

2025 오픈소스 페스티벌

# 오픈데이터 AI Compliance

책임변호사 조정원

Dec, 2025

## 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.



LG  
LG 계열사




외부 파트너사

글로벌 최신 AI 기술 연구를 선도하며,  
누구나 전문가로 성장하고  
전문가는 더 높은 가능성을 실현할 수 있는  
세상을 만들어갑니다.

Advancing AI for a better Life



정책기관



학계, 연구기관

## Contents

---

- 1 AI 개발 및 서비스 과정에서의 리스크
- 2 Data Compliance Framework
- 3 EXAONE NEXUS: Open Data Compliance Agent

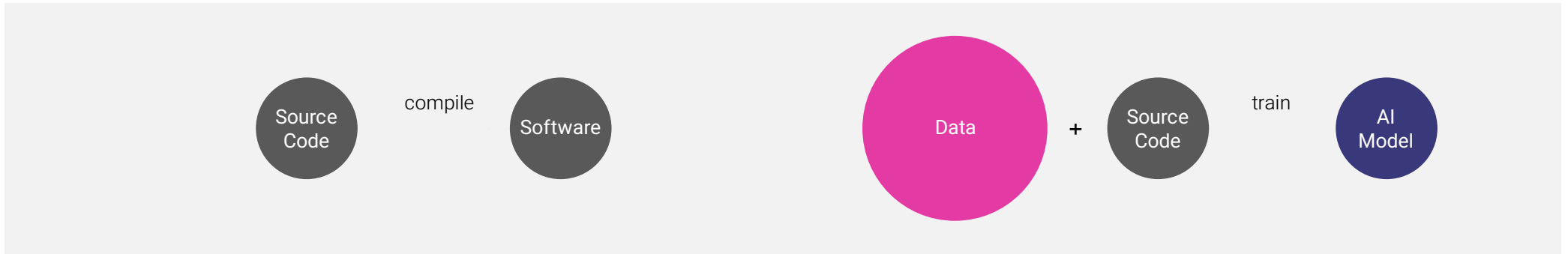
Chapter

1

AI 개발 및 서비스  
과정에서의 리스크

## — 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

## — 시개발사의 항변 - 공정이용 법리

- 피고의 항변 : 저작권법 상 저작물을 원저작자의 허락 없이도 이용할 수 있는 공정이용(Fair Use)의 법리

### 미국 17 U.S.C. 107

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The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

2012.3.15  
한-미 FTA  
발효시점

### 대한민국 저작권법 제35조의

5

제35조의5(저작물의 공정한 이용)

- ① 제23조부터 제35조의4까지, 제101조의3부터 제101조의5까지의 경우 외에 저작물의 일반적인 이용 방법과 충돌하지 아니하고 저작자의 정당한 이익을 부당하게 해치지 아니하는 경우에는 ~~보도·비평·교육·연구 등을 위하여~~ 저작물을 이용할 수 있다.
- ② 저작물 이용 행위가 제1항에 해당하는지를 판단할 때에는 다음 각 호의 사항 등을 고려하여야 한다.

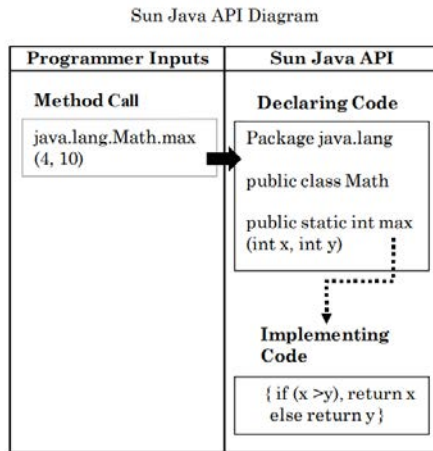
1. ~~영리성 또는 비영리성 등~~ 이용의 목적 및 성격
2. 저작물의 종류 및 용도
3. 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성
4. 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향

공정이용 4요소	고려할 사항
제1요소 : 원저작물 이용의 목적과 성격	원저작물의 이용이 상업적인가? 원저작물의 이용이 변형적인가? 원저작물의 이용이 악의나 고의에 의한 이용인가? 원저작물의 이용이 합리적이고 관습적인가?
제2요소 : 원저작물의 성격	원저작물이 창작적 저작물 혹은 사실적 저작물인가? 원저작물이 발행된 저작물인가?
제3요소 : 이용된 부분의 양과 중요성	원저작물을 이용한 양적, 질적 수준이 목적과 관련하여 합리적인 수준인가? (제1요소 및 제4요소와 상호 보완 관계)
제4요소 : 원저작물의 잠재적 시장 또는 가치에 미치는 영향	원저작물의 이용이 원저작물의 잠재적 시장 또는 가치에 얼마나 영향을(손해를) 미치는가?

## — 공정이용의 제1요소 : 이용의 목적과 성격

- 공정 이용의 제4요소와 함께 제1요소에서 그 이용 목적이 얼마나 변형적 이었는가에 대한 쟁점이 주목되고 있음.

### Oracle v. Google 사건



	Oracle v. Google 판결 (2021)	Goldsmith v. Andy Warhol Foundation 판결 (2023)
공정이용 인정	BREYER, Roberts, <u>Sotomayor</u> , Kagan, <u>Gorsuch</u> , <u>Kavanaugh</u>	Kagan, Roberts
공정이용 부정	THOMAS, Alito, Barret	<u>Sotomayor</u> , Thomas, Alito, <u>Kavanaugh</u> , Barret, <u>Goruch</u> , Jackson

- Oracle은 Java API의 “선언부(Declaring)” 코드를 Google이 안드로이드 운영체제에서 무단 사용한 것이 저작권 침해에 해당한다고 소를 제기함.
- 1심 1차 (2012) : Java API의 저작권 인정되지 않아 Google 승소.  
2심 1차 (2014) : API도 저작권 보호 대상이라는 판결로, Oracle 승소 → 파기환송.  
1심 2차 (2016) : 공정이용 인정으로 Google 승소. (배심원 평결)  
2심 2차 (2018) : 공정이용 부정으로 Oracle 승소.  
3심 (2021) : 대법원에서 공정이용이 최종 인정으로 Google 승소.
- 1. **이용의 목적과 성격** : Google의 사용은 Java의 기존 용도를 대체하는 것이 아니라 새로운 창의적 표현의 기초로 사용된 것으로서 **"변형적"** 성격이 있다며, 이 요인은 Google에게 유리하게 판단됨.
- 2. **원저작물의 성격** : 창작성이 낮은 기능적 코드이기 때문에, 이 요인은 Google에 유리하게 판단됨.
- 3. **이용된 부분의 양과 중요성** : 복제된 부분은 제한적이며(전체 Java API 중 전체의 0.4%인 Declaring Code 11,500줄만 복사), 기능적 구조를 위한 최소한의 사용이므로 Google에 유리.
- 4. **원저작물의 잠재적 시장 또는 가치에 미치는 영향** : Google이 만든 안드로이드는 새로운 시장을 창출한 것으로 보았고, Oracle의 Java 시장과 직접 경쟁하지 않는다고 판단.

## — 공정이용의 제4요소 : 원저작물의 잠재적 시장 또는 가치에 미치는 영향

- 미국 저작권청은 저작물의 AI학습이 '공정이용'에 해당하지 않을 수 있다는 취지의 Copyright and Artificial Intelligence, Part 3 : Generative AI Training의 Pre-publication을 지난 5월에 발표했습니다.

### 1. Lost Sales (판매 손실 가능성)

...  
There are instances, however, where the use of works in generative AI training can lead to a loss in sales. The use of pirated collections of copyrighted works to build a training library, or the distribution of such a library to the public, would harm the market for access to those works. And where training enables a model to output verbatim or substantially similar copies of the works trained on, and those copies are readily accessible by end users, **they can substitute for sales of those works.**

저작물의 불법 복제본을 대중에게 배포하면, 시장에 피해를 줄 수 있고, AI 모델이 실질적으로 유사한 복제본을 생성하게 하면 저작물의 판매를 대체할 수 있음.

**A potential loss of sales is particularly clear in the case of works specifically developed for AI training.** There is a thriving industry focused on developing training datasets that improve the ability of language models to follow instructions, format and structure outputs, use tools, act consistently with human values, or improve domain performance. Where the content of those datasets is copyrightable, or the datasets themselves evince human selection and arrangement of data, and the datasets are primarily or solely targeted at AI training, widespread unlicensed use would likely cause market harm.

특히 AI 학습을 위해 특별히 준비된 데이터인 경우, 잠재적인 판매 손실은 더욱 분명함.

Uses involving the retrieval of copyrighted works by **RAG can also result in market substitution.** As described above, RAG augments AI model responses by retrieving relevant content during the generation process, resulting in outputs that may be more likely to contain protectable expression, including derivative summaries and abridgments. A user for whom the augmented response "satisf[ies] the . . . need" for the original work will not pay to obtain it in the marketplace.

RAG의 경우에도 시장 대체 효과를 유발할 수 있으며, 사용자는 결국 원본을 구매하지 않을 가능성이 커짐.

### 3. Lost Licensing Opportunities (라이선스 기회의 상실)

...  
**Although licensing markets are still developing and factual contexts vary, available information shows that markets exist or are "reasonable" or "likely to be developed,"** for certain copyright sectors, types of training or uses, and models. Direct licensing is most common and most promising with respect to corporate entities with catalogs of high-quality and easily identifiable content. For example, content controlled by large stock photography companies, national news outlets, and major record companies or film studios may be more easily licensable. Such content likely has a higher training value because it is high-quality and curated, and the centralization of rights makes it easier to license without incurring substantial volume-related transaction costs.

라이선스 시장은 여전히 형성 중이지만, 일부 분야에서는 이미 시장이 존재하거나 합리적 혹은 개발될 가능성이 있다고 판단할 수 있음.

...  
As both the creative industries and AI technologies develop further, data needs and licensing markets will continue to evolve. **Where licensing markets are available to meet AI training needs, unlicensed uses will be disfavored under the fourth factor.** But if barriers to licensing prove insurmountable for parties' uses of some types of works, there will be no functioning market to harm and the fourth factor may favor fair use.

향후 라이선스 시장은 진화할 것이고, 이러한 AI 학습 라이선스 시장이 존재한다면, 허가 받지 않은 이용은 공정이용 제4요소 판단에서 불리하게 작용할 것.

## — 텍스트 및 데이터 마이닝(TDM) 면책?

- EU TDM(DSM) 면책 규정은 직접적인 AI 학습 데이터 입수, 학습, 배포 과정의 일부 영역만을 면책합니다.

### DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

#### Article 3

##### Text and data mining for the purposes of scientific research

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by **research organisations and cultural heritage institutions** in order to carry out, for the purposes of **scientific research**, text and data mining of works or other subject matter to which they have lawful access.

...

① Article 3에서는 과학적 연구인 경우에는 DSM 규정이 제한 없이 적용 됨을 명시함.

→ 독일 저작권법(Urheberrechtsgesetz, UrhG) 제60조d항

(1) In order to enable **scientific research**, it shall be permissible to reproduce works for text and data mining purposes (section 44b (1) and (2), first sentence) if: 1. the access to the works is lawful and 2. the reproductions are made by **research organisations or cultural heritage institutions**....

#### Article 4

##### Exception or limitation for text and data mining

1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

...

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph **has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.**

...

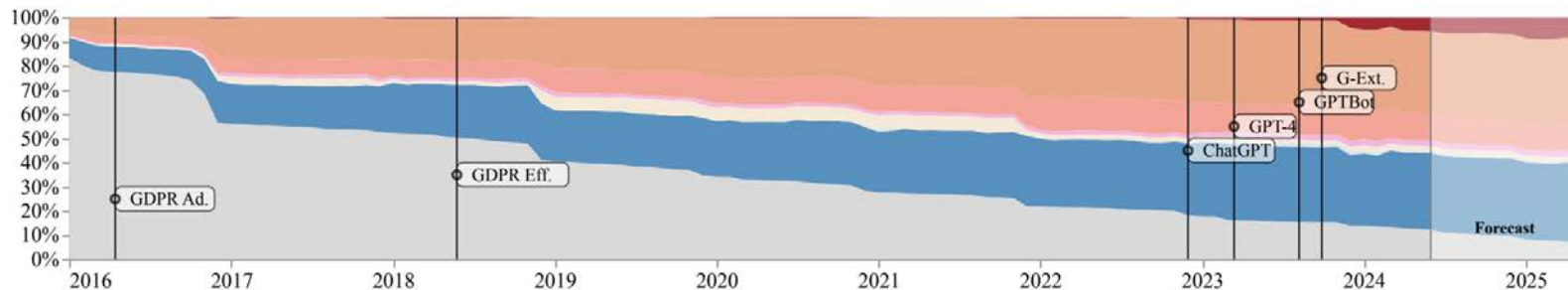
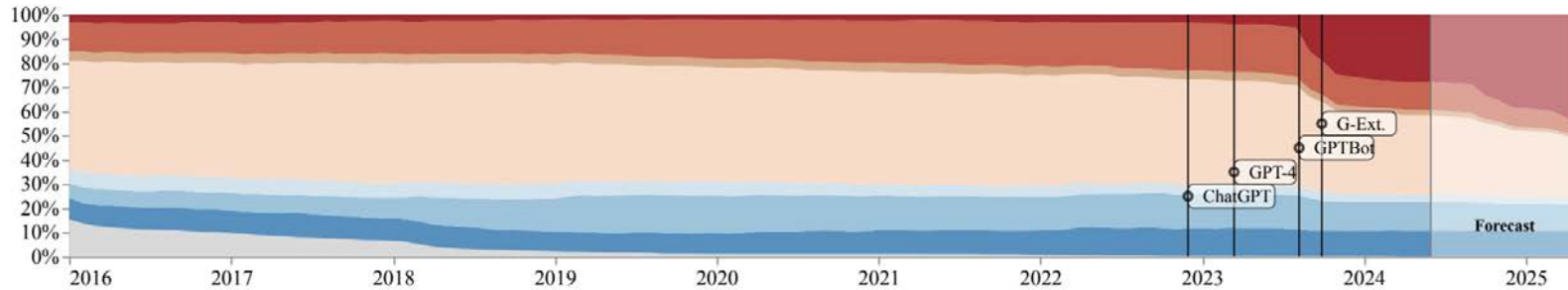
② Article 4에서는 상업적 목적까지 포함되는 경우(과학적 연구가 아닌 경우), 저작권자가 기계 판독 가능한 수단 등으로 의사표시하여 Opt-Out 할 수 있는 것으로 명시함.

→ 독일 저작권법(Urheberrechtsgesetz, UrhG) 제44조b항

...(3) Rightholders shall be entitled to **prohibit the use of works** for text and data mining purposes in accordance with subsection (2) **if they have expressly reserved this right in an appropriate manner.**

## — 텍스트 및 데이터 마이닝(TDM) 면책?

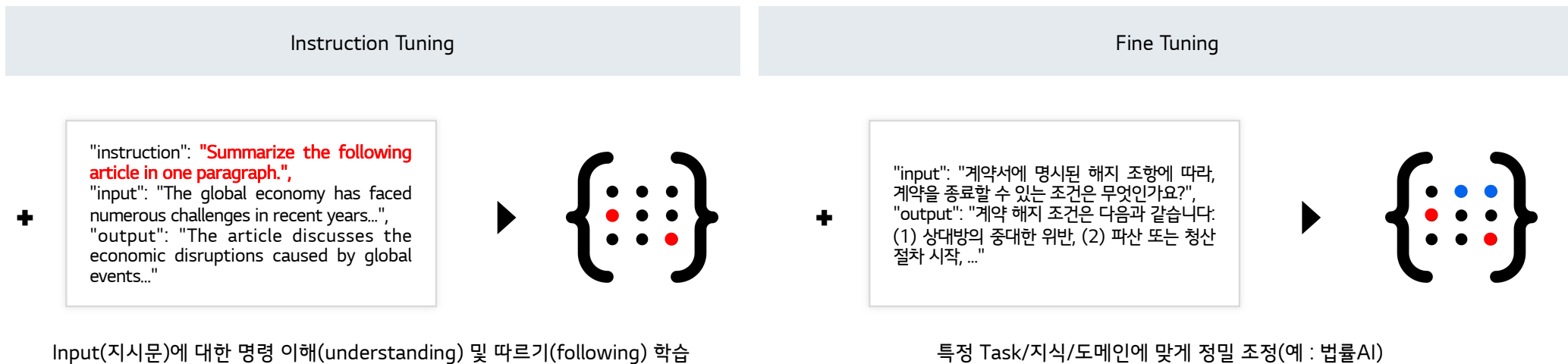
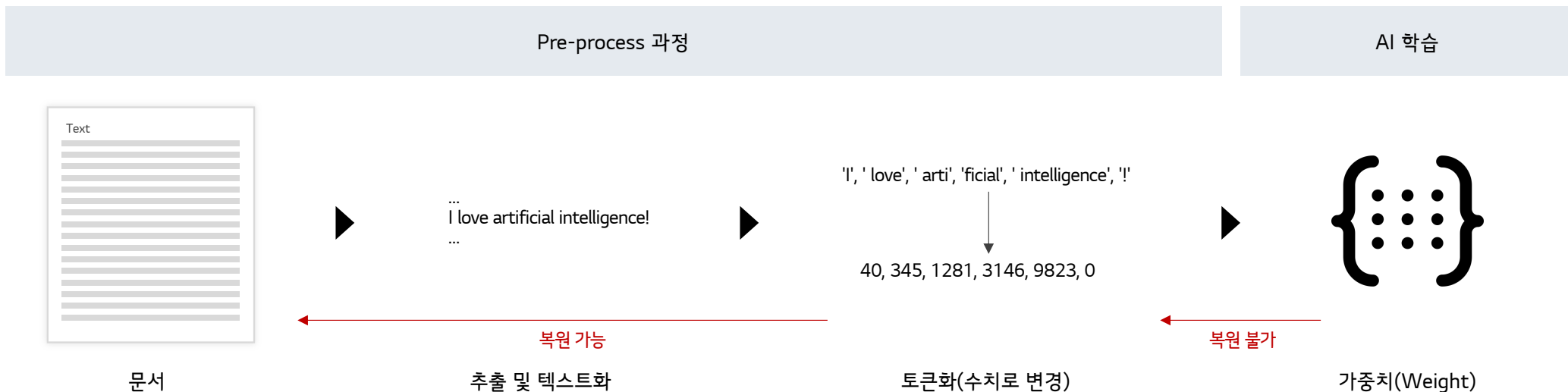
- 데이터에 대한 권리 보유자는 특히 주요 생성형AI가 공개된 이후 그들의 저작물을 지키기 위하여 Robots.txt, 이용약관 수정의 조치를 취하고 있습니다.



- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., Sileo, D., ... & Pentland, S. (2024). Consent in crisis: The rapid decline of the AI data commons

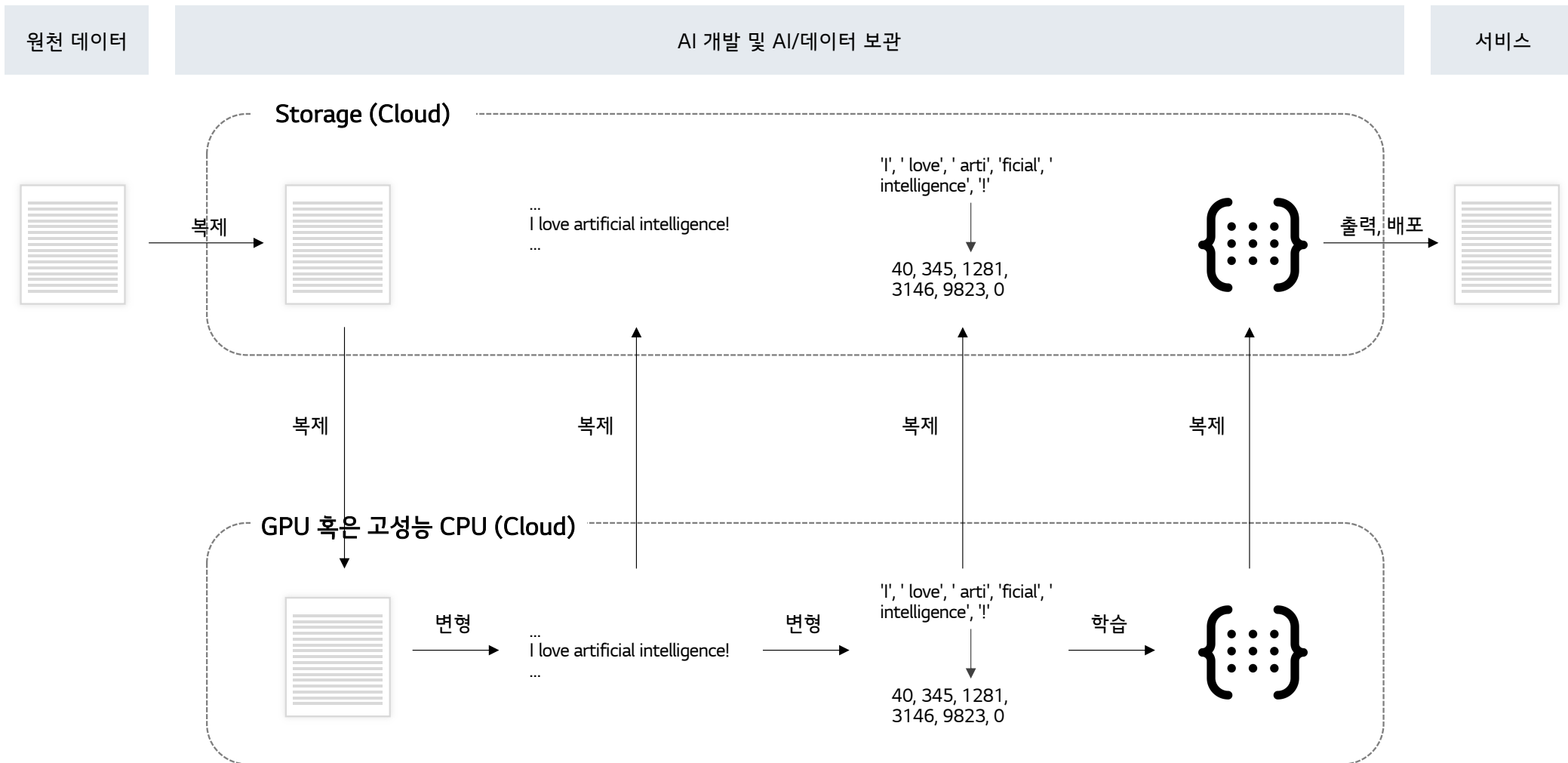
## AI 학습 과정에서의 Data의 변화 : Text에서 Weight으로 변화

- AI 학습 과정에서 저작물은 그 형태가 문서→Text →단위(토큰)→숫자→행렬로 변화하고, AI모델의 출력은 반대 과정을 거칩니다.



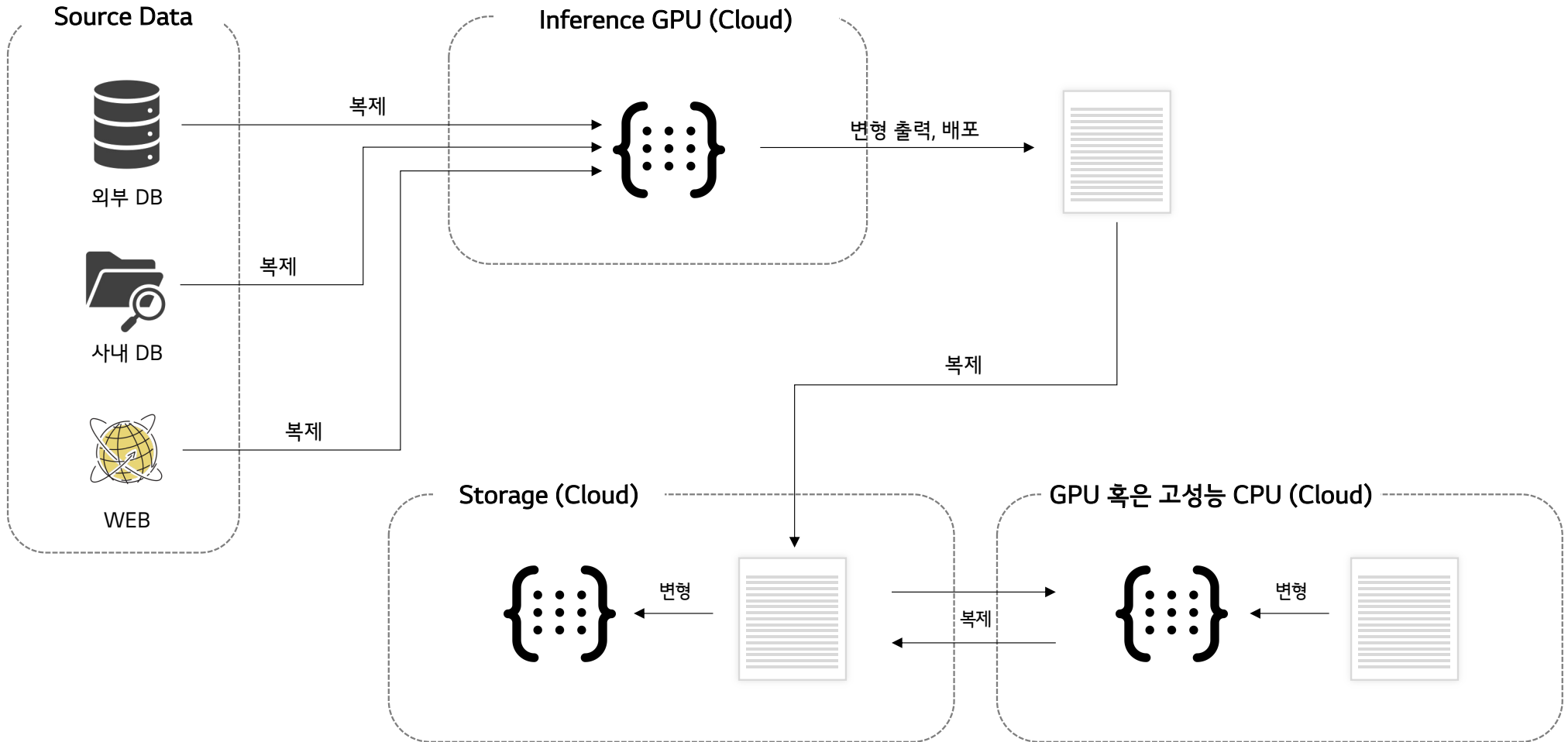
## — 학습과정의 리스크

- AI학습 과정에서 원천 데이터는 복제, 변형(혹은 2차적저작물 생성), 배포 됩니다.



## 서비스과정의 리스크

- AI서비스 과정에서도 소스 데이터가 복제, 변형(혹은 2차적저작물 생성), 배포 됩니다. 또한, 배포된 데이터가 다시 재학습 되어 또다른 AI모델에 학습됩니다.



피고	원고	소제기일	배경	이슈가 되는 데이터
OpenAI	Doe 1	2022.3.11	개발자들의 라이선스를 무시하고 Github을 통해 코드를 무단으로 학습하여 CODEX와 Copilot을 개발했다고 주장하며 소제기	GitHub의 Licensed Code
	Mark Walters	2023.6.5	사실이 아닌 내용을 생성하여 명예훼손을 했다고 주장하며 소제기	Privacy (Defamation)
	Paul Tremblay	2023.6.28	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었으며, GPT-4가 그들의 책에 대한 상세 요약을 제공했다고 주장하며 소제기	The Pile3 : Books2, Shadow Liabrary (Library Genesis, Z-Library, Sci-Hub, Bibliotik)
	Sarah Silverman	2023.7.7	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었으며, GPT-4가 그들의 책에 대한 상세 요약을 제공했다고 주장하며 소제기	Shadow Library (Z-Library, Bibliotik, Library Genesis)
	A.T.	2023.9.5	개인정보가 포함된 학습데이터를 사용했다고 주장하며 소제기	Privacy
	Chabon	2023.9.8	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었으며, GPT-4가 그들의 책에 대한 상세 요약을 제공했다고 주장하며 소제기	Shadow Library (Z-Library, Bibliotik, Library Genesis)
	Authors Guild	2023.9.19	작가들이 자신들의 책과 글들이 허락 없이 학습데이터로 활용되고, ChatGPT의 파생물로 현현했다고 주장하며 소제기	Common Crawl, Books1/Books2 Dataset
	Alter	2023.11.21	작가들이 자신들의 책과 글들이 허락 없이 학습데이터로 활용되고, ChatGPT의 파생물로 현현했다고 주장하며 소제기	Authors Guild v OpenAI와 병합
	The New York Times Co.	2023.12.27	NYT가 자신들의 저작물이 GPT에 학습되고, ChatGPT가 NYT의 저작물을 제공한다고 주장하며 소제기	News Data
	Basbanes	2024.1.5	두 작가가 그들의 책이 허락 없이 학습데이터로 활용 되었다고 주장하며 소제기	Authors Guild v OpenAI와 병합
	A.S.	2024.2.27	개인정보가 포함된 학습데이터를 사용했다고 주장하며 소제기	Privacy
	Raw Story Media, Inc.	2024.2.28	뉴스 사이트인 Raw Story와 Alter Net가 Copyright Management Information 삭제에 대한 DMCA 위반으로 소제기	News Data
	Intercept Media Inc.	2024.2.28	뉴스 사이트인 The Intercept가 Copyright Management Information 삭제에 대한 DMCA 위반으로 소제기	News Data
	Elon Musk	2024.2.29	OpenAI 설립 시의 합의 내용과 다르게 운영되는 내용에 대한 합의 위반을 주장하며 소제기	Founding Agreement (합의)
New York Daily News	2024.4.30	8개의 신문사가 자신들의 저작물이 GPT에 학습되고, ChatGPT가 신문사들의 저작물을 제공한다고 주장하며 소제기	News Data	
Meta	Richard Kadrey	2023.7.7	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었다고 주장하며 소제기	The Pile : Books3, Shadow Library (Bibliotik)
	Chabon	2023.9.12	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었다고 주장하며 소제기	Richard Kadrey v. Meta Platforms, Inc.와 병합
Alphabet	J.L.	2023.7.11	인터넷 이용자들을 대표하여 구글이 제3자의 재산이며 개인정보가 포함된 인터넷의 글을 허락 없이 스크래핑 했다고 주장하며 소제기	C-4 (Common Crwal의 파생)
Stability AI	Sarah Andersen	2023.1.13	Visual Art작가들이 그들의 허락없이 저작물이 복제되고 Output으로 현현되었다고 주장하며 소제기	Laion Dataset
	Getty Images	2023.2.3	저작권자의 허락을 받지 않고 사진을 무단으로 이용하여 학습데이터로 사용했다고 소제기	Laion Dataset
Anthropic	Concord Music Group, Inc.	2023.10.18	Universal Music Group의 음반 저작권자들이 Claude가 음반(가사)에 대한 저작권을 침해하여 AI Model을 개발하고 운영했다고 주장하며 소제기	음악의 가사
Databricks, Mosaic ML	Stewart O'Nan	2024.3.8	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었다고 주장하며 소제기	The Pile : Books3
NVIDIA	Abdi Nazemian	2024.3.8	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었다고 주장하며 소제기	The Pile : Books3
Microsoft, The Eleuther AI Institute, Meta, Bloomberg	Mike Huckabee	2023.10.17	작가들이 자신의 책이 허락 없이 학습데이터로 활용되었다고 주장하며 소제기	The Pile : Books3
Ross Intelligence	Thomson Reuters Enterprise Centre GMBH	2020.5.6	Westlaw의 Research Database 권리자가 AI 스타트업이 Westlaw의 headnote를 불법적으로 복제하여 AI학습데이터로 사용했다고 주장하며 소제기	Article(Legal Research)
Bloomberg	Huckabee	2023.12.27	피고가 작가들의 책이 저작권법으로 보호받는 저작물인 것을 인지도 있음에도 불구하고 AI학습데이터로 이용했다고 주장하며 소제기	The Pile : Books3
Suno, Inc.	Universal Music Group	2024.6.24	Suno가 저작권이 있는 음악을 무단으로 AI 학습에 사용하고 AI 음악 생성 서비스를 제공했다고 주장하며 소제기	음악

## — Li v. Liu (베이징인터넷법원판결 : (2023)경0497민초11279호)

- 원고인 Li가 Stable Diffusion을 이용해 생성한 이미지를 SNS에 올린 후, 피고 Liu가 이를 무단으로 사용한 것에 대해 '저작권 침해 및 정보네트워크 보급권 분쟁 혐의'로 소송을 제기함.

Stable Diffusion으로 생성한 원고의 이미지



### 법원의 판결

- AI는 도구일 뿐이며, 사람의 지적 투자가 창작과정에 반영되었으므로 생성된 이미지의 저작물성이 인정됨.
- 원고가 프롬프트 입력과 파라미터 설정 등 창작 과정에 관여했으므로, 원고가 생성된 이미지의 저작자로 인정됨.
- 피고가 원고의 저작물을 무단 사용하고 워터마크를 제거했으므로 저작권 침해가 인정됨.
- 피고는 원고에게 공개 사과문을 게시해야 하며, 500위안(한화 약 10만원)을 지급해야 함)

## — 신창화 문화발전 유한공사 v. AI회사 (광저우인터넷법원판결 : (2024)월0192민초1호)

- 중국 울트라맨 독점 저작권 사용 허가를 받은 원고인 상하이 신창화 문화발전 유한공사가 피고 A사의 생성형 AI 서비스가 제공되는 사이트에서 울트라맨 이미지가 생성되는 것을 보고 저작권이 침해되었다며 소송을 제기함.

피고 A사의 서비스로 생성된 이미지

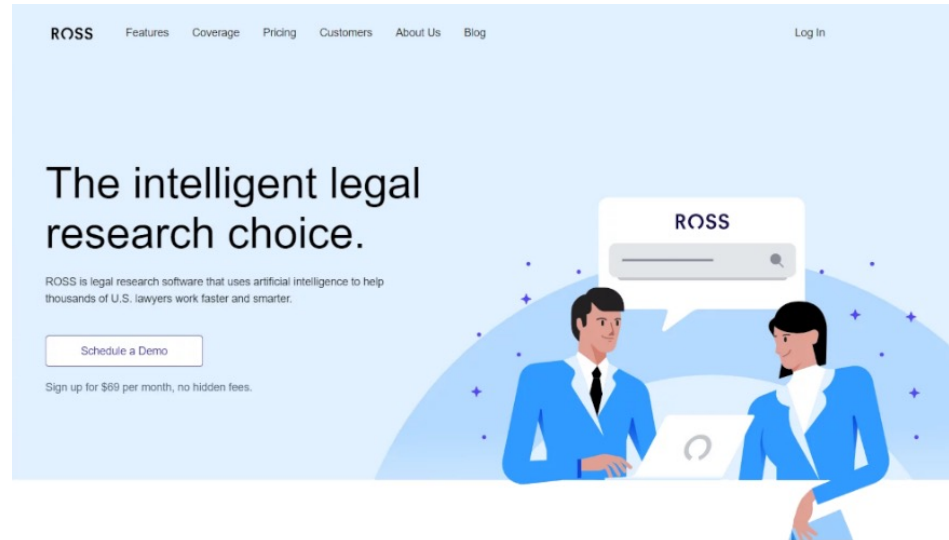


### 법원의 판결

- 피고의 웹사이트에서 '울트라맨' 키워드 입력만으로 저작권으로 보호되는 울트라맨 이미지와 실질적으로 유사한 이미지가 출력됨.
- 생성형AI 서비스 제공자인 피고는 저작권법 상 원고의 복제권 및 각색권을 침해함.
- 피고는 '생성형 인공지능 잠정관리 방법'(중국의 법률)에 따른 '민원 제기 및 신고 체제 구축'을 하지 않았고, '사용자 규정, 약관 등으로 사용자가 타인의 지식재산권을 침해하지 않도록 고시'를 하지 않았으며, 'AI 산출물에 대한 표시 의무'를 다하지 않았음.
- 피고는 원고에게 10,000위안(약 200만원)을 지급해야 하며, 울트라맨 관련 키워드로 실질적으로 유사한 이미지가 생성되지 않도록 예방 조치를 취해야 함.

## Thomson Reuters v. ROSS Intelligence (미국에서의 AI 저작권 소송 첫 판결)

- 2020년 원고인 Thomson Reuters는 ROSS사가 자사의 서비스인 Westlaw의 자료를 무단 복제하여 학습했다고 주장하며 소를 제기하였고, 법원에서 2025년 2/11에 Summary Judgement 판결을 내리면서 원고인 Thomson Reuters의 승소 판결을 내림.

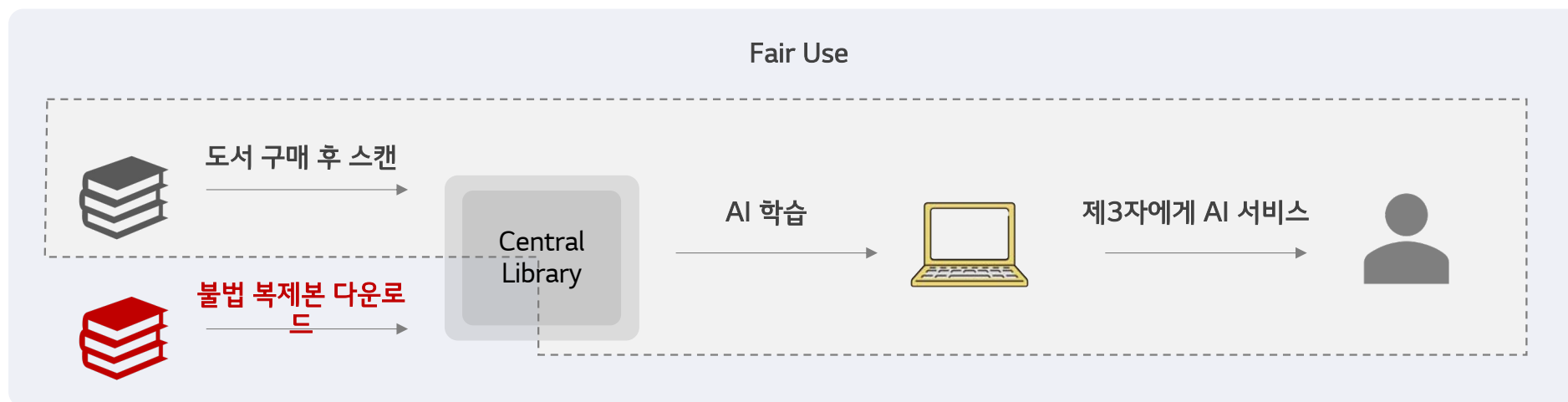


### 법원의 판결

- 공정이용의 네가지 요소인 1) 사용 목적, 2) 원본 저작물의 성격, 3) 사용된 저작물의 양, 4) 해당 사용이 원본 저작물의 시장 가치에 미치는 영향 중 특히 4)에 대해서 시장에서 **Thomson Reuters에 손해를 미쳤다고 판단하며 ROSS의 공정이용 주장을 기각함.**
- ROSS가 Westlaw의 대체제를 개발함으로써 경쟁하려 했으며, 이 행위가 AI 학습 데이터 시장에 영향을 미칠 수 있다고 판단함. **ROSS가 독자적으로 개발할 수도 있었던 것을 Thomson Reuters의 저작권을 침해하며 사용한 것은 정당화될 수 없다고 판결했습니다.**
- 판결이 나온 지 몇 시간 만에, 음악 저작권자(출판사)가 Anthropic을 상대로 제기한 소송의 원고 측은 법원에 이 판결을 제출할 수 있도록 허가를 요청함.

## — 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.

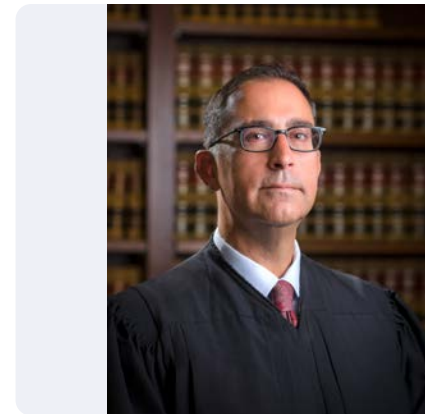


공정이용 4요소	1) LLM 학습을 위한 복제	2) 구매한 책의 복제(디지털화)	3) 불법 다운로드한 도서 복제(저장)
제1요소 : 원저작물 이용의 목적과 성격	피고 유리	피고 유리	원고 유리
제2요소 : 원저작물의 성격	원고 유리	원고 유리	원고 유리
제3요소 : 이용된 부분의 양과 중요성	피고 유리	피고 유리	원고 유리
제4요소 : 원저작물의 잠재적 시장 또는 가치에 미치는 영향	피고 유리	중립	원고 유리
공정이용?	공정이용 O	공정이용 O	공정이용 X

## — 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 0



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..”

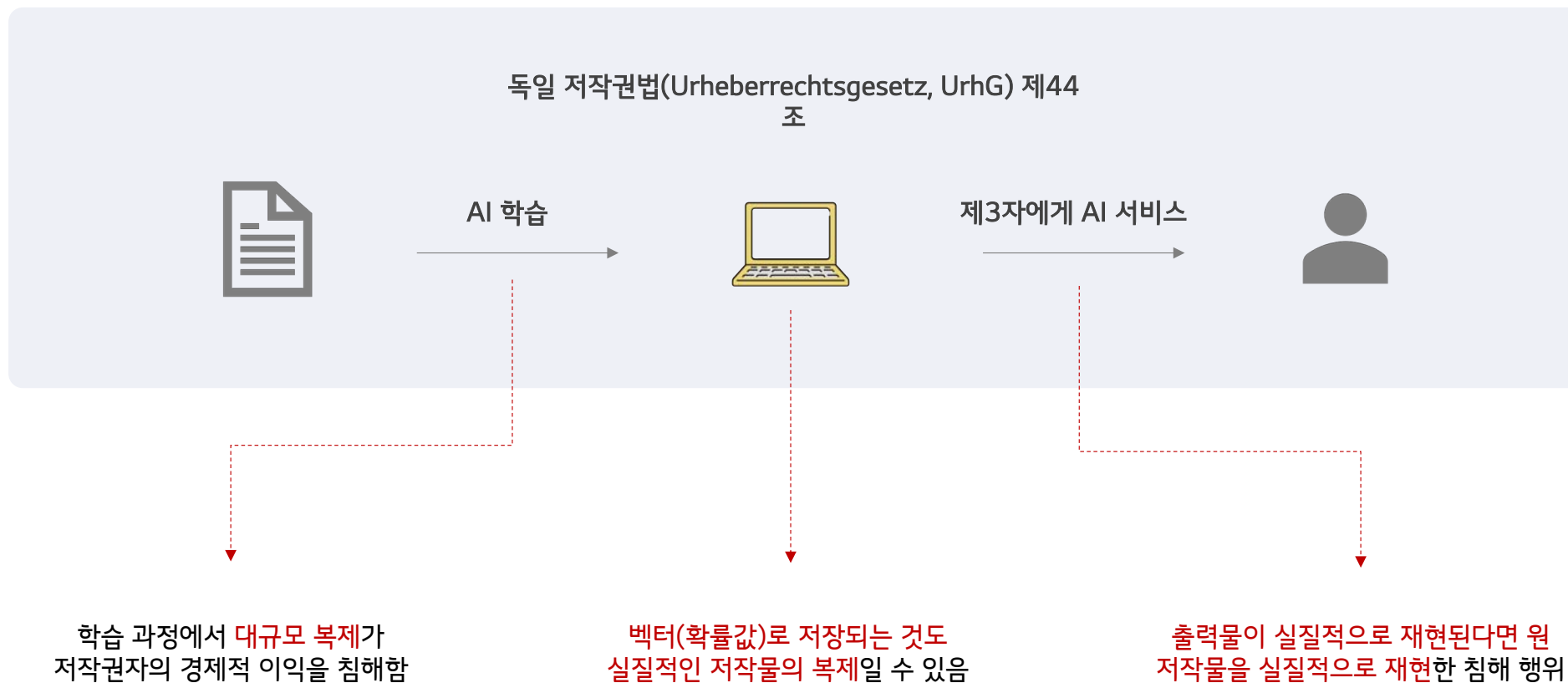
“.. 법원은 일반적 이해에 근거해 판결할 수 없고, 당사자들이 제출한 증거에 근거하여 사건을 판단해야 한다.. 원고의 주장은 명백히 패소할 주장이었다..”

“.. 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..”

## — 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 초과한다고 판결함.



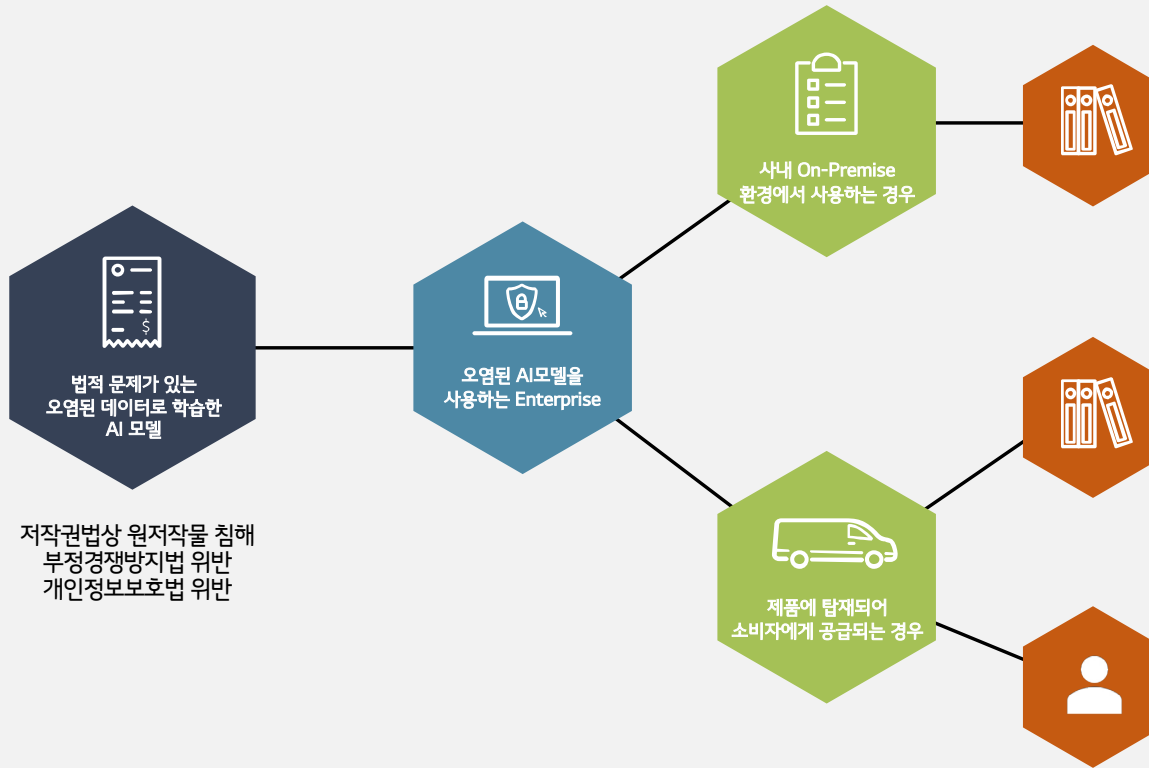
Chapter

2

Data Compliance  
Framework

## — LG에게 발생할 수 있는 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.



### 데이터 저작자와의 분쟁

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

### 데이터 저작자와의 분쟁

- 저작권법상 저작권 침해 방조행위에 대한 법적 책임<sup>5</sup>
- 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 피해를 끼친 것, 결함에 대해 고지의 의무를 다하지 않은 것에 대한 불법행위<sup>6</sup>
- 민법상 법률적 제한 내지 장애로 인해 하자에 해당하여 손해배상 책임 발생<sup>7</sup>
- 형법상 해당 하자가 발생할 것을 알고도 판매한 것에 대한 사기죄<sup>8</sup>

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

## — Data의 투명성과 관련된 법적, 윤리적 움직임

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

### 워싱턴주 입법중

HB 1168 - AI 학습 데이터 투명성 법안

- 생성형 AI 시스템 또는 서비스를 개발하는 기업은 해당 시스템의 훈련에 사용된 데이터에 대한 문서를 공개해야 합니다.
- 공개 문서에는 데이터 출처, 데이터 유형 및 수량, 개인정보 사용 여부, 데이터 수정 과정, 합성 데이터 사용 여부 등이 포함되어야 합니다.

### 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 상업성 AI 기업이 무단으로 저작물을 활용하여 수익을 창출하면서도 저작권자에게 대가를 지불하지 않는다면, 이는 시장 왜곡을 초래하고, 공정하지 않다고 평가됨.

### 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.

### 버지니아주 입법중

HB 2250 - AI 학습 데이터 투명성 법안

- 소비자는 자신의 개인정보가 AI 모델 훈련에 사용되는 것을 거부할 수 있는 권리를 가집니다.
- 개발자는 소비자의 요청에 따라 훈련 데이터에 대한 확인 및 삭제 요청을 처리해야 하며, '훈련 금지(Do Not Train)'로 지정된 데이터를 AI 훈련에 사용해서는 안 됩니다.

### EU

EU AI ACT 실천강령 공개 예정

- 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

## — 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.

## 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)

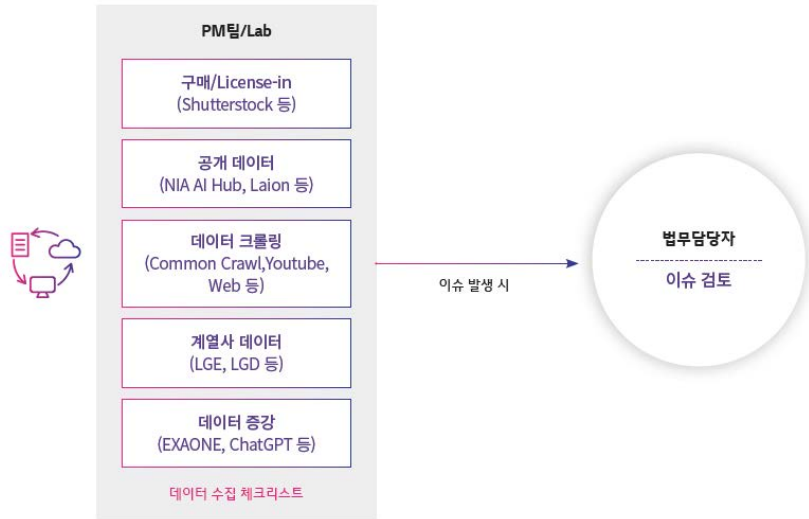
### DATA COMPLIANCE REPORT

- Data Compliance Score 및 Class 평가 기준(ENG)

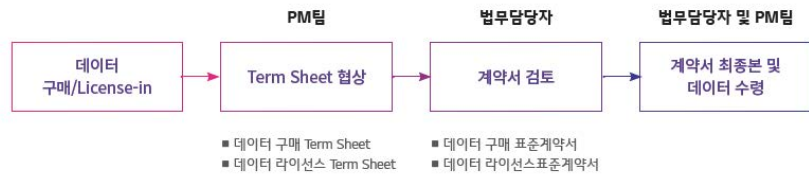
## Data Compliance

- Data의 수집, 보관, 파기, 이전, 구매, 이용허락(License-in) 과정에서 발생할 수 있는 법적 리스크를 탐지합니다.
- 법적 리스크가 존재하는 경우, 법무담당자가 개입하여 올바른 IP Transaction이 발생하도록 조치합니다.

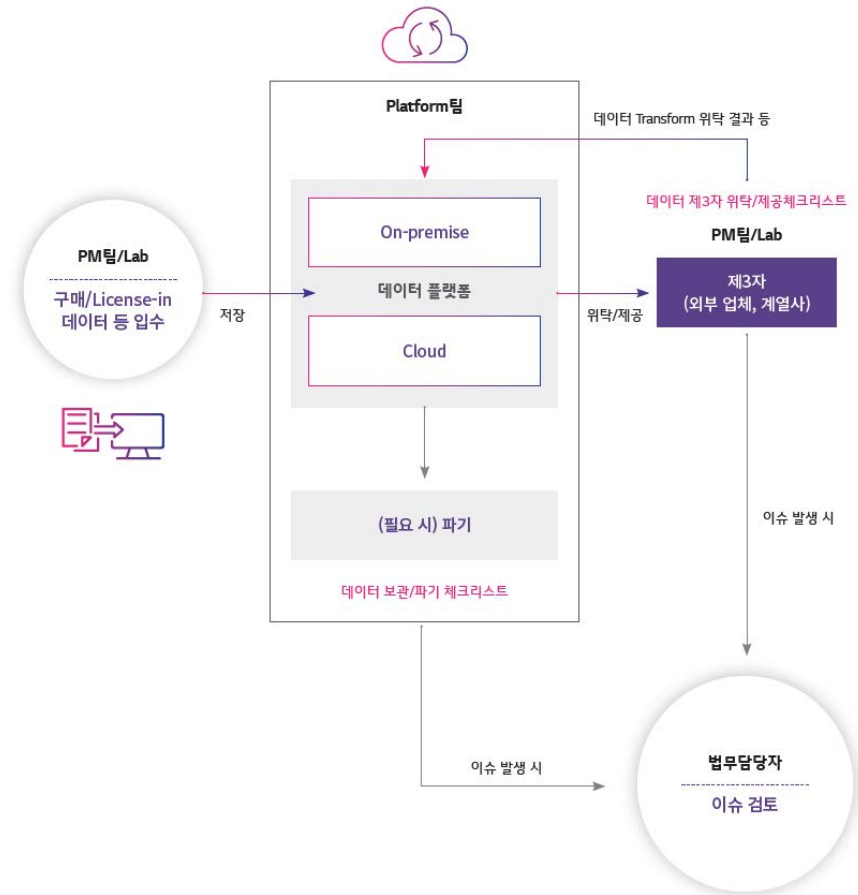
### Data 수집 프로세스



### Data 구매/License-in 프로세스



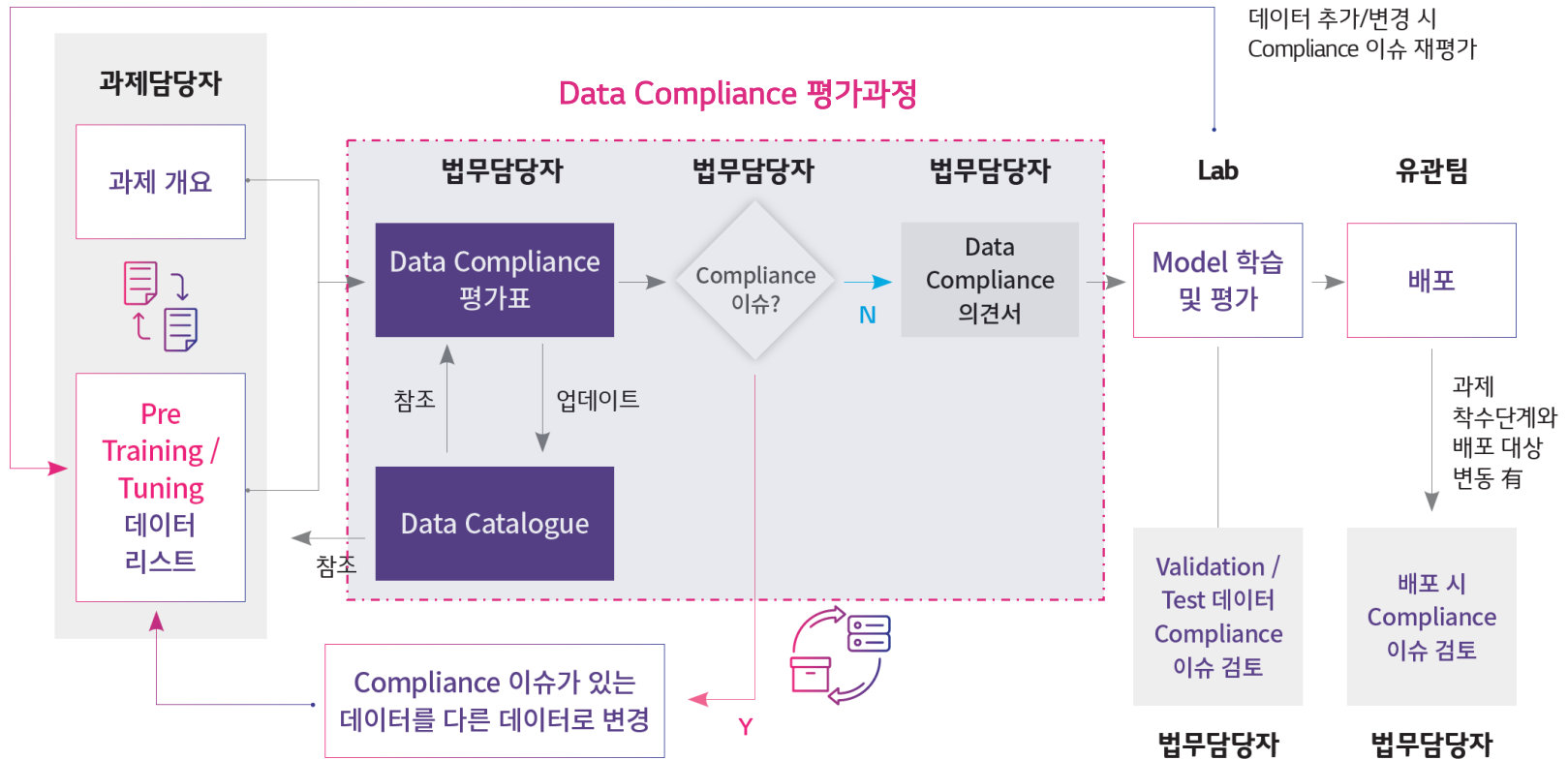
### Data 보관/파기/이전 프로세스



## Data Compliance

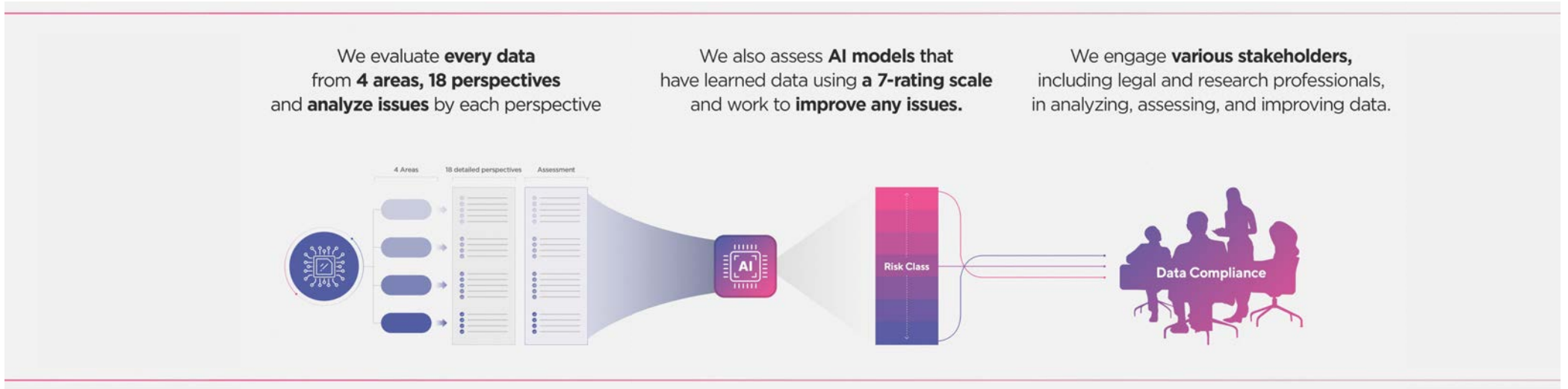
- Data Compliance 프로세스를 통해, 모든 학습 데이터를 검토하고 평가합니다.

### Data Compliance 프로세스



## Data Compliance

- 학습 데이터에 법적 리스크가 존재하는 경우, 법무담당자와 연구원이 함께 대안을 고민하고 해결책을 마련합니다.



### Permission to Use Data

- Granting of License Status
- Permission to Modify Data and Authority to Create Derivative Works
- Potential Copyright Infringement of Generated Outputs
- Status of Rights Assignment for Outputs
- Existence of Data Attribution Obligations

### Data Usage Period and Territory

- Limitations on Data Usage Period
- Possibility of License Revocation
- Limitations on AI Model Service Period
- Geographic Restrictions on Data Usage

### Personal Information and Data Security

- Inclusion of Personal Information or Pseudonymized Data
- Consent Status of Data Subjects
- Possibility of Third-Party Data Transfer/Provision
- Restrictions on Specific User Access to Data

### Additional Legal Risks

- Issues in the Data Collection Process
- Known Disputes Regarding AI Models Using the Data
- Additional Risks Present in License Agreements
- Types of License Conditions

## 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		리스크				
		小		大		
Data에 대한 License 부여 여부	1-1) License 부여 여부	상업적 목적으로 사용하지 사용 가		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 저작물 생성에 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 등의 받거나, 산출물이 생성되지 않음	원저작물과 다른 형태의 산출물 생성됨	원저작자의 저작물과 유사한 산출물 생성 가능성 있으나, 높지 않음	원저작자의 저작물과 유사한 산출물 생성 가능성과 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		리스크			
		小	중	대	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 위반하거나 적법하지 않은 방법으로 획득
Data 분쟁 여부	4-2) Data가 이용된 시모델에 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 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

- 항목별로 평가된 점수를 바탕으로, 각 항목별 가중치를 통해 해당 Data의 Risk를 정의합니다.
- Class, 총점으로 정의되며, Weight의 결정 기준은 각 국가의 입법 현황, 사법적인 판단에 기인합니다.

$R_{1-1}$  = Mandatory Data Class

$$R_i = 0.1R_{1-2} + 0.15R_{1-3} + 0.05R_{1-4} + 0.03R_{1-5} + 0.07R_{2-1} + 0.03R_{2-2} + 0.05R_{2-3} + 0.04R_{2-4} + 0.09R_{3-1} + 0.03R_{3-2} + 0.03R_{3-3} + 0.05R_{3-4} + 0.02R_{3-5} + 0.06R_{4-1} + 0.05R_{4-2} + 0.1R_{4-3} + 0.05R_{4-4}$$

$R_i$  = Risk score of data #i

### Category A. Permission to Use Data

$R_{1-2}$  = Permission to Modify Data and Authority to Create Derivative Works

$R_{1-3}$  = Potential Copyright Infringement of Generated Outputs

$R_{1-4}$  = Status of Rights Assignment for Outputs

$R_{1-5}$  = Existence of Data Attribution Obligations

### Category B. Data Usage Period and Territory

$R_{2-1}$  = Limitations on Data Usage Period

$R_{2-2}$  = Possibility of License Revocation

$R_{2-3}$  = Limitations on AI Model Service Period

$R_{2-4}$  = Geographic Restrictions on Data Usage

### Category C. Personal Information and Data Security

$R_{3-1}$  = Inclusion of Personal Information

$R_{3-2}$  = Consent Status of Data Subjects

$R_{3-3}$  = Inclusion of Pseudonymized Data

$R_{3-4}$  = Possibility of Third-Party Data Transfer/Provision

$R_{3-5}$  = Restrictions on Specific User Access to Data

### Category D. Additional Legal Risks

$R_{4-1}$  = Issues in the Data Collection Process

$R_{4-2}$  = Known Disputes Regarding AI Models Using the Data

$R_{4-3}$  = Additional Risks Present in License Agreements

$R_{4-4}$  = Types of License Conditions

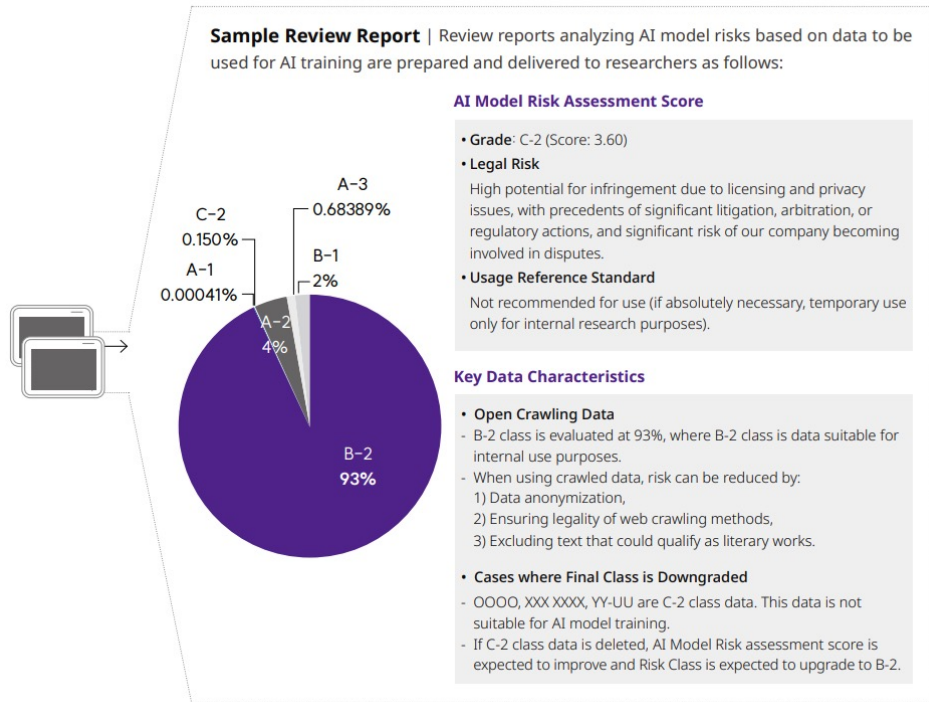
## Data Compliance

- 생성형 AI의 경우 각 Data의 Token의 가중치를 바탕으로 AI모델의 Risk Category와 Score가 계산 됩니다.
- 용도 참고별로, 서비스 전부터 각 AI모델의 Risk별 적합한 용도(제3자 제공)의 범위를 인식합니다.

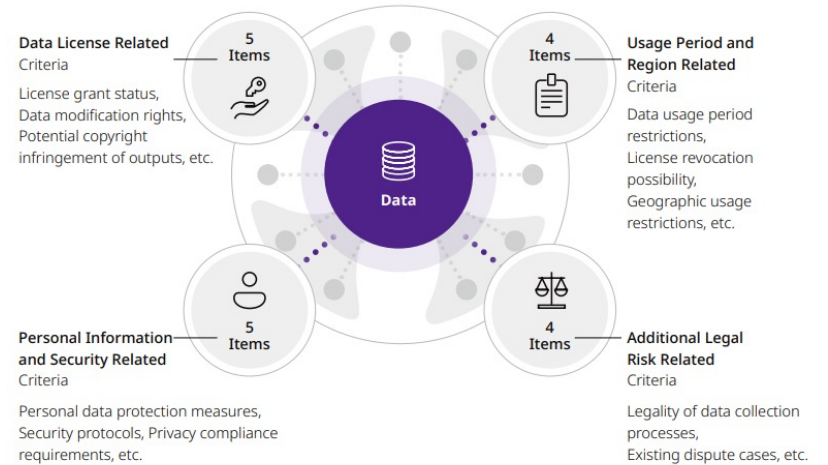
Category	Score	분류	LEGAL RISK	용도 참고
A-1	4.90	License/개인정보 -Free	자체서비스, Public Cloud에서 AI Model 서비스 등 Public하게 공개되어도 원저작자나 Licensor, 정보주체, 유관기관으로부터 Data와 관련한 분쟁의 가능성이 없음	B2C
A-2	4.57	License/개인정보 위험도 小	License/개인정보 이슈가 존재하여 침해의 가능성이 약간 존재 가능하나 소송이나 중재 등 분쟁 또는 유관기관의 처분 등으로 심화된 알려진 Case 없음	
A-3	4.22	License/개인정보 위험도 小	License/개인정보 이슈가 존재하여 침해의 가능성이 상당히 존재 가능하나 소송이나 중재 등 분쟁 또는 유관기관의 처분 등으로 심화된 알려진 Case 없음	
B-1	3.73	License/개인정보 위험도 中	License/개인정보 이슈가 존재하여 침해의 가능성이 높으며, 소송이나 중재 등 분쟁 또는 유관기관의 처분 등이 발생한 Case가 존재하고, 당사도 분쟁에 휘말릴 가능성이 약간 존재함	B2B
B-2	3.51	License/개인정보 위험도 中	License/개인정보 이슈가 존재하여 침해의 가능성이 높으며, 소송이나 중재 등 분쟁 또는 유관기관의 처분 등이 발생한 Case가 수 건 있고, 당사도 분쟁에 휘말릴 가능성이 약간 존재함	
C-1	3.18	License/개인정보 위험도 大	License/개인정보 이슈가 존재하여 침해의 가능성이 크게 높으며, 소송이나 중재 등 분쟁 또는 유관기관의 처분 등이 발생한 Case가 상당히 있고, 당사도 분쟁에 휘말릴 가능성이 존재함	Internal
C-2	-	License/개인정보 위험도 大	License/개인정보 이슈가 존재하여 침해의 가능성이 크게 높으며, 거액의 소송이나 중재 등 분쟁 또는 유관기관의 처분 등이 발생한 Case가 있고, 당사도 분쟁에 휘말릴 가능성이 크게 존재함	

## 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.



**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.

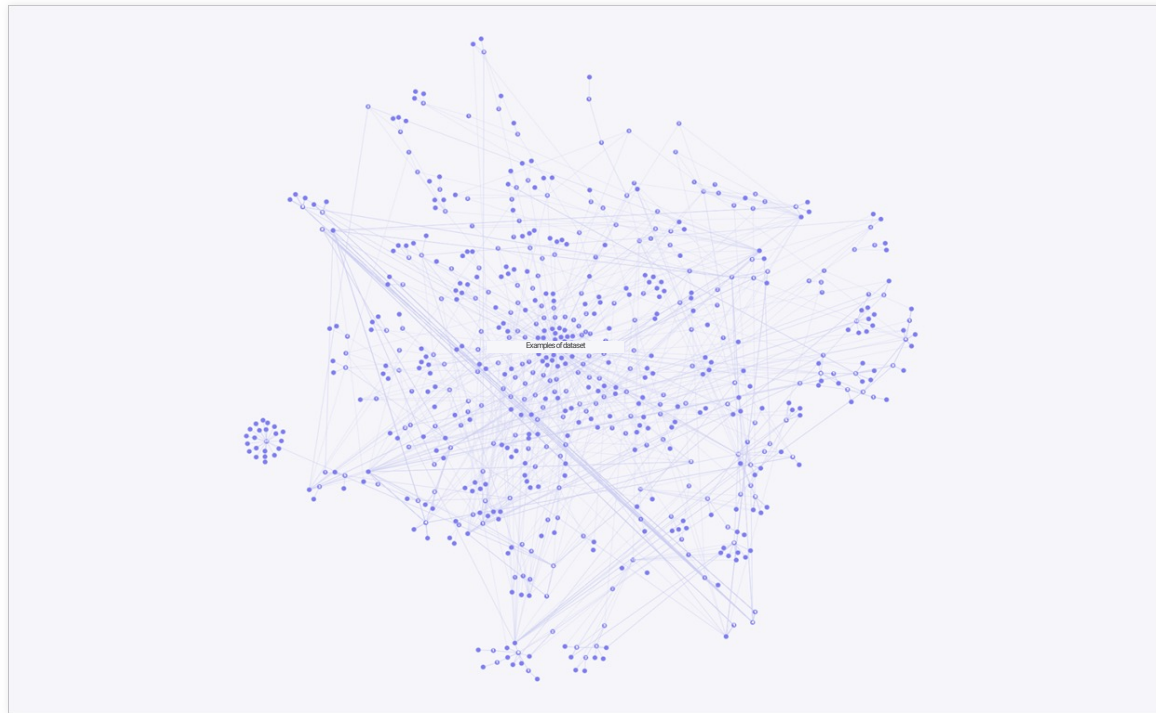
Chapter

3

EXAONE NEXUS:  
Open Data  
Compliance Agent

## — 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 소스에서 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) 통계

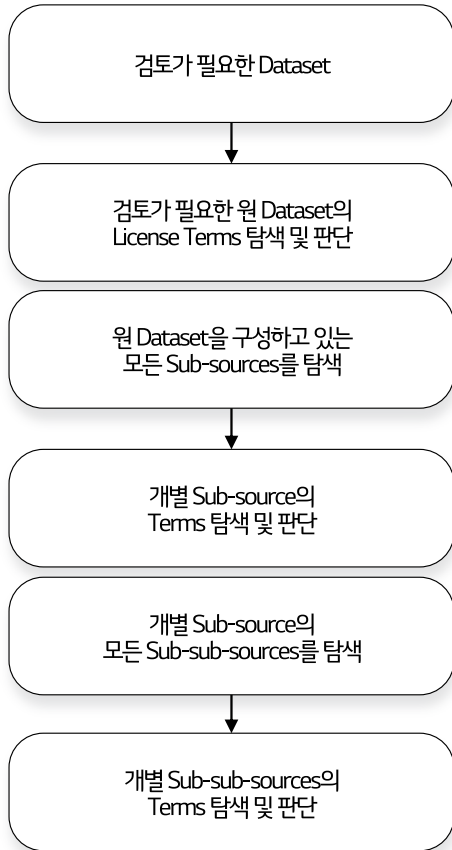
Mean	Std	Min	25%	50%	75%	Max
23.18	87.15	1	2	5	11	1691

3,612 데이터셋의 Data Source 통계

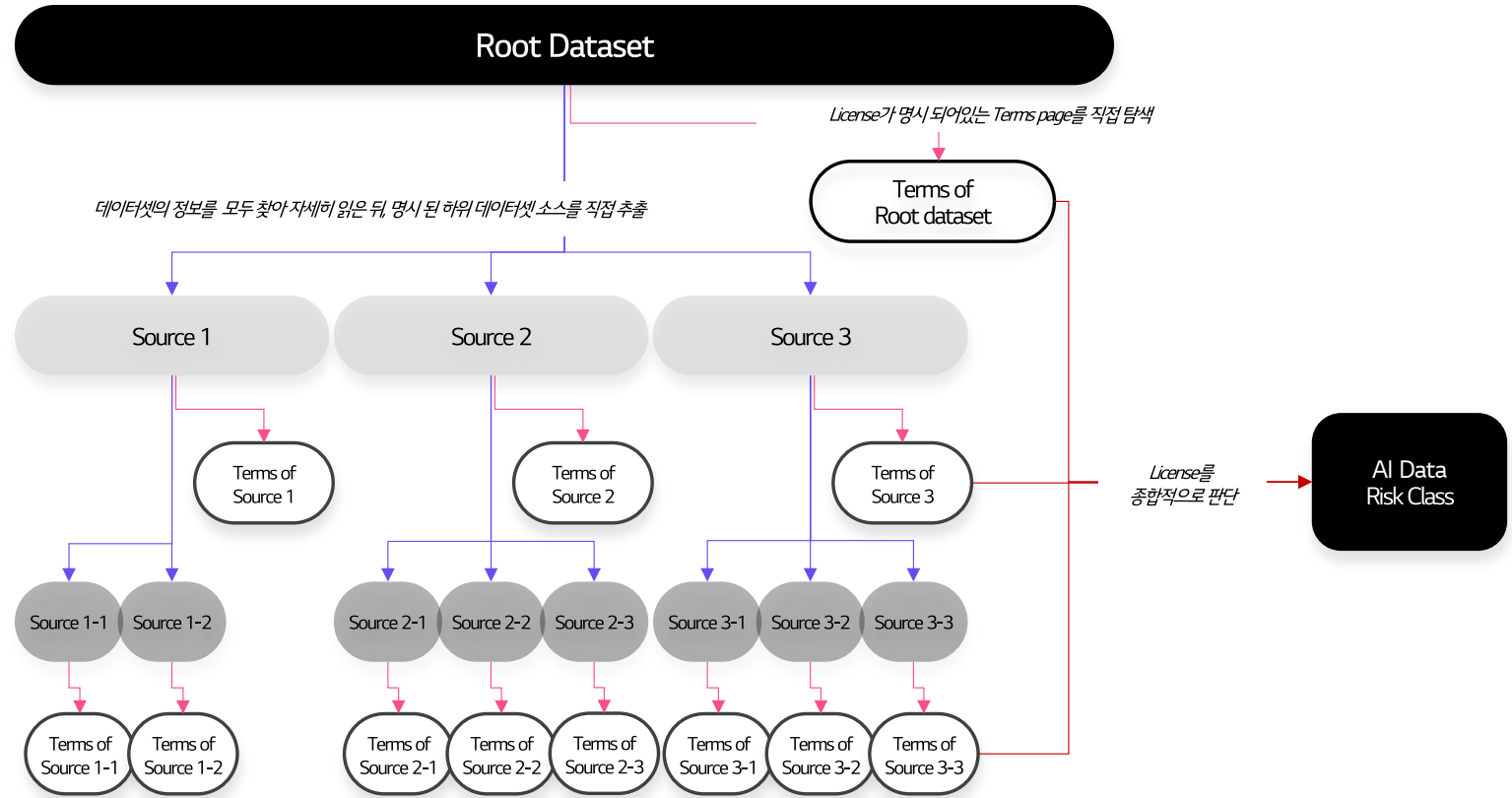


## — 현대 AI 학습 데이터 검토의 어려움

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

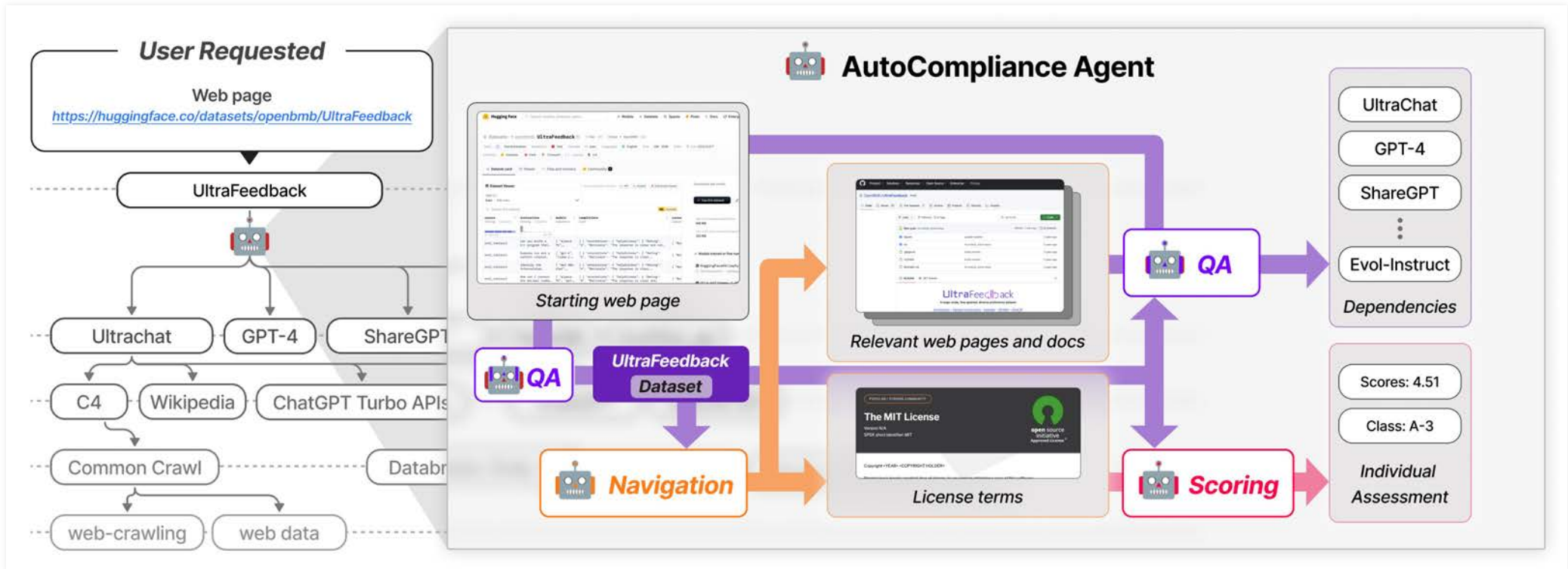


모든 법적 검토가 필요한 소스들을 다 찾을 때까지 반복



## EXAONE NEXUS를 활용한 Data Compliance 자동화 구조

- EXAONE NEXUS는 EXAONE 3.5 32B를 기반으로 개발되었습니다.
- QA, Navigation, Scoring 모듈을 통해 데이터셋의 정보를 바탕으로 어떤 하위 데이터 소스로 데이터셋이 구성되어 있는지, 각 데이터 셋의 라이선스 정보가 어떻게 되는지를 자동적으로 탐색합니다.



PKU-Alignment/PKU-SafeRLHF

huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF

Chrome is being controlled by automated test software.

Hugging Face Search models, datasets, users... Models Datasets Spaces Posts Docs Pricing Log In Sign Up

Datasets: PKU-Alignment PKU-SafeRLHF like 109

Tasks: Text Generation Modalities: Tabular Text Formats: json Languages: English Size: 100K-1M

Tags: safe safety ai-safety ilm lm human-feedback +2 Libraries: Datasets pandas Croissant +1 License: cc-by-nc-4.0

Dataset card Viewer Files and versions Community

### Dataset Card for PKU-SafeRLHF

**Warning:** this dataset contains data that may be offensive or harmful. The data are intended for research purposes, especially research that can make models less harmful. The views expressed in the data do not reflect the views of PKU-Alignment Team or any of its members.

Homepage | Single Dimension Preference Dataset | Q-A Dataset | Prompt Dataset

#### Dataset Summary

This dataset is a sibling project of PKU-SafeRLHF-v0 and BeaverTails.

We provide a high-quality dataset consisting of 83.4K preference entries, which is annotated across two dimensions: harmlessness and helpfulness. Specifically, each entry in this dataset includes two responses to a question, accompanied by safety meta-labels and preferences based on their helpfulness and harmlessness. For a more fine-grained labeling of Q-A pairs in this dataset, see PKU-SafeRLHF-QA.

In this work, we performed SFT on Llama2-7B and Llama3-8B with Alpaca 52K dataset, resulting in Alpaca2-7B and Alpaca3-8B. This dataset contains responses from Alpaca-7B, Alpaca2-7B, and Alpaca3-8B in the corresponding folders under /data.

The data collection pipeline for this dataset is depicted in the following image:

#### Labels of Each Q-A Pair

##### Harm Categories

Each Q-A pair in the dataset is labeled with one or more harm categories. Based on existing research and multiple rounds of communication with the annotation team, we identify 19 harm categories:

- S1: Endangering National Security
- S2: Insulting Behavior
- S3: Discriminatory Behavior

Downloads last month: 8,868

Use this dataset Edit dataset card

Size of downloaded dataset files: 465 MB

Size of the auto-converted Parquet files: 146 MB Number of rows: 164,236

#### Models trained or fine-tuned on PKU-Alignment/...

- NCSOFT/Llama-3-OffsetBias-8B Text2Text Generation Updated ... 1.13k
- chargodard/piano-medley-7b Text Generation Updated Jan 4 843
- NCSOFT/Llama-3-OffsetBias-RM-8B Text Classification Updated 29 da... 750
- chargodard/servile-harpsichord-cdpo Text Generation Updated Dec 10, 2023 603
- mradermacher/piano-medley-7b-11-GGUF Updated Aug 2 402
- mradermacher/piano-medley-7b-GGUF Updated Jun 4 378

Browse 22 models trained on this dataset

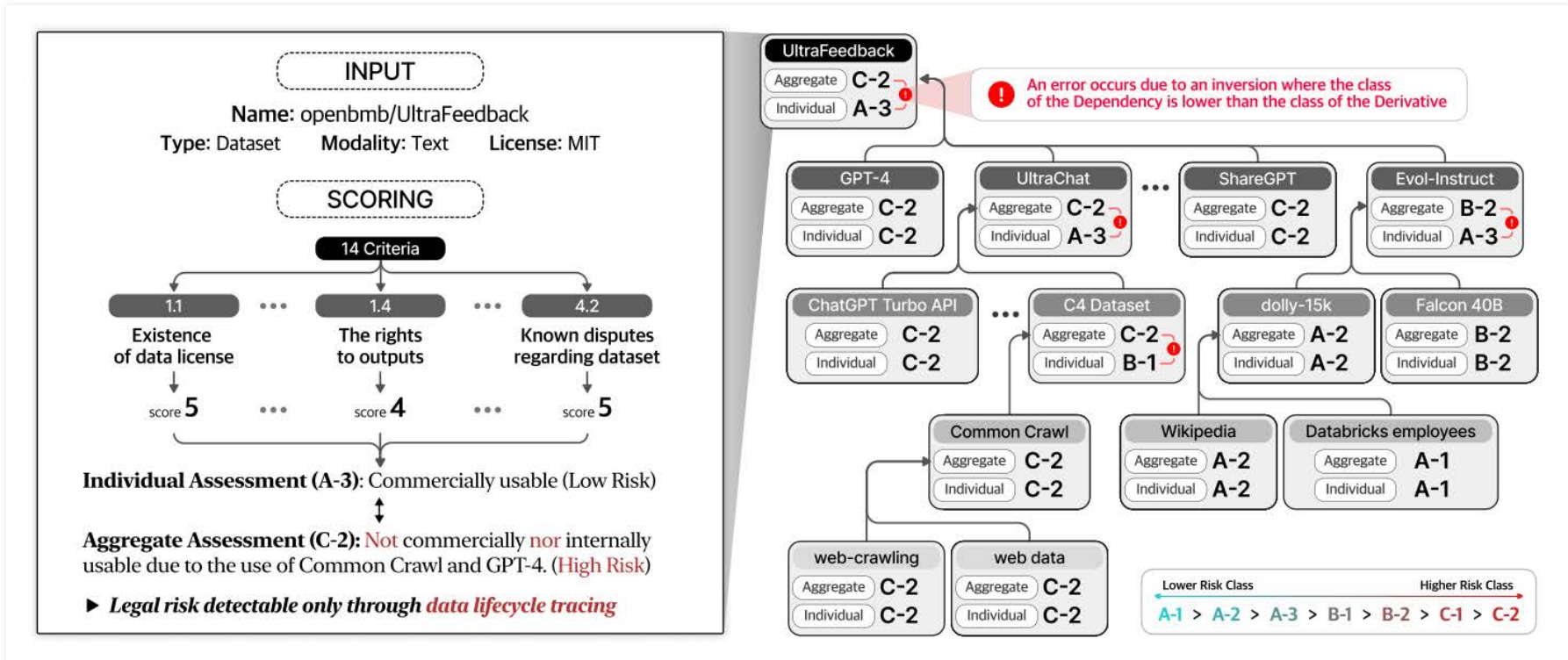
#### Collection including PKU-Alignment/PKU-SafeRL...

PKU-SafeRLHF Collection

A safety alignment preferen... 4 items Upd... 1

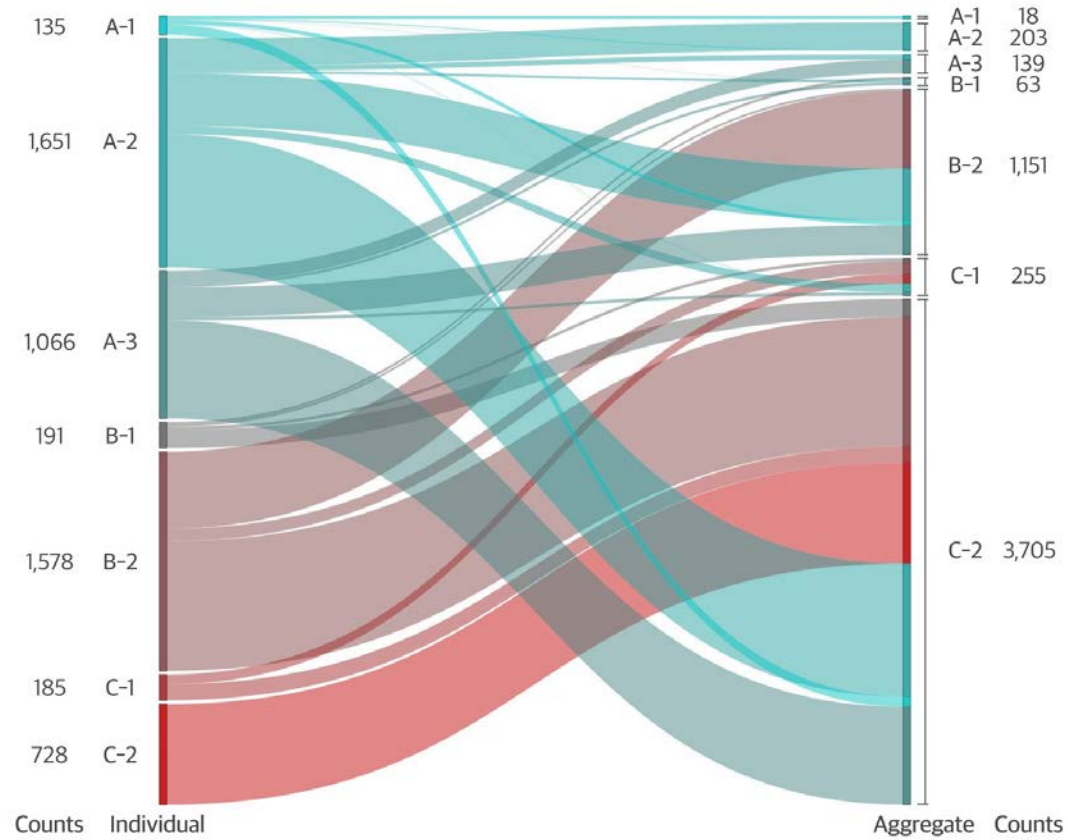
— EXAONE NEXUS를 활용한 Data Compliance 자동화 구조

- 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 성능

- EXAONE NEXUS는 인간 변호사와 비교했을 때 45배 빠른 속도, 0.1% 수준의 비용, 26% 빠른 정확도를 보였습니다.

Round	작성자	라이선스 데이터셋 작성	소요 시간	소요 금액
1	법무법인 변호사 5인	200개 작성 및 검토, 녹화 (HF Most Download 1-200)	29시간 / 인	62,634,000
2	법무법인 변호사 5인		(Total : 145시간)	
3	법무법인 변호사 5인	50개 Motherset만 재작성 (HF Most Download 1-50)	2.26시간 / 인 (Total : 11.3시간)	5,130,000
4	법무법인 변호사 5인	45개 (HF Most Download 201-245)	3.94시간 / 인 (Total : 19.7시간)	8,190,000
	EXAONE NEXUS	45개 (HF Most Download 201-245)	Total 0.9시간	\$2.5
5	법무법인 변호사 5인	155개 (HF Most Download 246-400)	10.02시간 / 인 (Total : 50.1시간)	20,900,000

Name	Set Accuracy (↑)	
	Dependencies	License terms
<b>AutoCompliance Agent</b>	<b>81.04%</b>	<b>95.83%</b>
Human expert	64.19%	87.73%
ChatGPT-4o	25.00%	39.81%
Perplexity Pro	28.24%	22.22%

Name	Time (sec)	Cost (\$)
<b>AutoCompliance Agent</b>	<b>53.1</b>	<b>0.29</b>
Human expert	2,418	207



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

- AI Dataset에 대한 검토는 변호사의 최종 검토가 여전히 필요합니다.

AI TRAINING DATASET LEGAL ASSESSMENT REPORTS Ultra Feedback **C-2**  
AIC-A00010-V1

## A.J. DataTrust Initiative

Certificate of Compliance

This dataset has undergone legal review by a qualified expert to ensure compliance with applicable laws and regulations.

Certified by:  
Saige Lim, Attorney-at-Law  
Date of Certification: 2025-08-04





AIC-A00010-V1  
Aug 11 2025



Assessment

Definitive Risk Class

# C-2

High Risk

Ultra Feedback's Compliance Results indicate a C-2, reflecting a high likelihood of violation, with some cases already escalated to litigation, arbitration, or regulatory action involving substantial financial stakes. The company faces a significant risk of becoming involved in such disputes.

License	Hierarchy*	Direct datasource*	Surface Risk*	Highest Risk*
MIT	14	7	A-3	C-2

\* Hierarchy: A hierarchy of sources with defined parent-child relationships, enabling full traceability from the root to all lowest-level.  
 \* Datasource: The exact source from which the data entry originates, including its repository or provider.  
 \* Surface Risk: Risk assessment based solely on the dataset information of the dataset requested by the user.  
 \* Highest Risk: Risk assessment based on the combined information of all datasources comprising the dataset requested by the user.

1

AI TRAINING DATASET LEGAL ASSESSMENT REPORTS Ultra Feedback **C-2**  
AIC-A00010-V1

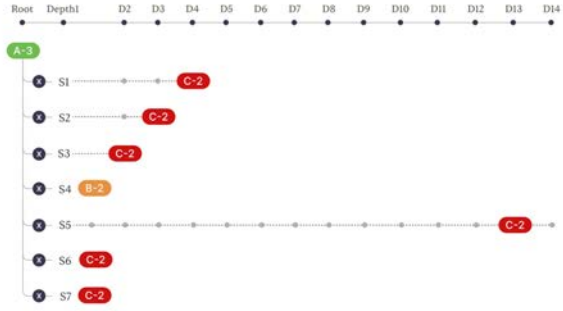
### Legal Risk

Direct Source

Code	Name	Type	License	Class
S1	Ultra Chat	Dataset	MIT	C-2
S2	Evol-Instruct	Dataset	Unknown	C-2
S3	truthful_qa	Dataset	Apache-2.0	C-2
S4	FalseQA	Dataset	Unknown	B-2
S5	FLAN	Dataset	CC-BY-4.0	C-2
S6	GPT-4	AI Model	Custom	C-2
S7	ShareGPT	SW/API	Apache-2.0	C-2

\* See Section 4 (6p) for the full hierarchy of all data.

Hierarchy & Risk Tracking



**Class Inversion: S1, S2, S3, S4, S5, S6, S7**  
 The class derived solely from Ultra Feedback's own information does not sufficiently reflect the all risks present across the underlying data sources. Consider trimming the dataset to reduce risks.

2

AI TRAINING DATASET LEGAL ASSESSMENT REPORTS Ultra Feedback **C-2**  
AIC-A00010-V1

### Results of 18 Risks

- 1. Data License:**
  - 1-1 The existence of a license to use the data Very High Risk
  - 1-2 BAuthorization to modify data and produce derivative works Very High Risk
  - 1-3 The potential for dispute arising from the outputs Very High Risk
  - 1-4 The rights to outputs Middle Risk
  - 1-5 The existence of an obligation to notify data usage High Risk
- 2. Data Use Period and Territory:**
  - 2-1 Restriction on data use period Middle Risk
  - 2-2 Whether the data license is revocable Middle Risk
  - 2-3 Restriction on AI model service period Middle Risk
  - 2-4 Restriction on data use territory Low Risk
- 3. Personal Information and Data Security:**
  - 3-1 Whether personal data is included in AI training data Very High Risk
  - 3-2 Whether data subjects have consented to the use of their data Very Low Risk
  - 3-3 Whether pseudonymized data is included in AI training data Very Low Risk
  - 3-4 Whether personal data may be entrusted or provided to third parties Very Low Risk
  - 3-5 Whether the scope of data users is limited Middle Risk
- 4. Additional Legal Risk:**
  - 4-1 Risks in the data collection process High Risk
  - 4-2 Known disputes involving the use of the same dataset in AI models Very High Risk
  - 4-3 Other contract risks associated with license agreements Middle Risk
  - 4-4 Types of license terms Very High Risk

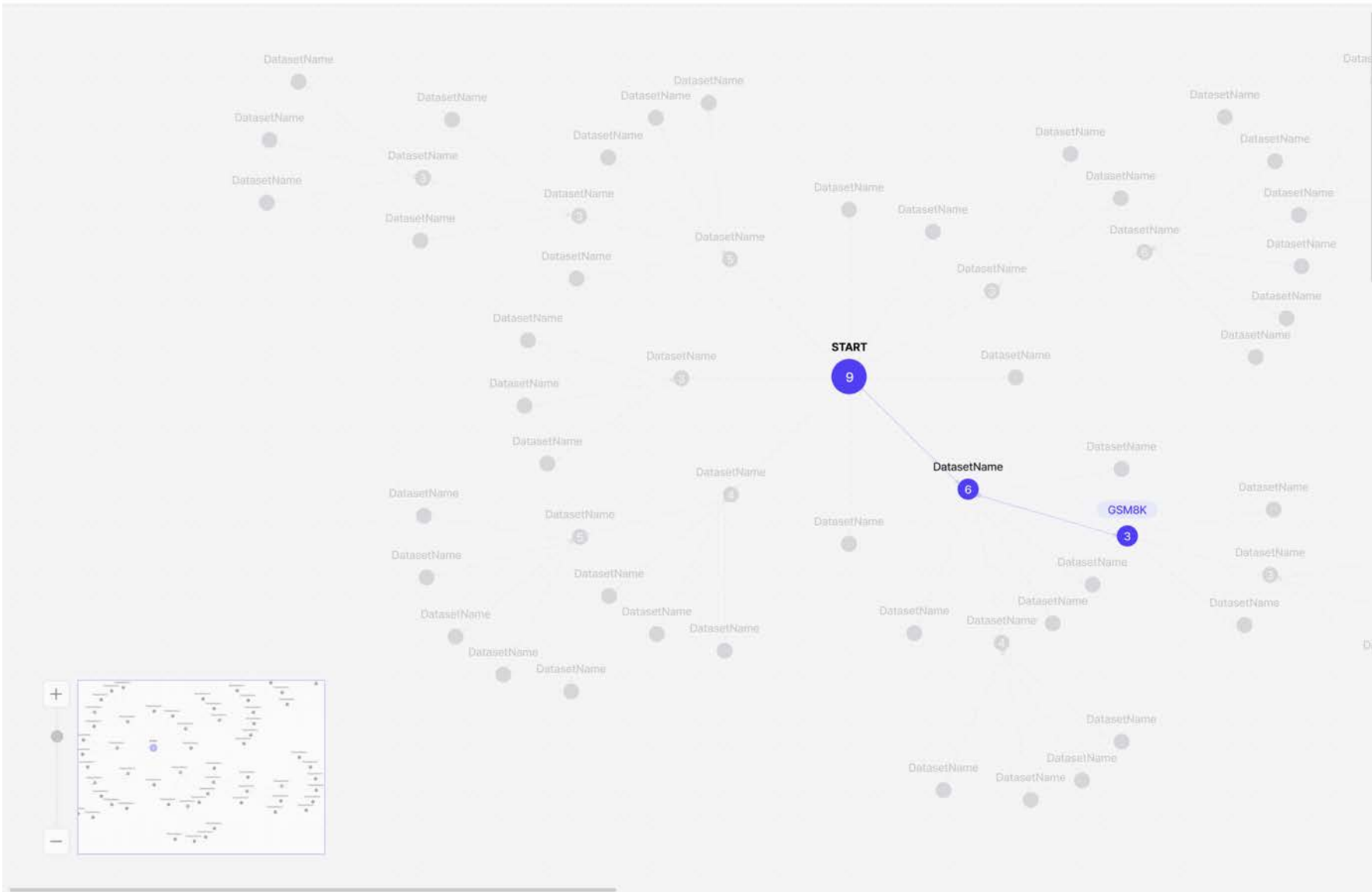
**Formula of Class Determination:**

$$B^{0.1} + C^{0.15} + D^{0.08} + E^{0.03} + F^{0.07} + G^{0.03} + H^{0.05} + I^{0.04} + J^{0.09} + K^{0.03} + L^{0.03} + M^{0.05} + N^{0.02} + O^{0.08} + P^{0.1} + Q^{0.05}$$

**Score and Condition:**  
 With a score of 2.27, there are no applicable conditions to downgrade the class.

A-1	A-2	A-3	B-1	B-2	C-1	C-2
						2.27

3



GSM8K

License 정보

Data URL [GSM8K](#)

License URL [MIT](#)

Source Name [news articles](#) [news](#) [blog posts](#)

Subset 정보 (7)

- 1 [AQuA-RAT](#) [Apache 2.0](#)
- 2 [TheoremQA](#) [Apache 2.0](#)
- 3 [GSM8K-RFT](#) [Non Listed](#)
- 4 [Out Curated](#) [Apache 2.0](#)
- 5 [NumGLUE](#) [Apache 2.0](#)
- 6 [AQuA-RAT](#) [Apache 2.0](#)
- 7 [AQuA-RAT](#) [Apache 2.0](#)

# Thank you

[legal@lgresearch.ai](mailto:legal@lgresearch.ai)