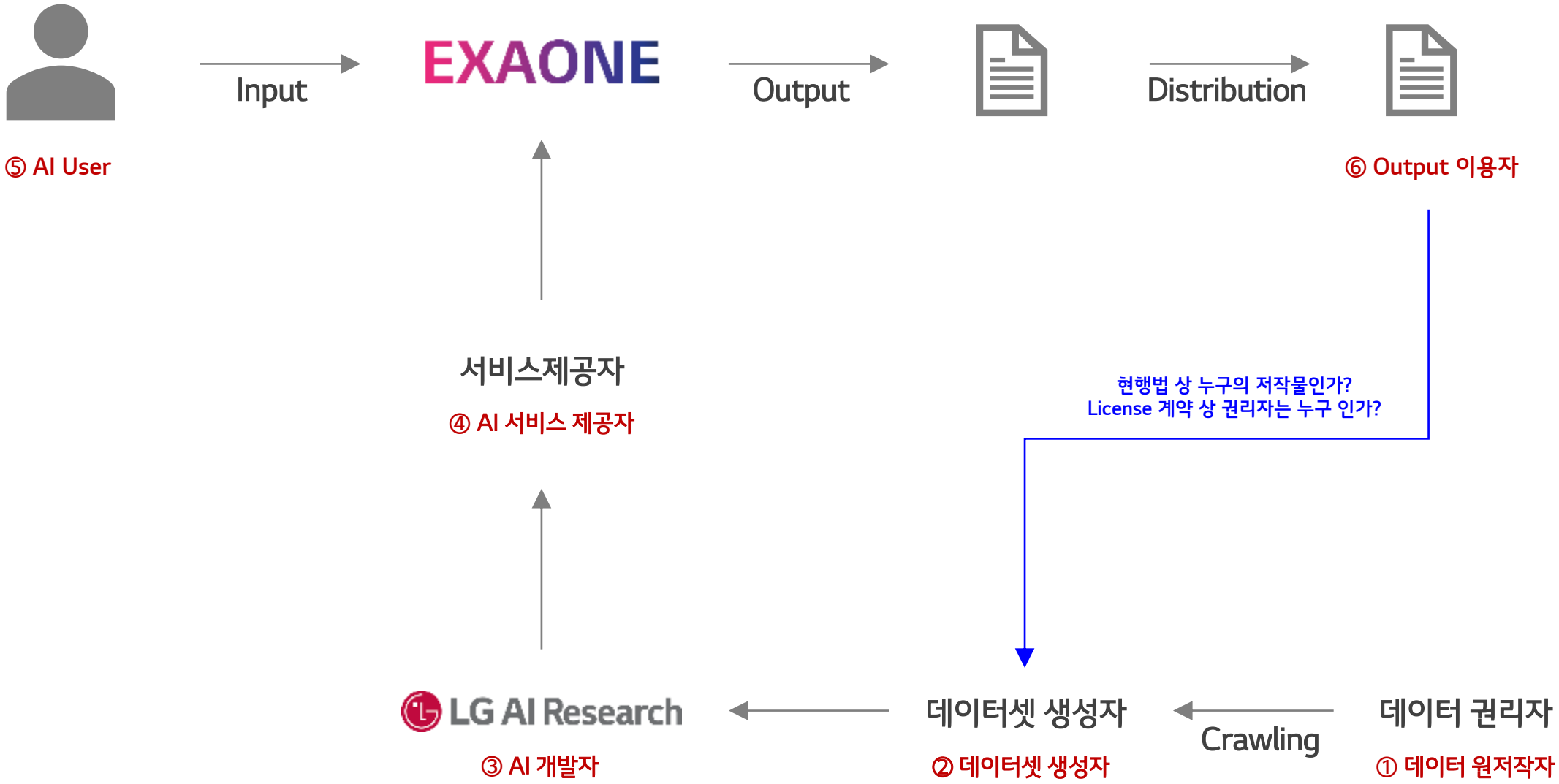
② 데이터셋 생성자

Crawling

데이터 권리자

① 데이터 원저작자

생성형AI Output의 저작권 침해 책임



생성형AI Output의 저작권 침해 책임

01	02	03	04
학습데이터 저작자	**AI 회사**	**프롬프트 입력자**	**누구도 아님**
원저작물 · 데이터셋	*시스템 설계자*	*창작적 개입 주체*	*Public Domain*

※ 후보 1의 '학습데이터 저작자'는 두 층위 — 원저작물의 저작자(1-a) + 데이터셋을 편집·제작한 자(1-b)

후보	법적 근거	주요 예시	기여의 성질
1-a. 원저작물 저작자	복제권 2차적저작물작성권	GEMA v. OpenAI (DE)	원재료 공급 사전적·간접적
1-b. 데이터셋 제작자	편집저작물 (저작권법 제6조) Feist 기준	Synthetic Data 쟁점 등	편집적 설계 선댁·배열·구성의 표현성
2. AI 회사	s.9(3) · 편집저작물 Stern	영국 CDPA Tencent (中, 2019)	시스템 구축 표현 득성 동제
3. 프롬프트 입력자	창작적 기여 도구돈	Li v. Liu (中) 강보현 (韓)	창작적 개입 현재적·직접적
4. 누구도 아님 (Public Domain)	인간 저작자성 부재 Public Domain	Thaler (US) USCO Part 2	기여 X (귀속 주체 부재)

'인간 창작' 기준의 방법론적 한계

BCI(Brain-Computer Interface) 기술은 이 문제를 이론에서 물리로 옮김.

뇌 신호와 AI 연산이 물리적으로 결합되는 순간, '인간이 창작했는가'라는 질문은 법 해석의 문제가 아니라 존재론적 판단 불가능의 문제가 됨.

역사적 패턴 · 기술이 등장할 때마다 반복되는 저작권 논쟁

19세기 후반 · 사진 → "셔터 누른 것이 창작인가" → 구도·조명 선택에서 창작성 인정 (해결)
20세기 후반 · 컴퓨터 프로그램 → "기계어가 저작물인가" → sui generis로 우회 해결 (각국 입법)
21세기 초반 · Photoshop / 디지털 아트 → "디지털 보정이 창작인가" → 인간 선택 인정 (해결)
2020년대 · 생성형 AI → 현재 논쟁 중 (Thaler·Li·Théâtre 판례 대립, 아직 미해결)
2030년대~ · BCI (Neuralink·Synchron) → **또 다시 새로운 논쟁이 시작될 것이며, 이번엔 답이 불가능할 수 있음**

BCI 시나리오 · 왜 '인간 창작' 기준이 붕괴하는가

화가가 BCI를 착용하고 '생각만으로' 이미지 생성. 뇌 신호가 AI 모델의 입력이 되고, AI가 그 신호를 이미지로 변환.

여기서 '인간이 창작했는가'의 답을 구하려면:
· 뇌 신호와 AI 연산은 **물리적으로 불가분**의 상태
· '어디까지가 인간이고 어디부터가 기계인가' — 답이 없음

방법론적 문제 제기

기술이 등장할 때마다 *"이게 인간 창작인가"*를 매번 새로 논쟁하는 방식의 법이 과연 유지 가능한가?

BCI 앞에서는 기준 자체가 판단 불가능해진다. 이건 해석의 실패가 아니라 **기준 설계 자체의 실패**다.

문제 제기 · 매번 기술이 바뀔 때마다 "저작물이냐 아니냐" 논쟁?

'인간 창작' 기준에 매몰되기보다, 법적으로 새로운 장치를 모색해야 할 때.

AI 기업들의 항변 – 저작권법상 면책

- 각 국가별로 AI 개발자가 주장할 수 있는 저작권법 상 면책규정은 서로 달리 규정되어 있습니다.
- 특히, AI 학습 자체를 면책 시키는 규정에 대한 저작권자와 AI개발자 간의 치열한 논의가 이어지고 있습니다.

국가	항변 근거	핵심 조항
미국	Fair Use	17 U.S.C. § 107 (4요소 형량)
한국	공정이용	저작권법 제35조의5
EU	TDM 예외	DSM Directive Arts. 3 & 4
일본	정보분석 예외	저작권법 제30조의4
싱가포르	전산데이터분석(CDA) 예외	Copyright Act 2021 §§ 243–244
영국	Fair Dealing	CDPA 1988 §§ 29–30

AI 학습 데이터 관련 사건 판결

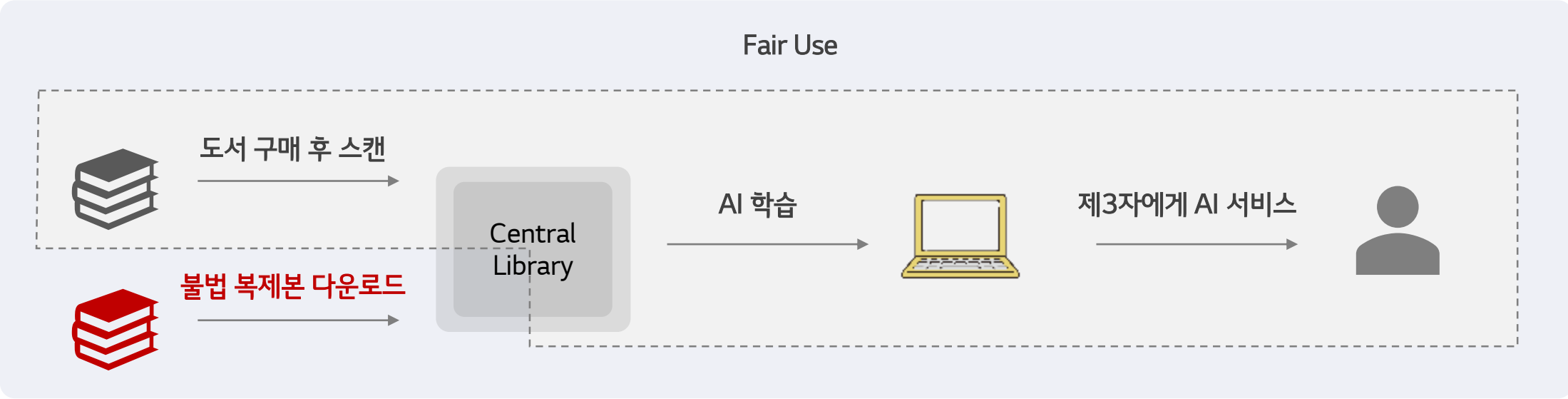
- 2024년부터, 전세계에서 각 국가의 저작권법에 따라서 다양한 판결이 이어지고 있습니다.
- 중국, 미국, 영국에서 AI회사가 패소하는 등, 학습데이터 컴플라이언스 중요성이 증대되고 있습니다.

사건	국가	년도	Prevailing Party	법적 판단 배경
Li v. Liu	China	2024	Plaintiff	Copyright law
Sin Changhwa Cultural Development LLC v. AI Company	China	2024	Plaintiff	Copyright law
Thomson Reuters v. ROSS Intelligence	U.S.	2025	Plaintiff	Copyright law (Fair use doctrine)
Andrea Bartz v. Anthropic	U.S.	2025	Plaintiff	Copyright law (Fair use doctrine)
Kadrey v. Meta	U.S.	2025	Defendant (AI Company)	Copyright law (Fair use doctrine)
Getty Images v. Stability AI	United Kingdom	2025	Defendant (AI Company)	Copyright law
GEMA v. OpenAI	Germany	2025	Plaintiff	Copyright law (EU DSM)

Plaintiff: 저작권자, Defendant: AI 회사

Andrea Bartz v. Anthropic (생성형AI 관련 미국의 첫번째 저작권 판결)

- 2024년 8월, 원고인 Bartz를 비롯한 저작자들은 Anthropic이 저작자들의 저작물을 허락 없이 LLM 학습에 이용하였다고 소를 제기하였음.
- 2025년 6월, Oracle 사건의 1심 판사이며 본 사건 판사인 Alsup은 1) 저작물을 LLM 학습을 위해 변형적으로 이용한 것은 공정이용, 2) 합법적으로 구매한 도서를 디지털화하여 복제한 것도 공정이용, 다만 3) 불법적으로 다운로드한 도서를 저장한 것은 추후에 LLM 학습에 사용되는 것과 관계 없이 불법을 구성할 가능성이 있다고(추후 Trial 진행) 판결함.

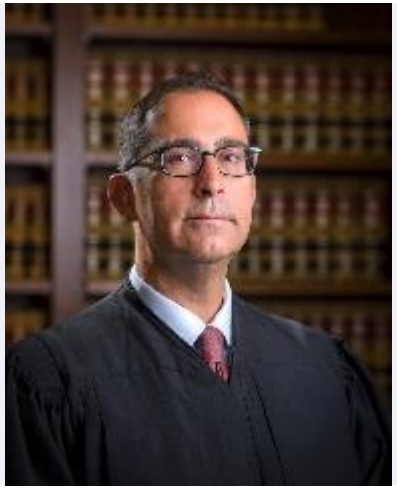


공정이용 4요소	1) LLM 학습을 위한 복제	2) 구매한 책의 복제(디지털화)	3) 불법 다운로드한 도서 복제(저장)
제1요소 : 원저작물 이용의 목적과 성격	피고 유리	피고 유리	원고 유리
제2요소 : 원저작물의 성격	원고 유리	원고 유리	원고 유리
제3요소 : 이용된 부분의 양과 중요성	피고 유리	피고 유리	원고 유리
제4요소 : 원저작물의 잠재적 시장 또는 가치에 미치는 영향	피고 유리	중립	원고 유리
공정이용?	**공정이용 O**	**공정이용 O**	**공정이용 X**

Kadrey v. Meta (생성형AI 관련 미국의 두번째 저작권 판결)

- 2023년 7월, 원고인 Kadrey를 비롯한 13인의 저작자들은 Meta가 저작자들의 저작물을 허락 없이 LLM 학습에 이용하였다고 소를 제기하였음.
- 사건 진행 중, 미국 연방지방법원 판사인 Vince Chhabria는 원고 측 변호인들이 "소송을 제대로 수행할 의지나 능력이 없어 보인다"고 이례적으로 질책함.
- 2025년 6월, Vince 판사는 Meta가 원고측 저작물을 이용하여 LLM을 개발한 행위는 공정이용이라고 판결함.

공정이용 4요소	LLM 학습을 위한 전 과정에서의 이용
제1요소 : 원저작물 이용의 목적과 성격	피고 유리
제2요소 : 원저작물의 성격	원고 유리
제3요소 : 이용된 부분의 양과 중요성	피고 유리
제4요소 : 원저작물의 잠재적 시장 또는 가치에 미치는 영향	원고 유리
공정이용?	**공정이용 O**



Vince 판사는 공정이용 제4요소를 바탕으로 LLM이 간접적으로 시장을 파괴한다는 판단을 했는가?

".. Alsup 판사는 생성형 AI의 변형적 이용에 주목하여, 학습 대상 저작물의 시장에 끼칠 수 있는 해악은 가볍게 여겼다.."

".. 많은 경우에서, 저작권 보호 저작물을 허가 없이 복제하여 생성형 AI를 학습시키는 것은 불법이 될 것이다.."

".. 법원은 일반적 이해에 근거해 판결할 수 없고, 당사자들이 제출한 증거에 근거하여 사건을 판단해야 한다.. 원고의 주장은 명백히 패소할 주장이었다.."

".. Meta가 불법적으로 원고의 책을 다운로드(복제)한 행위나 학습에 사용한 행위는 별개의 행위이나, 결국 모두 고도로 변형적인 최종 목적을 위한 것이었다.."

"..공정이용 4요소가 가장 중요하며.. 원고가 승소할 수 있는 논점은 유사한 작품을 시장에 쏟아낼 가능성이 있는 제품을 만든 것, 즉 Market dilution 이었다.. 그러나 원고는 관련 증거를 제대로 제시하지 못했다.."

Getty Images v. Stability AI

❑ Getty Image는 2023년 1월 및 2월, 각각 영국과 미국에서 자신들에게 소유권이 있는 이미지들이 Stability AI로부터 무단으로 AI모델에 학습되고, Stability AI가 유사한 Image들을 사용자에게 제공하면서 Getty Image와 직접적으로 경쟁하게 되었다며 소 제기.

Stable Diffusion의 Output

원고측 주요 주장 (Getty Images)

- 저작권 침해 (Copyright Infringement)

  Stability AI가 허락없이 Getty Images의 이미지를 학습 및 재생산하면서 2차적 저작물을 무단으로 생성하였음
- 불공정 경쟁 (Unfair Competition)

  Stability AI가 허락 없이 Getty Images의 이미지를 사용하면서, 공공이 이것이 허락된 행위인 것처럼 착각하도록 했음
- 상표권 침해 및 희석화(Dilution)

  Stability AI가 생성한 이미지에는 허락받지 않은 Getty Images의 상표가 포함되어 있으며, 기괴한 이미지와 함께 제공되면서 상표의 가치를 희석시킴

피고측 주요 주장 (Stability AI, UK)

- 공정 거래 (Fair Dealing)

  생성된 이미지는 pastiche(모방 작품)일 뿐이며, 이것은 공정 거래의 범주에 속할 수 있음
- 학습 및 개발 과정에서의 Jurisdiction

  Stable Diffusion 모델의 AI학습과 개발이 이루어 진 곳은 영국이 아니었음
- Output의 저작권 침해 부인

  Output은 프롬프트로 이미지를 생성할 때 무작위 노이즈를 시작점으로 사용하며, 단순히 AI 모델의 파라미터 값과 연결되어 이미지를 생성할 분, 기존 이미지의 직접적인 저작권 상 "복제" 행위가 발생하지 않음

Getty Images v. Stability AI

❑ Getty Images의 영국/미국 동시 소송의 실익 1 : Governing Law에 따른 법리적 판단 (영국의 Fair Dealing vs 미국의 Fair Use)

영국 저작권법(Copyright, Designs and Patents Act 1988) 상의 공정 취급 (Fair Dealing)

29. Research and private study
(1) Fair dealing with a work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement.]

30. Criticism, review and news reporting.
(1) Fair dealing with a work for the purpose of criticism or review, of that or another work or of a performance of a work, does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise) and provided that the work has been made available to the public.

30A. Caricature, parody or pastiche
(1)Fair dealing with a work for the purposes of caricature, parody or pastiche does not infringe copyright in the work.

VS

미국 저작권법(United States Code Title 17-Copyrights) 상의 공정 이용 (Fair Use)

107. Limitations on exclusive rights: Fair use
Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

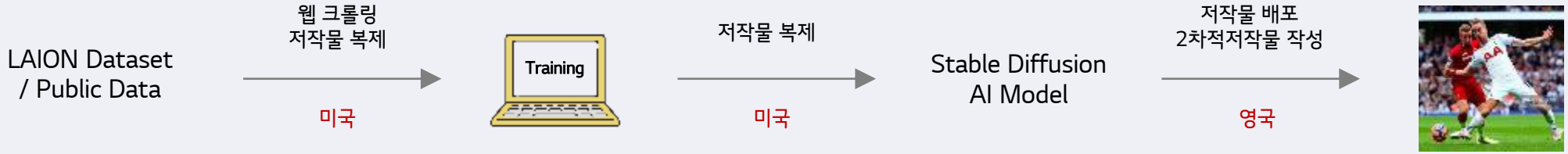
(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
(2) the nature of the copyrighted work;
(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
(4) the effect of the use upon the potential market for or value of the copyrighted work.

The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

지정된 목적 중 하나인 경우에만 적용될 수 있음

목적은 단지 예시로, 해당 목적에 포함되지 않아도 공정하지 않다고 단정할 수 없음

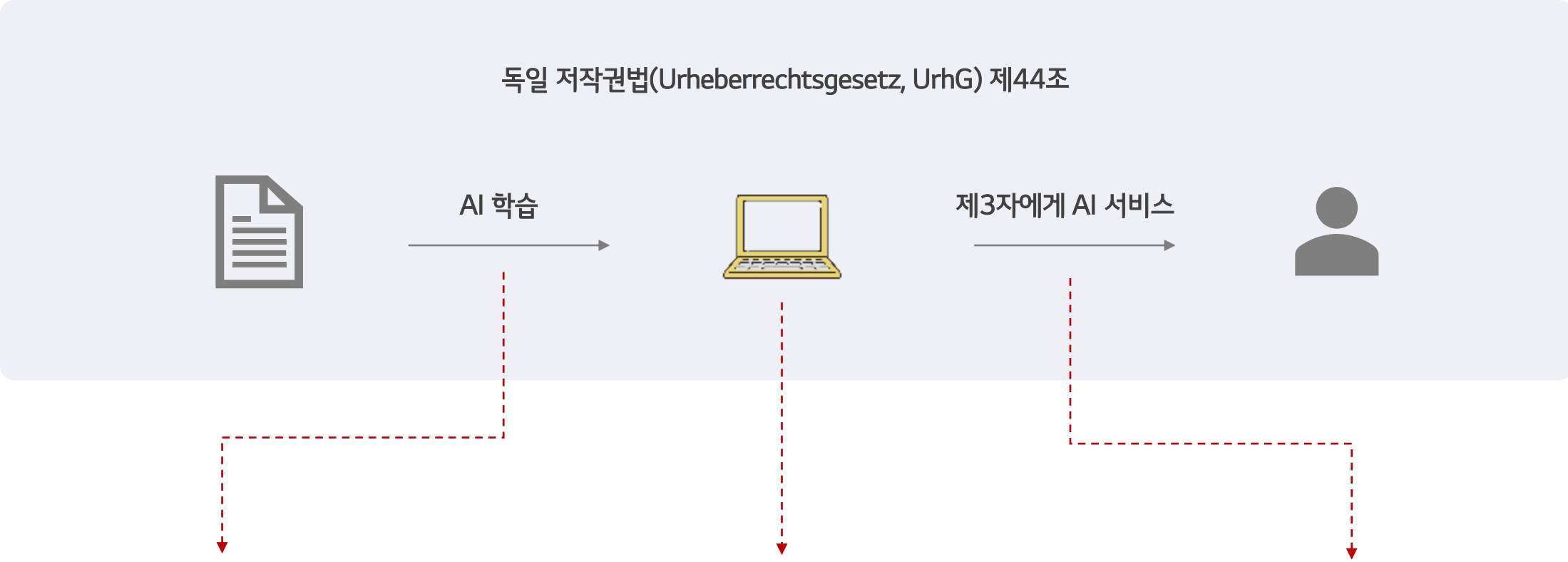
❑ Getty Images의 영국/미국 동시 소송의 실익 2 : 실질적인 침해 행위는 어디서?



- 2023년 12월, 영국 고등법원은 Stability AI가 신청한 약식판결(Summary Judgment) 신청을 기각함.
- 2025년 6월, Getty Images는 영국 고등법원에서의 주요 저작권 침해 주장을 철회함 : 침해 행위가 대부분 미국에서 발생했고, UK 관할권 논리와 증거 부족 때문.
- 2025년 11월, **저작권 침해 청구 기각, 상표권 침해는 초기 버전 워터마크 출력에 한해 극히 제한적으로 인용.**

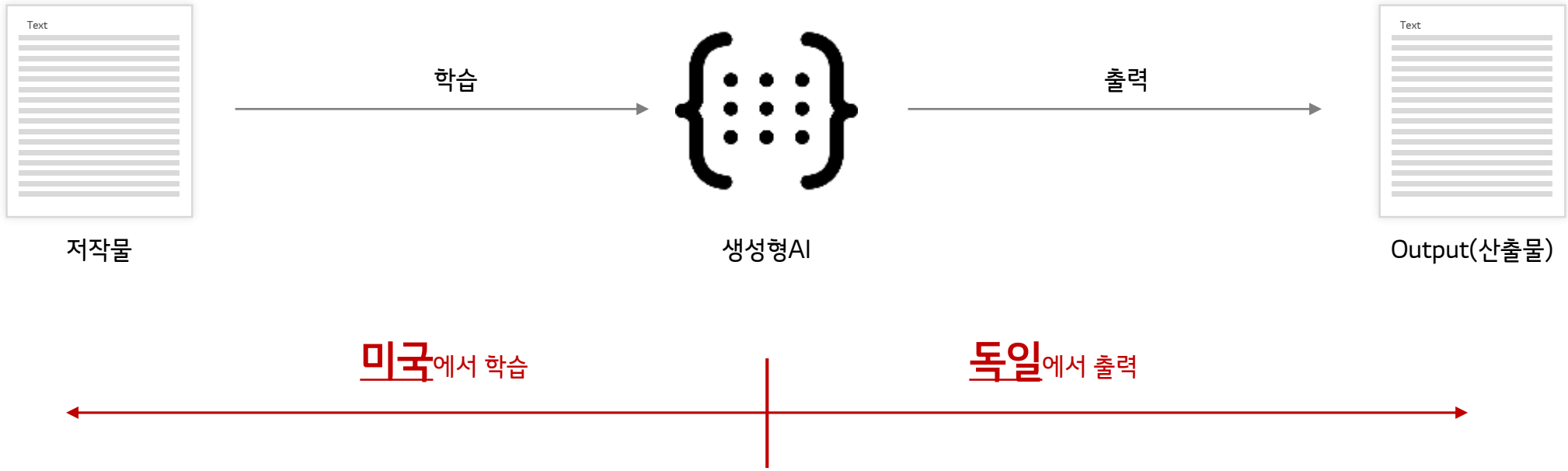
GEMA v. OpenAI (생성형AI 관련 독일의 첫번째 저작권 판결)

- 2024년 11월, 원고인 음악저작권단체 GEMA는 독일 뮌헨 지방법원에 OpenAI가 음악 가사를 허락 없이 LLM 학습에 이용하였다고 소를 제기하였음.
- 2025년 11월, 뮌헨 지방법원은 OpenAI가 저작물을 학습하고 재현(Output 출력)하는 과정에서 **저작권법을 위반하였고, 이는 독일 저작권법 §44b TDM 예외 범위를 초과**한다고 판결함.



GEMA v. OpenAI (생성형AI 관련 독일의 첫번째 저작권 판결)

- **관할:** OpenAI Ireland(EU 법인)에 대해서는 브뤼셀 I Recast(Regulation 1215/2012) 제7조 제2호의 불법행위 특별관할, 구체적으로 독일이 결과발생지라는 점에서 뮌헨 관할이 성립.
- **준거법:** 준거법은 로마 II Regulation(Regulation 864/2007) 제8조 제1항의 보호국법 원칙(lex loci protectionis)에 따라 독일법이 적용.



GEMA v. Suno

- Suno의 AI 모델이 GEMA 회원 작곡가들의 음악 저작물을 무단으로 학습하고 이를 기반으로 유사한 음악을 생성·출력한다는 것이 핵심 주장.
- GEMA v. OpenAI의 후속 사건으로 대상이 가사에서 음악 편곡으로 확장됐고 양 당사자가 훈련이 미국에서 이루어졌음을 인정한 상태에서 독일 법원이 관할을 행사하려 하려 했음.
- **관할:** Suno는 비EU 법인(미국)이므로 브뤼셀 I Recast가 아예 적용되지 않음. 따라서 GEMA v. OpenAI에서 쓴 브뤼셀 I 제7조 제2호 구성 자체가 불가능하여, GEMA가 원용한 것이 VGG §131인데, 집중관리단체의 권리 침해에 관한 분쟁에서 침해행위지 또는 침해자 주소지 법원에 전속관할을 부여하는 조항.
- **준거법:** 학습은 미국에서 이루어졌으므로 침해행위지는 미국임. 로마 II Regulation 제8조를 그대로 적용하면 학습 단계 침해에 대한 보호국은 미국이 되고, 따라서 준거법은 미국 저작권법이 되어야 한다는 논리가 성립함.

Open Source v. Open Data

Open Source

Open Data

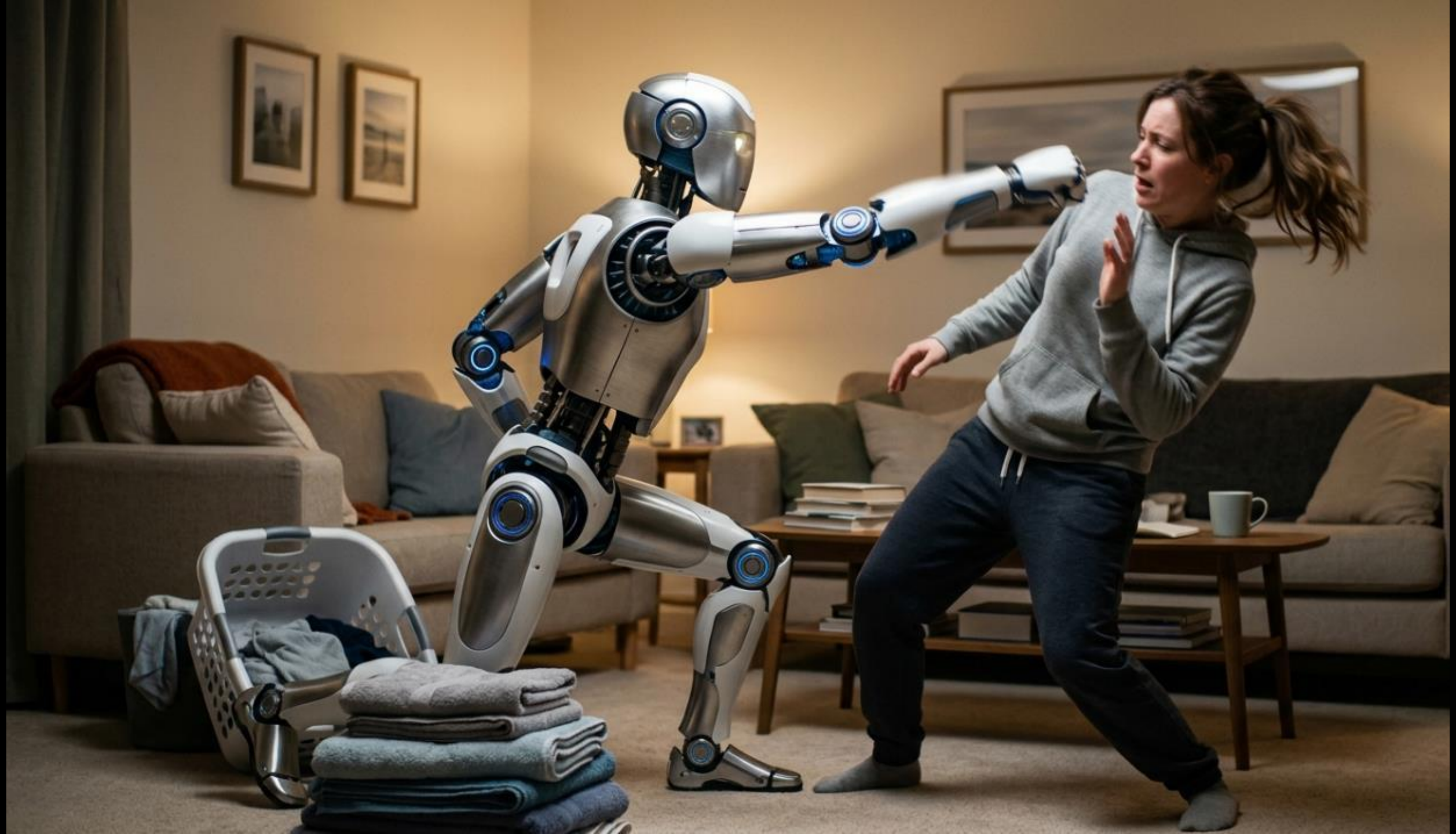
우리는 오픈데이터 컴플라이언스에 준비되어 있는가?

Chapter

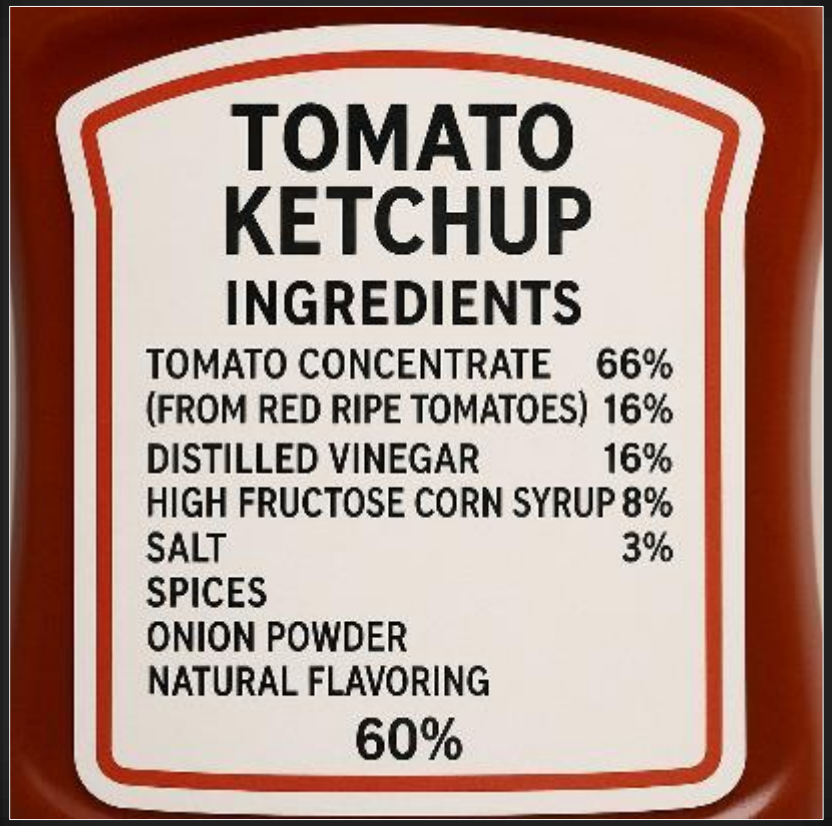
2

AI-BOM의 필요성과 Data Compliance

LG AI Research

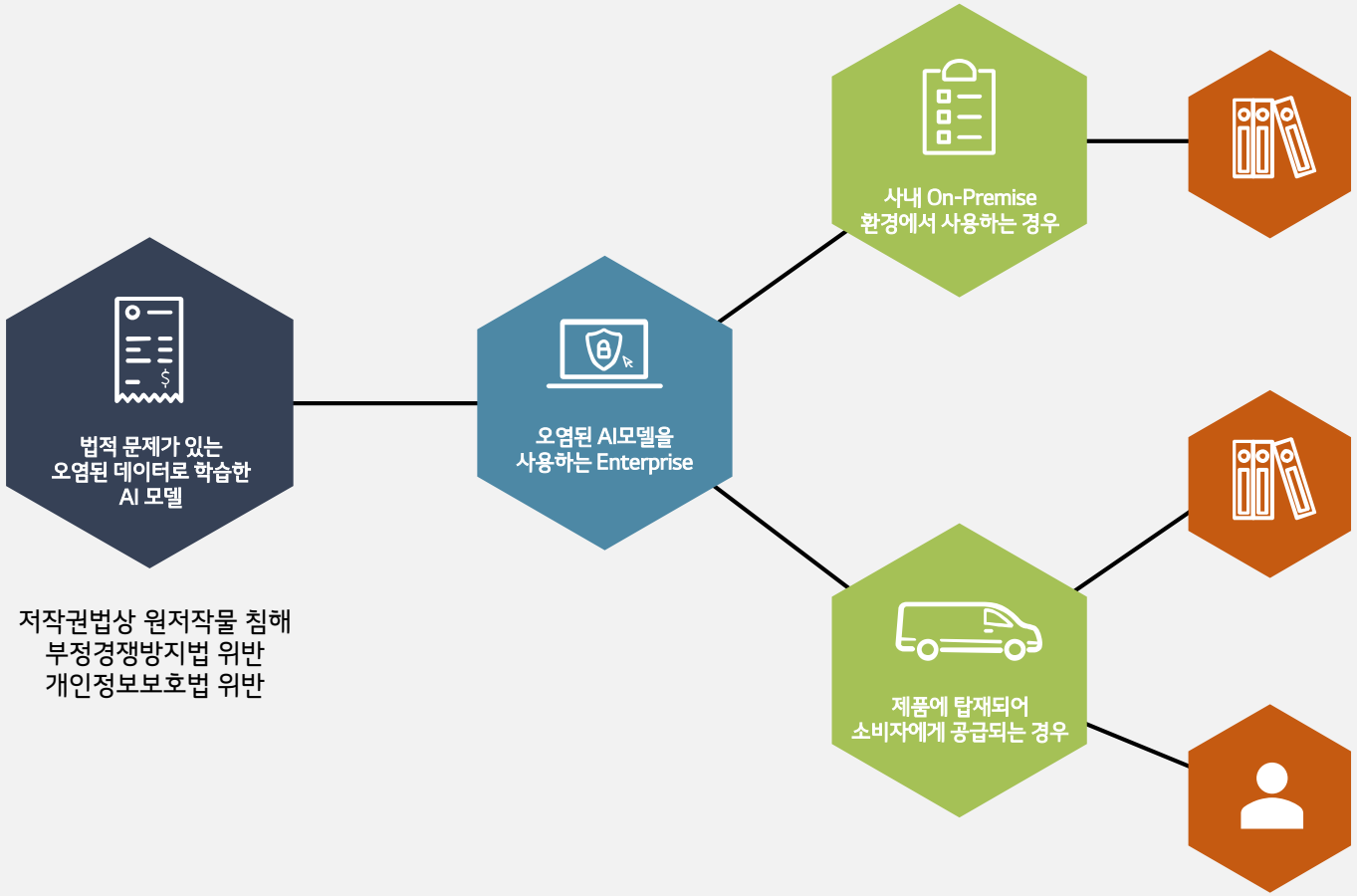


"구매자가 제품을 살 때, 그것이 무엇으로 만들어졌는지 명확히 알고자 하는 욕구는 인간의 본성 중 하나다."



Supply Chain에서 발생할 수 있는 법적 리스크

- 법적 문제가 있는 데이터를 학습데이터로 사용한 AI모델을 Enterprise에서 사용하게 되면, 데이터 저작자와의 분쟁으로 손해배상, 형사처벌, 과태료, AI모델 폐기 청구 등의 분쟁이 발생할 수 있으며, Enterprise 소비자에 대해 리콜을 시행하거나, 손해배상 및 형법상 고소될 수 있습니다.



저작권법상 원저작물 침해
부정경쟁방지법 위반
개인정보보호법 위반

데이터 저작자와의 분쟁

- 저작권법상 저작권 침해에 대한 민·형사상의 책임[1]
- 부정경쟁방지법상 부정사용행위, 퍼블리시티권 침해, 성과도용행위에 따른 민·형사상의 책임[2]
- 영업비밀 침해에 따른 민·형사상의 책임[3]
- 위법한 개인정보 이용에 따른 법적 책임[4]

데이터 저작자와의 분쟁

- 저작권법상 저작권 침해 방조행위에 대한 법적 책임[5]
- 부정경쟁행위 내지 영업비밀 침해행위의 방조행위에 대한 법적 책임
- 영업비밀 침해에 따른 민·형사상의 책임
- 위법한 개인정보 이용에 따른 법적 책임

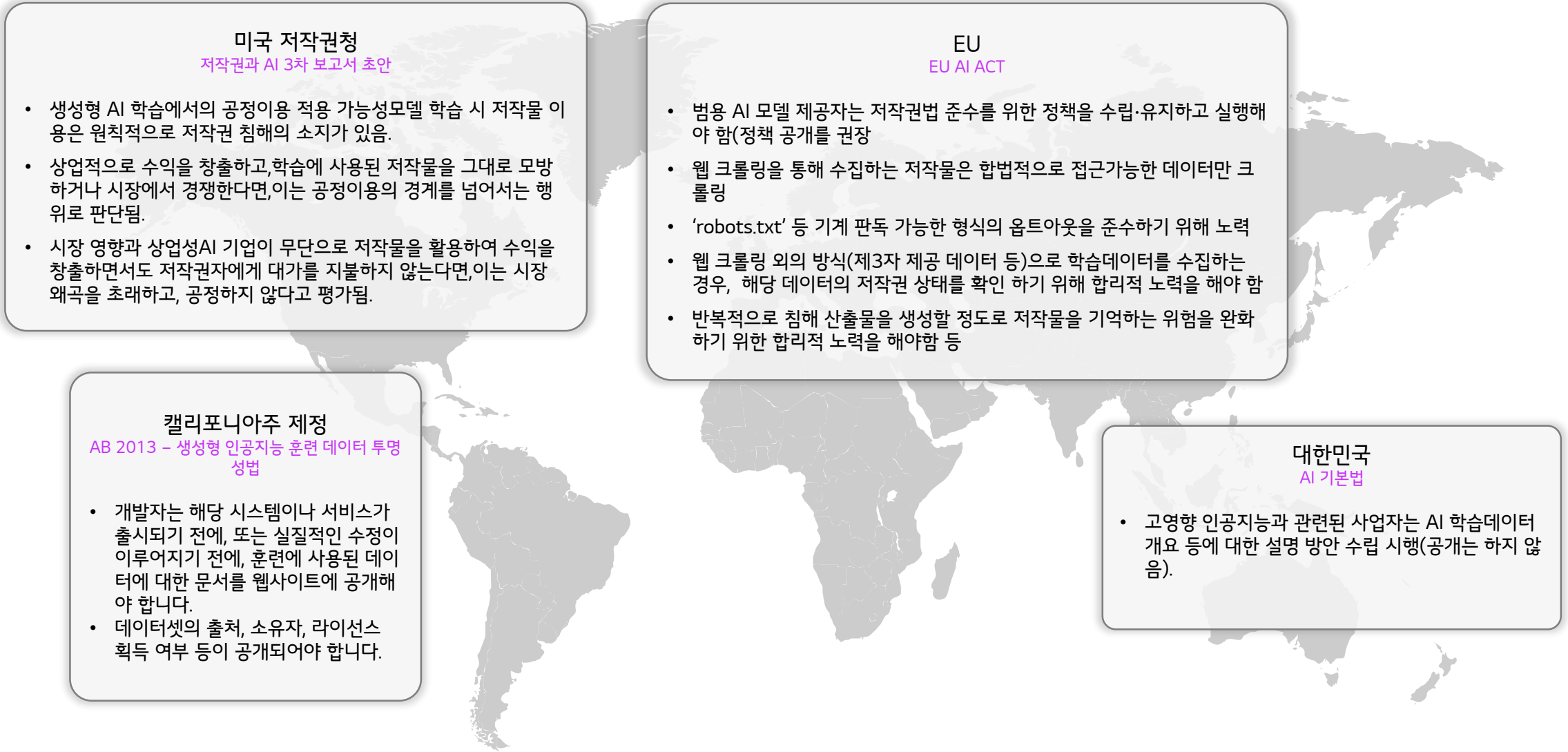
소비자와의 분쟁

- 소비자기본법상 소비자 재산에 위해를 끼친 것, 결함에 대해 고지의 의무를 다하지 않은 것에 대한 불법행위[6]
- 민법상 법률적 제한 내지 장애로 인해 하자에 해당하여 손해배상 책임 발생[7]
- 형법상 해당 하자가 발생할 것을 알고도 판매한 것에 대한 사기죄[8]

1: 저작권법 전반 및 제123조침해행위 정지청구 참고, 2: 부정경쟁방지법 제2조 제1호 (카)목 3), 3: 부정경쟁방지법 제2조 제3호 (나)목 및 (라)목 및 부정경쟁방지법 제2조 제3호 (다)목 및 (바)목 , 4: 개인정보보호법 제39조, 5: 대법원 2021. 11. 25. 선고 2021도10903 판결 참조, 6: 소비자기본법 제19조 제1항, 7: 민법 제580조, 민법 제582조 및 대법원 2000. 1. 18.선고 98다18506 판결, 대법원 2018. 7. 12. 선고 2015다64315 판결 참조, 8: 대법원 2008. 5. 8. 선고 2008도1652 판결 참조

AI-BOM의 필요성 - Data의 투명성과 관련된 법적, 윤리적 움직임

- 각 국가는 AI학습에 이용되는 데이터의 출처나 저작권법 준수 여부와 관련한 법률을 제정하거나 입법을 추진하고 있습니다.



미국 저작권청

저작권과 AI 3차 보고서 초안

- 생성형 AI 학습에서의 공정이용 적용 가능성모델 학습 시 저작물 이용은 원칙적으로 저작권 침해의 소지가 있음.
- 상업적으로 수익을 창출하고,학습에 사용된 저작물을 그대로 모방하거나 시장에서 경쟁한다면,이는 공정이용의 경계를 넘어서는 행위로 판단됨.
- 시장 영향과 상업성AI 기업이 무단으로 저작물을 활용하여 수익을 창출하면서도 저작권자에게 대가를 지불하지 않는다면,이는 시장 왜곡을 초래하고, 공정하지 않다고 평가됨.

EU

EU AI ACT

- 범용 AI 모델 제공자는 저작권법 준수를 위한 정책을 수립·유지하고 실행해야 함(정책 공개를 권장
- 웹 크롤링을 통해 수집하는 저작물은 합법적으로 접근가능한 데이터만 크롤링
- 'robots.txt' 등 기계 판독 가능한 형식의 옵트아웃을 준수하기 위해 노력
- 웹 크롤링 외의 방식(제3자 제공 데이터 등)으로 학습데이터를 수집하는 경우, 해당 데이터의 저작권 상태를 확인 하기 위해 합리적 노력을 해야 함
- 반복적으로 침해 산출물을 생성할 정도로 저작물을 기억하는 위험을 완화하기 위한 합리적 노력을 해야함 등

캘리포니아주 제정

AB 2013 – 생성형 인공지능 훈련 데이터 투명성법

- 개발자는 해당 시스템이나 서비스가 출시되기 전에, 또는 실질적인 수정이 이루어지기 전에, 훈련에 사용된 데이터에 대한 문서를 웹사이트에 공개해야 합니다.
- 데이터셋의 출처, 소유자, 라이선스 획득 여부 등이 공개되어야 합니다.

대한민국

AI 기본법

- 고영향 인공지능과 관련된 사업자는 AI 학습데이터 개요 등에 대한 설명 방안 수립 시행(공개는 하지 않음).

미국의 입법 (State law)

❑ 캘리포니아 AB 2013(제정), 워싱턴 HB 1168(입법 中)에서 **학습 세부 요약 내용 공개가 요구**되고, 버지니아 HB 2250에 학습 세부 요약 공개 및 학습 금지/삭제 요구권이 존재함.

캘리포니아 AB 2013

...

2026년 1월 1일부터, 2022년 1월 1일 이후에 발표된 AI시스템이나 서비스에 대해 생성형 AI 학습데이터 목록을 웹사이트에 고지해야 함.

3111. On or before January 1, 2026, and before each time thereafter that a generative artificial intelligence system or service, or a substantial modification to a generative artificial intelligence system or service, released on or after January 1, 2022, is made publicly available to Californians for use, regardless of whether the terms of that use include compensation, the developer of the system or service shall post on the developer's internet website documentation regarding the data used by the developer to train the generative artificial intelligence system or service, including, but not be limited to, all of the following:

(a) A high-level summary of the datasets used in the development of the generative artificial intelligence system or service, including, but not limited to:

(1) The sources or owners of the datasets.

(2) A description of how the datasets further the intended purpose of the artificial intelligence system or service.

(3) The number of data points included in the datasets, which may be in general ranges, and with estimated figures for dynamic datasets.

(4) A description of the types of data points within the datasets. For purposes of this paragraph, the following definitions apply:

데이터셋 출처, 데이터 포인트 수, 저작권 보호 여부, 라이선스 구매 여부, 개인정보 포함 여부, 정제/처리/수정 여부, 합성 데이터 생성 여부 등을 공개해야 함.

(A) As applied to datasets that include labels, "types of data points" means the types of labels used.

(B) As applied to datasets without labeling, "types of data points" refers to the general characteristics.

(5) Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.

(6) Whether the datasets were purchased or licensed by the developer.

(7) Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140.

(8) Whether the datasets include aggregate consumer information, as defined in subdivision (b) of Section 1798.140.

(9) Whether there was any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the artificial intelligence system or service.

(10) The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing.

(11) The dates the datasets were first used during the development of the artificial intelligence system or service.

(12) Whether the generative artificial intelligence system or service used or continuously uses synthetic data generation in its development. A developer may include a description of the functional need or desired purpose of the synthetic data in relation to the intended purpose of the system or service.

...

미국의 입법 (State law)

❑ 캘리포니아 AB 2013(Civil Code Section 3111) 시행에 따라, AI 기업들이 학습데이터 요약본을 공개하고 있음.

Training Data Summary Pursuant to California Civil Code Section 3111

Updated: 어제

OpenAI offers publicly available generative AI systems in the state of California. We develop these systems using a variety of data sources, including publicly available data, data that we partner with third parties to access, and information that our users or human trainers and researchers provide or generate. We also develop our systems using synthetic data.

We use data to help our systems better understand human language and the world. That, in turn, allows our systems to enhance human creativity, advance scientific discovery and medical research, and enable hundreds of millions of people to improve their daily lives. Our systems are developed on datasets containing trillions of tokens of textual, image, audio, and audiovisual content.

We use a diverse set of data to develop our systems, including data that may be protected by copyright and data in the public domain. Although we take steps to reduce the amount of personal information in our training datasets, some of our data may include personal information and aggregate consumer information as defined in California Civil Code Section 1798.140. Our users have the ability to opt-out of their content being used for training, as well as to request removal of certain personal information from ChatGPT responses, using our Privacy Portal. We use a variety of techniques to process our datasets to improve the performance and accuracy of our models.

OpenAI

Training Data Documentation

Pursuant to California Civil Code Section 3111 (AB 2013)

Anthropic, PBC — Claude Model Family

1. Dataset Sources and Ownership

Claude models are trained on a proprietary mix of data from the following source categories:

- **Publicly Available Internet Data:** Content from publicly accessible repositories as well as content crawled from the public web using Anthropic's general-purpose web crawler, which follows industry-standard practices with respect to robots.txt instructions. The crawler operates transparently and does not access password-protected pages, sign-in pages, or bypass CAPTCHA controls.
- **Data Acquired from Third-Parties:** Non-public data obtained from third-party providers through commercial arrangements and data partnerships.
- **Data Labeling Services and Contractors:** Data provided by data labeling services and paid contractors, including human preference selection, safety evaluation, and adversarial testing data.
- **User Data:** Data from Claude users who have not opted-out from model training.
- **Internally Generated Data:** Data generated internally at Anthropic, including synthetic data created through various methodologies.

Anthropic

G7 Vision - SBOM for AI

❑ G7의 Hiroshima AI Process 이후 사이버보안 실무그룹의 SBOM for AI 공개 등 AI의 투명성에 대한 논의가 지속되고 있으며, SBOM의 형태 로서의 논의가 진행되고 있습니다.

A SHARED G7 VISION ON SOFTWARE BILL OF MATERIALS FOR AI (Transparency and Cybersecurity along the AI Supply Chain)

3. Software Bill of Materials for AI

Properties

To allow an SBOM for AI to be effective, it needs to ensure that the three following properties are satisfied:

학습, 테스트, 평가에 사용된 데이터셋에 대한 SBOM

• being able to capture the static and dynamic aspects of AI systems (e.g., datasets used for training, testing and validation during the lifecycle of the system or learning outcomes) that distinguish them from traditional software systems;

• being able to be easily processed automatically and tool generated in a machine-readable format;

• being able to leverage structured data formats as much as possible, to ensure that the relevant information is available transparently upon demand to all the stakeholders.

Furthermore, it is equally important to clearly define the information set that an SBOM for AI should include, defined as its "minimum elements".

Example minimum elements

An SBOM for AI should be composed of a set of minimum elements to capture the distinctive features of an AI system, ensuring compatibility and providing an adequate level of transparency for all the stakeholders. It should automatically build upon information captured by each of the AI components, providing an understanding of the flow between the AI elements of the system. While some transparency mechanisms exist, this effort aims to highlight a core set of data fields that are machine-generatable and machine processable. It is important to highlight that these minimum elements represent recommendations to a reasonable extent and should be decided accordingly to the specific context of use. Here below is an exemplary set of high-level minimum elements for a G7 SBOM for AI framework, which may extend the information used for traditional software bill of materials (e.g., supplier name, component version), listed as clusters that can embed more detailed information on:

• Models used by the AI system, including basic information to identify the model, describe how the model was created, and spell out how the model is intended to be used.

• Learning, including the description of the training techniques and pipelines and information about training datasets in, e.g., datasheets for datasets.

• Datasets used during the whole lifecycle of the model, including basic information that documents the identity, creation, use, and provenance of data.

• Safety and security characteristics, such as a link or reference to the safeguards or guardrail implementations, safety alignment, compliance attestations and cybersecurity best practices adopted during the AI lifecycle.

• System level characteristics, such as a link or reference to a description of the flow between the AI elements and how the model consumes input data.

• Key Performance Indicators of an AI system, including model benchmark evaluation results.

• Licensing information about the components of an AI system.

• Infrastructure used by the AI system, including the software components specifically required to deliver an AI system.

The list is open for further expansion of the clusters in the future to keep pace with the rapid development of technology.

To increase trustworthiness and to avoid giving a false sense of security, an SBOM for AI should be verifiable as a whole. This implies not only the verification of its individual components - e.g., via cryptographic hashes or digital signatures from the corresponding manufacturers - but also of the entire SBOM for AI. In order to achieve this goal, a viable SBOM for AI should at least be digitally signed by its manufacturer. While individual components are signed within the SBOM for AI, the signature of the entire SBOM for AI has to be verifiable from the outside.

AI-BOM 표준

❑ 각 단체들은 AI BOM 표준을 준비하고 있으며, OpenChain의 AI BOM에서는 직접적으로 Governance 체계에 EU AI Act를 지목하는 등 AI BOM은 AI기본법과 맞물려 발전할 것으로 판단됨.

SPDX 3.X의 데이터셋 field

AI 학습데이터와 관련하여 이용목적, 주소, 타입, 사이즈, 접근가능성, 개인정보포함여부, 익명화 여부 등 학습데이터로 부터 발생할 수 있는 쟁점에 대해 리스트업 되어 specification으로 작성되어 있음.

Properties

Property	Type	minCount	maxCount
anonymizationMethodUsed	xsd:string	0	*
confidentialityLevel	ConfidentialityLevelType	0	1
dataCollectionProcess	xsd:string	0	1
dataPreprocessing	xsd:string	0	*
datasetAvailability	DatasetAvailabilityType	0	1
datasetNoise	xsd:string	0	1
datasetSize	xsd:nonNegativeInteger	0	1
datasetType	DatasetType	1	*
datasetUpdateMechanism	xsd:string	0	1
hasSensitivePersonalInformation	/Core/PresenceType	0	1
intendedUse	xsd:string	0	1
knownBias	xsd:string	0	*
sensor	/Core/DictionaryEntry	0	*

External properties cardinality updates

Property	minCount	maxCount
/Core/Artifact/builtTime	1	
/Core/Artifact/originatedBy	1	1
/Core/Artifact/releaseTime	1	
/Software/Package/downloadLocation	1	
/Software/SoftwareArtifact/primaryPurpose	1	

OpenChain AI System BOM

AI 학습데이터와 관련하여 이용목적, 주소, 타입, 사이즈, 접근가능성, 개인정보포함여부, 익명화 여부 등 학습데이터로 부터 발생할 수 있는 쟁점 등에 대한 거버넌스 체계.

3.9 AI System Bill of Materials

A process shall exist for creating and managing an AI SBOM, this can be in any format e.g. SPDX, CycloneDX, or another format.

The AI SBOM shall account for inbound materials from third-parties.

Verification material(s):

• A documented procedure for identifying, tracking, reviewing, approving, and archiving information related to the components of an AI system (e.g., model, datasets, etc).

• Records for the supplied system that demonstrates the documented procedure was properly followed.

Rationale:

• To ensure a process exists for creating and managing an AI SBOM used to construct the supplied system. A bill of materials is needed to support the systematic review and approval of the system to understand the obligations and restrictions

OWASP GenAI Security Project

OWASP 에서는 AI-BOM을 오픈소스로 추출&생성할 수 있는 tool을 공개하는 등, AI Supply chain에서 Vulnerability 관점에서 AI-BOM을 관리하는 Tool을 제공하고 있음.

AI-BOM의 포맷은 동일한 OWASP 내의 CycloneDX 포멧을 따름.

* CycloneDX 1.7에서는 ML-BOM 포맷을 간접적으로 지원함

AI 개발자의 고민

법적 리스크를 어떻게 헷징해야 하나?

우리는 어떤 데이터를 이용해야 하나?

우리는 어떻게 해야하는가..?

.

.

그런데, 무엇보다 우리 회사가 어떤 데이터를 학습시키는지 알고 있는지?

법원의 판단

데이터 투명성 법률 제정

AI 개발자

Data Compliance

- Data Compliance 가이드라인을 통해, Data lifecycle 관점에서 각 담당자가 Data의 이용 과정에서 발생할 수 있는 risk를 탐지합니다.
- Data Compliance Report(평가 기준)을 통해, 학습 데이터에 대한 법적 리스크를 평가합니다.

DATA COMPLIANCE 가이드라인

- AI 연구·개발 과정에서 발생할 수 있는 데이터 관련 법률 리스크
- LG AI연구원의 데이터 취급 관련 표준 업무 프로세스
- Data Compliance 체크리스트 (수집, 보관, 파기, 제3자 제공/위탁)

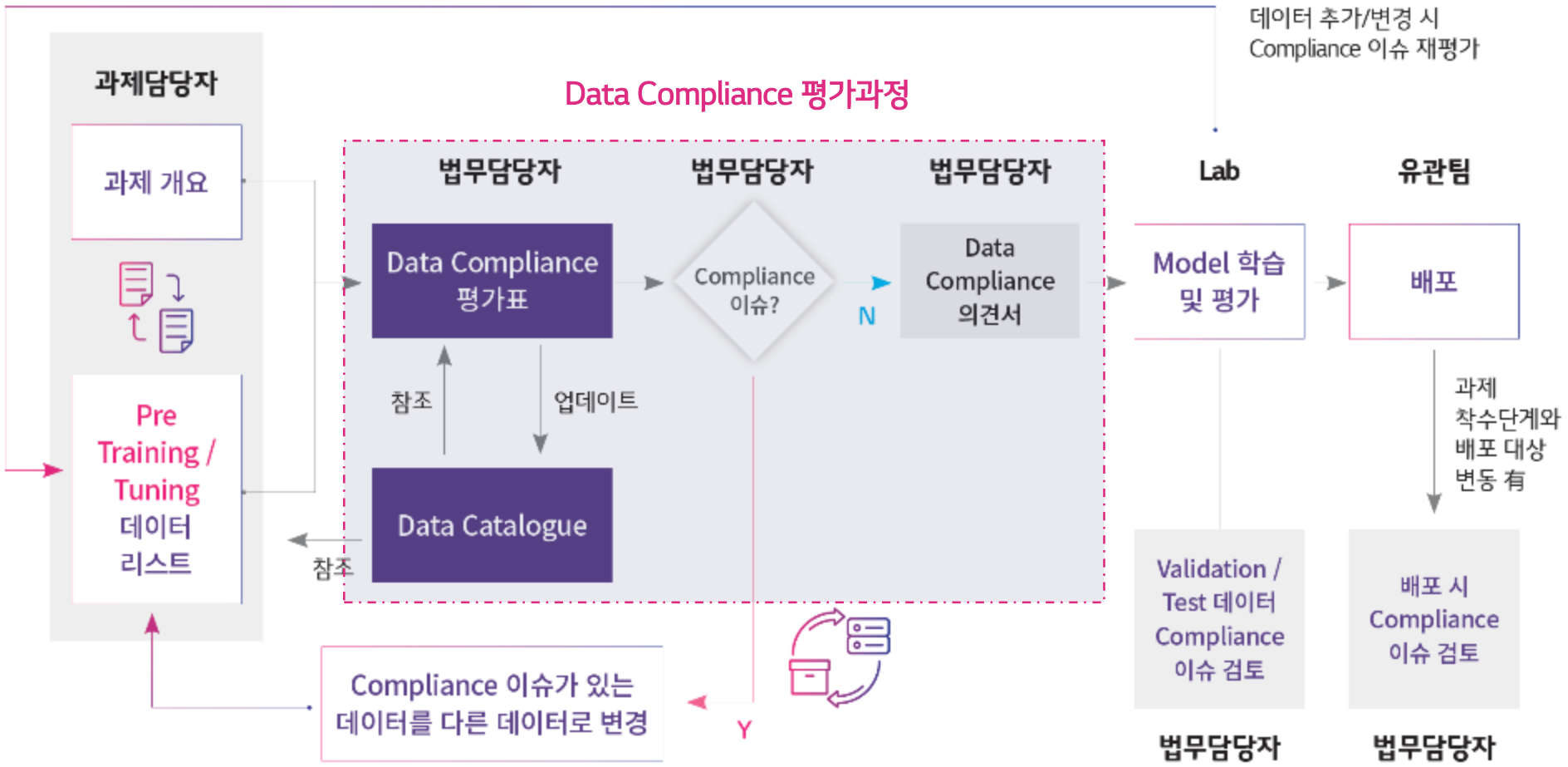
DATA COMPLIANCE REPORT

- Data Compliance Score 및 Class 평가 기준(ENG)

Data Compliance

- Data Compliance 프로세스를 통해, 모든 학습 데이터를 검토하고 평가합니다.

Data Compliance 프로세스



Data Compliance

- 평가의 카테고리는 총 4가지로, 1. Data License 항목, 2. Data 사용 기간 및 지역 관련 항목, 3. 개인정보 및 데이터 보안 관련 항목, 4. 추가적인 법적 리스크 관련 항목으로 분류됩니다.

1. Data License 항목		小 ← 리스크 → 大				
Data에 대한 License 부여 여부	1-1) License 부여 여부	상업적 목적으로 제한없이 사용 가능		명시적으로 내부 연구 목적으로 사용 가능	Unknown	명시적으로 사용 불가
	1-2) 데이터 수정 가능 권한 및 2차적 저작물 작성 권한	2차적 저작물 작성을 포함한 모든 수정과 변형 권한 부여		2차적 저작물 작성은 불가능하나, 2차적 저작물에 이르지 않는 수준의 수정 또는 변형 가능	Unknown	명시적으로 2차적 저작물 작성을 포함한 여하한 수정과 변형 불가
원저작자의 권리 침해 가능성	1-3) 산출물(Output)의 원저작권 침해 가능성	원저작자 동의 받거나, 산출물이 생성되지 않음	원저작물과 다른 형태의 산출물 생성됨	원저작자의 저작물과 유사한 산출물 생성 가능성 있으나, 높지 않음	원저작자의 저작물과 유사한 산출물 생성가능성과 무관하게, 원저작자의 저작물 중 일부가 산출물에 포함될 가능성이 있음	원저작자의 저작물과 유사한 산출물 생성 가능성 높음
산출물에 대한 권리	1-4) Prompt, Output에 대한 권리 부여 여부	Prompt, Output에 대해 당사에 소유권, 지재권 있음	Prompt, Output에 대해 당사에 사용권한 있음	Prompt, Output에 대한 당사의 권리 Unknown		명시적으로 Prompt, Output에 대한 당사의 권리 없음
데이터 고지의 의무	1-5) 데이터 고지의 의무 존재 여부	(고지가능 시) 데이터 고지의 의무 존재 혹은 부존재	(고지불가 시) 데이터 고지의 의무 미존재		(고지불가 시) 별도 데이터 고지의 의무 존재	
2. Data 사용 기간 및 지역 관련 항목						
Data 사용 기간	2-1) Data 사용 기간의 제한	데이터 영구적으로 사용 가능	운영에 문제 없을 수준의 데이터 기간 제한 존재하거나 데이터 기간 존재하지 않음	데이터는 기간 제한 있으나 AI 모델에 대한 제한 Unknown		데이터 사용 기간이 이미 경과함
	2-2) Data 라이선스 부여의 철회 가능성	라이선스 부여철회불가	Unknown	라이선스 부여 철회가능		
AI모델 사용 기간	2-3) AI모델 서비스 기간의 제한	AI모델 서비스 영구적인 제공 가능		Unknown		AI모델 서비스 영구적인 제공 불가
Data 사용 지역	2-4) Data 사용 지역의 제한	Worldwide	Unknown		특정지역(한국, 미국, EU를 포함)에서만 사용가능	특정지역(한국, 미국, EU를 포함하지 않음)에서만 사용가능

Data Compliance

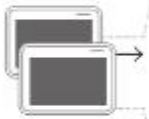
- 평가의 카테고리는 총 4가지로, 1. Data License 항목, 2. Data 사용 기간 및 지역 관련 항목, 3. 개인정보 및 데이터 보안 관련 항목, 4. 추가적인 법적 리스크 관련 항목으로 분류됩니다.

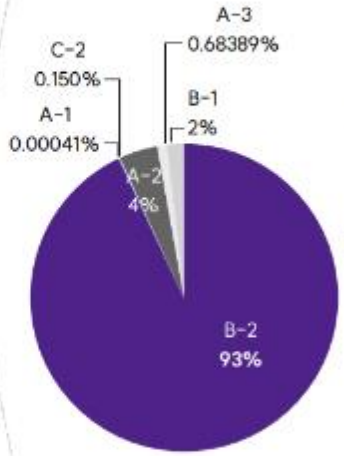
3. 개인정보 및 데이터 보안 관련 항목		小 ← 리스크 → 大				
개인정보 포함여부	3-1) 개인정보의 포함 여부	개인정보 포함 되어 있지 않거나, 익명 비식별화 처리 계획이 있음	개인정보 포함 되었으나, 가명 처리 계획이거나 Unknown		개인정보 포함 가능성이 높음	개인정보 포함 되었음
정보주체의 동의 수취 여부	3-2) 정보주체의 동의 여부	개인정보 포함 되어 있지 않거나, 개인정보 포함되어 있으나 정보주체의 동의를 받음				개인정보 포함 되어 있으나 정보주체의 동의를 받지 않음
가명정보 포함여부	3-3) 가명정보의 포함 여부	가명정보 포함 되어 있지 않거나, 가명정보 포함되어 있으나 정보주체의 동의를 받음		개인정보 포함 되어 있고 이를 가명 처리할 계획임		가명정보 포함 되어 있음
Data 위탁/제3자 제공 가능 여부	3-4) Data 위탁/제3자 제공 가능 여부	Data 제3자 위탁/제공가능 권리 있거나, 명시적 제한 없음		Unknown		명시적으로 Data 제3자 위탁/제공 불가
Data 사용권한 제한 여부	3-5) Data에 대한 특정 이용자 사용 제한	Data 이용에 대해 특정 이용자에게만 사용권한이 부여되어 있지 않음		Data 이용에 대해 특정 이용자에게만 사용권한이 부여되어 있음		
4. 추가적인 법적 리스크 관련 항목						
Data 수집 신뢰성	4-1) Data 수집 과정에서의 적법성	적법한 방법으로 획득		Web Crawling 등을 통하여 Data 획득	Unknown	Robots.txt를 회피하거나 적법하지 않은 방법으로 획득
Data 분쟁 여부	4-2) Data가 이용된 AI모델에 대한 알려진 분쟁	알려진 Data 분쟁 없음	알려진 Data 분쟁 존재하나, 개인 대상 소액 분쟁	10억 이상의 분쟁 존재함		100억 이상의 분쟁 존재함
계약적 Risk	4-3) License 계약의 추가 리스크 존재함	알려진 추가 계약 리스크 없음	(데이터 보안, 비밀유지의무 등) 까다로운 관리 체계 요구함	책임한도 무한, 자유로운 Audit 가능한 경우 등의 리스크 존재함		
License 조건의 유형	4-4) License 계약의 추가 리스크 존재함			재배포가 가능함	Share-Alike 수준의 요구 사항	재배포가 불가함

Data Compliance

- Data 뿐만 아니라, 평가된 Data를 이용하여 개발된 모든 AI 모델의 리스크를 평가하여 리스크를 Classification 하고, 용도에 맞게 AI를 이용합니다.
- 지속적인 Data 및 AI의 Data Compliance 체계 수행으로 저작권법, 개인정보보호법, 부정경쟁방지법 준수와 다가올 국내외 AI 기본법을 대비하고자 합니다.

Sample Review Report | Review reports analyzing AI model risks based on data to be used for AI training are prepared and delivered to researchers as follows:





AI Model Risk Assessment Score

- **Grade**: C-2 (Score: 3.60)
- **Legal Risk**

High potential for infringement due to licensing and privacy issues, with precedents of significant litigation, arbitration, or regulatory actions, and significant risk of our company becoming involved in disputes.

- **Usage Reference Standard**

Not recommended for use (if absolutely necessary, temporary use only for internal research purposes).

Key Data Characteristics

- **Open Crawling Data**
  - B-2 class is evaluated at 93%, where B-2 class is data suitable for internal use purposes.
  - When using crawled data, risk can be reduced by:
    1) Data anonymization,
    2) Ensuring legality of web crawling methods,
    3) Excluding text that could qualify as literary works.
- **Cases where Final Class is Downgraded**
  - OOOO, XXX XXXX, YY-UU are C-2 class data. This data is not suitable for AI model training.
  - If C-2 class data is deleted, AI Model Risk assessment score is expected to improve and Risk Class is expected to upgrade to B-2.

Evaluation Criteria: 18 Legal Perspectives | Data is protected by various laws including copyright law, personal information protection law, and unfair competition prevention law in each country, with legally permissible uses varying by jurisdiction. Considering these characteristics, our AI-based data compliance system reviews potential risks across 18 legal perspectives:

Data Risk Classification Table | Datasets are classified into three major grades (A, B, C) and seven sub-categories based on risk assessment results.

Category	License/Privacy	Key Legal Risks
Grade A	Risk-Free	Very low likelihood of legal disputes or risks.
Grade B	Medium Risk	High likelihood of violations related to licenses or privacy issues.
Grade C	High Risk	Very high likelihood of legal dispute escalation.

현대 AI 학습 데이터 검토의 어려움

- 엄청난 파라미터 사이즈를 갖는 모델들이 주류가 되면서, 학습 데이터셋 또한 완전 새로운 데이터로만 구성 된 단순 구조 형태가 아닌, 성능 향상에 필요한 다양한 소스부터 수집 된 데이터들을 복합적으로 섞은 거대한 수직 계층형 구조를 갖게 되었습니다.

Mean	Std	Min	25%	50%	75%	Max
2.20	1.83	0	1	2	3	16

3,612 데이터셋의 계층 레벨(Depth) 동계

Mean	Std	Min	25%	50%	75%	Max
23.18	87.15	1	2	5	11	1691

3,612 데이터셋의 Data Source 동계

현대 AI 학습 데이터 검토의 어려움

- 실제로 Open-sourced 학습데이터를 이용하는 경우, 해당 학습데이터의 모든 원본 데이터에 대한 법적 리스크를 검토해야 하는 어려움이 존재합니다.

검토가 필요한 Dataset

↓

검토가 필요한 원 Dataset의
License Terms 탐색 및 판단

↓

원 Dataset을 구성하고 있는
모든 Sub-sources를 탐색

↓

개별 Sub-source의
Terms 탐색 및 판단

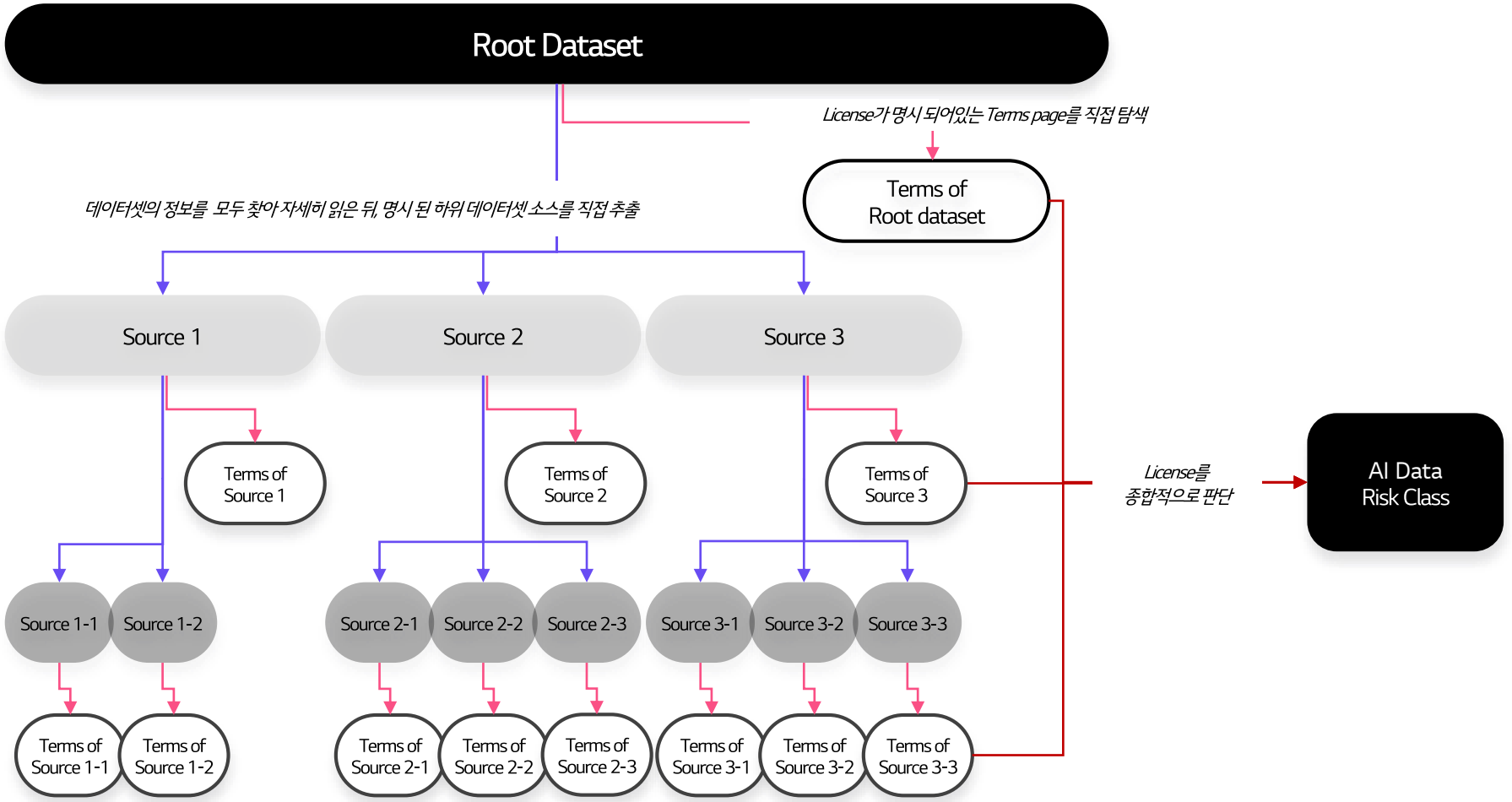
↓

개별 Sub-source의
모든 Sub-sub-sources를 탐색

↓

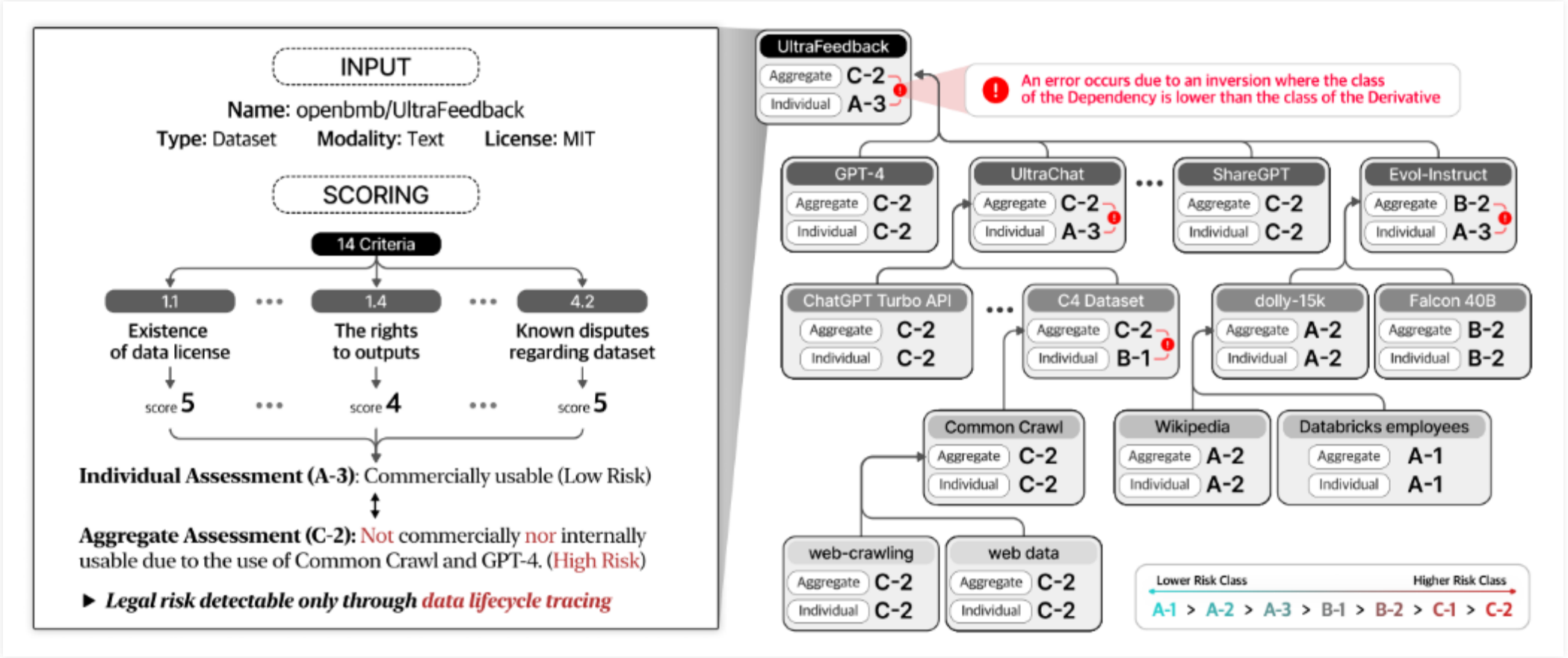
개별 Sub-sub-sources의
Terms 탐색 및 판단

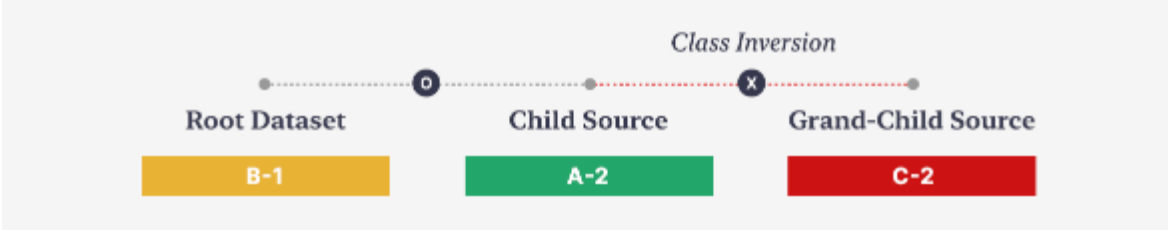
모든 법적 검토가 필요한 소스들을 다 찾을 때 까지 반복



EXAONE NEXUS를 활용한 Data Compliance 자동화 구조

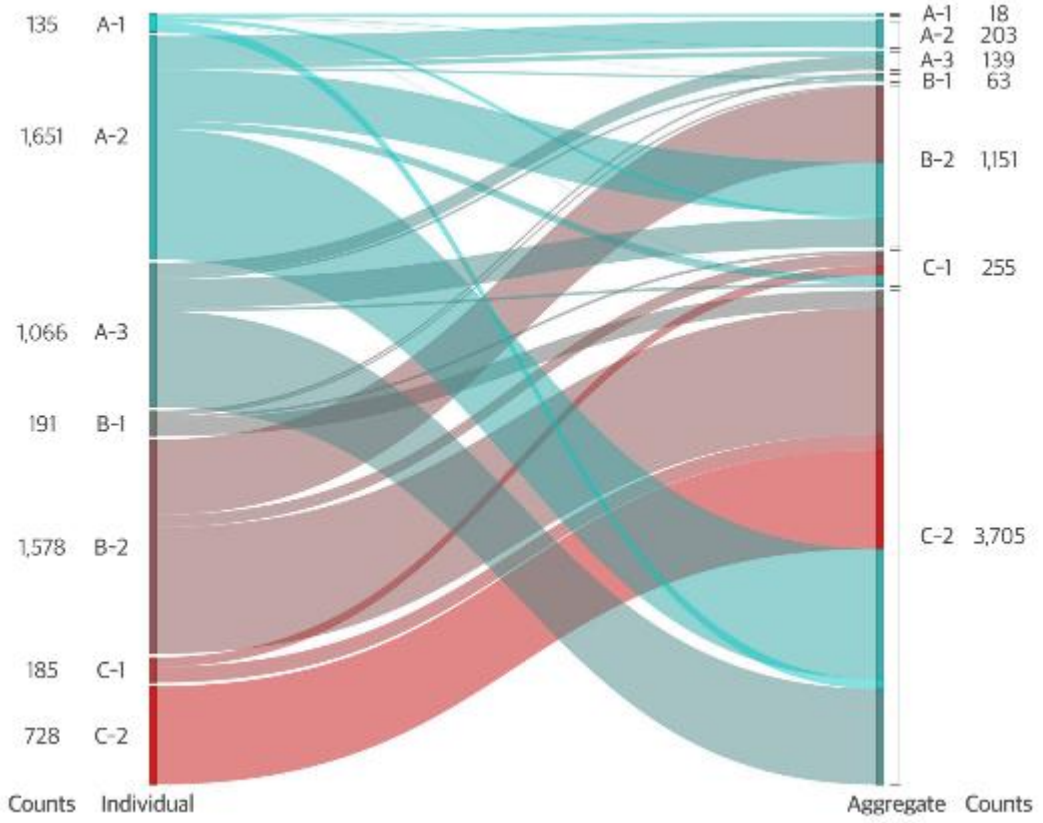
- EXAONE NEXUS는 사용하고자 하는 단일 데이터 셋이 아닌, 데이터 셋이 갖고 있는 모든 출처가 되는 데이터 셋의 법적 위험을 Data의 Life Cycle 측면에서 탐지합니다.

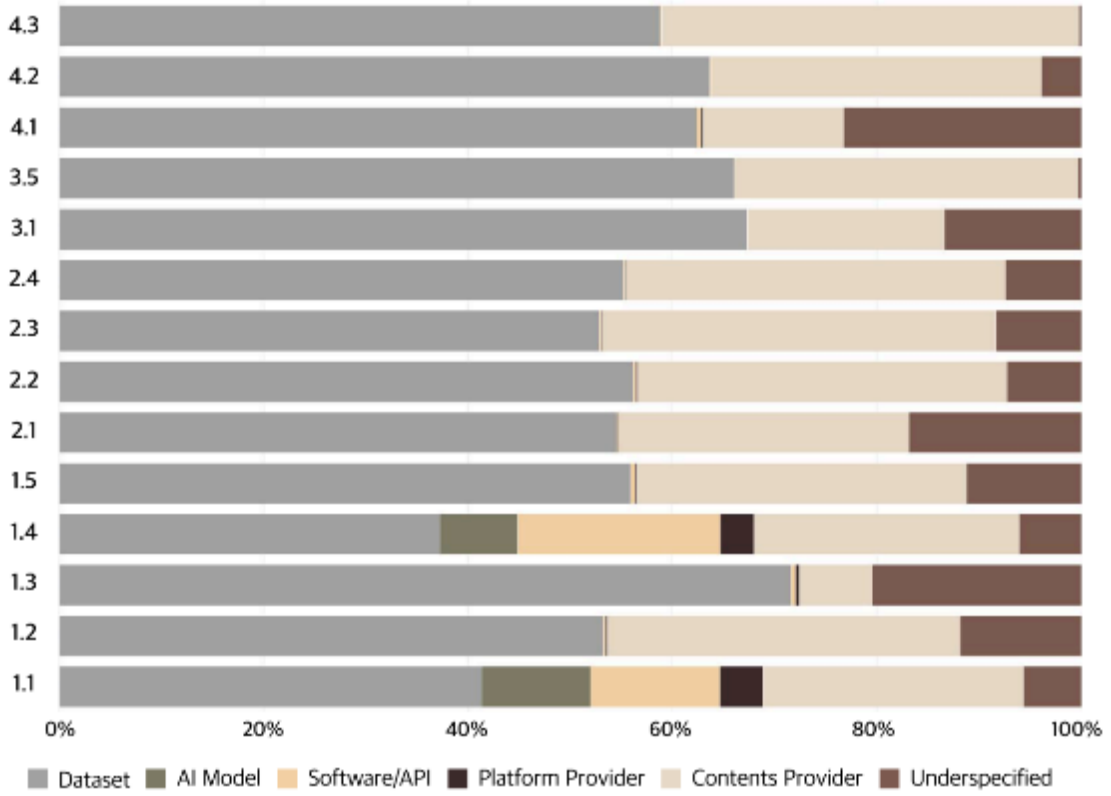




잠재적 위험

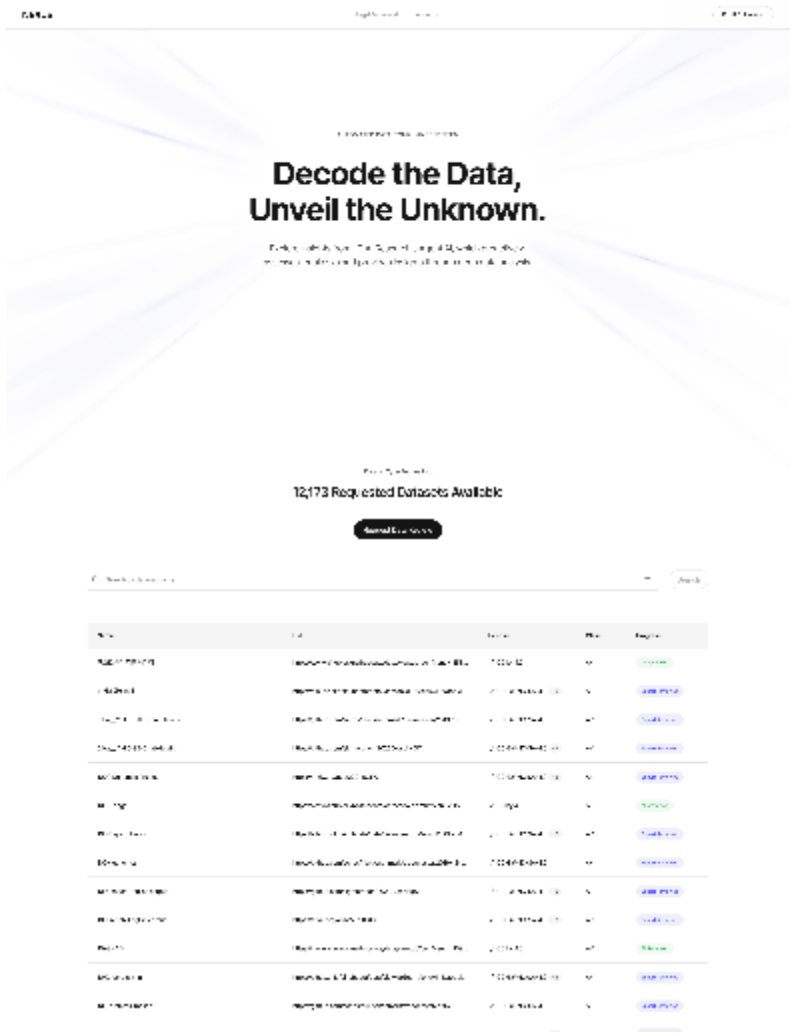
- 상업적으로 이용 가능하다고 판단된 2,852개의 AI 학습 데이터셋 중 종속 데이터의 리스크를 모두 고려해 본 결과, **21.21%인 605개의 데이터셋만 상업적으로 이용 가능했습니다.**



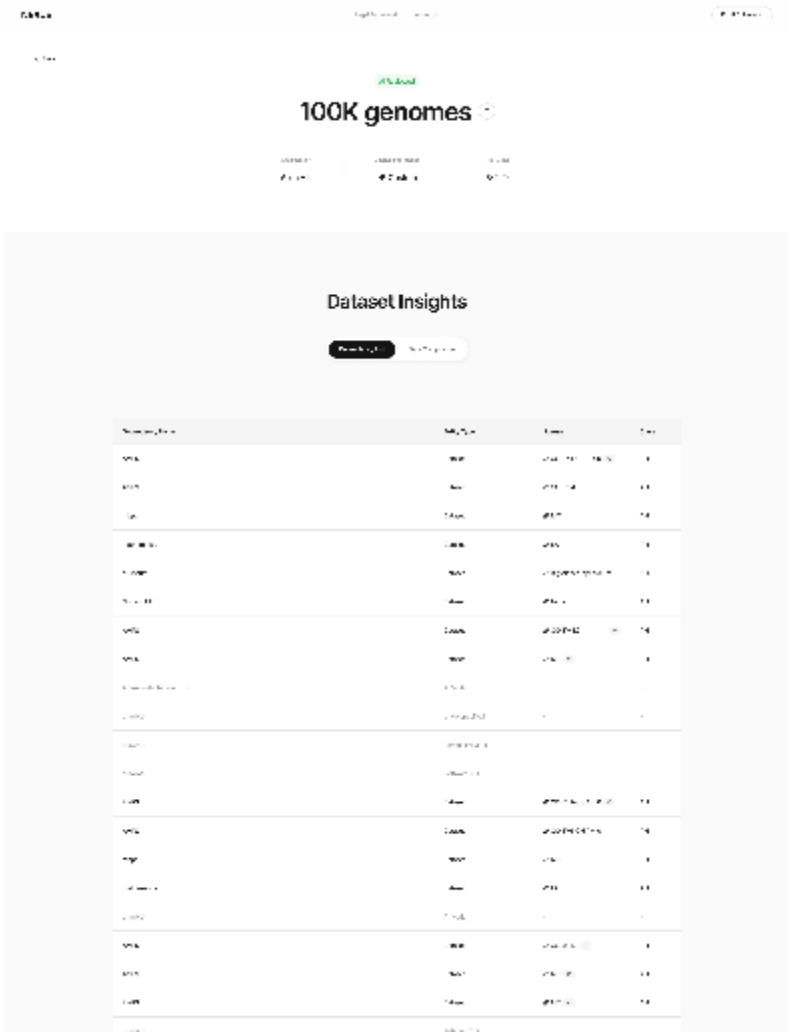


EXAONE NEXUS (http://nexus.lgresearch.ai/)

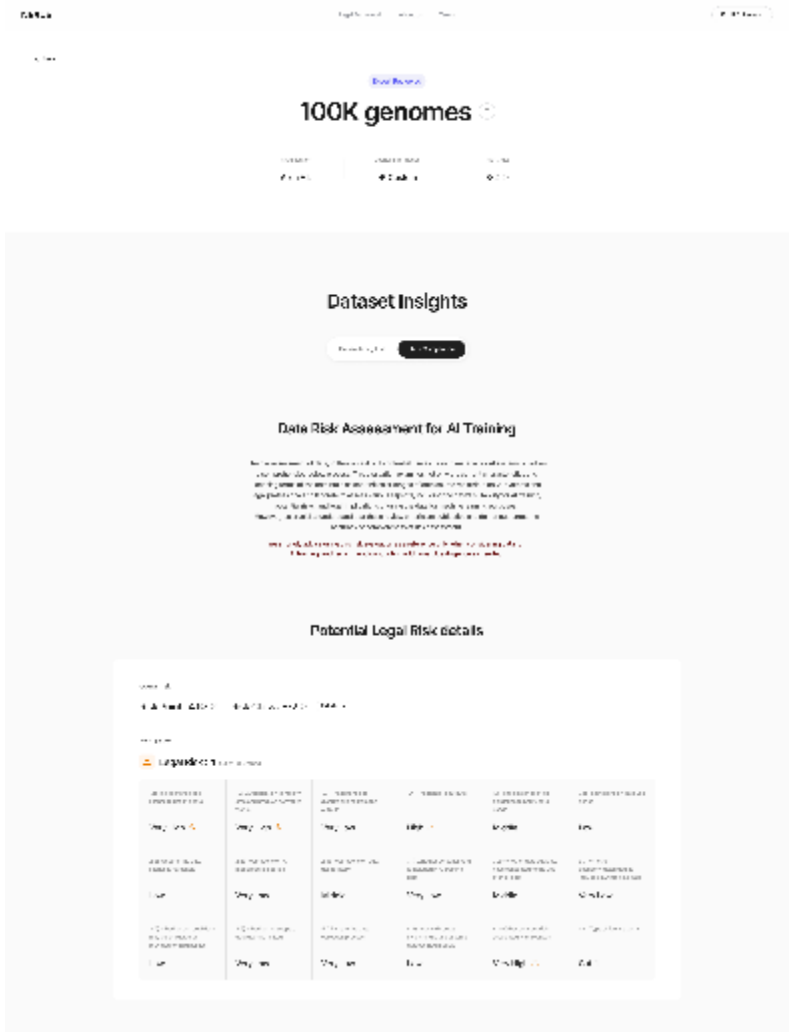
- EXAONE NEXUS를 통해 Agent가 검토한 데이터셋의 검토 결과를 직접 확인 할 수 있습니다.



Dataset List



Data Dependency List



Data Compliance Results

Thank you

legal@lgresearch.ai

LG AI Research