LG AI Research

Data Compliance

February 2025

Contents

Introduction

The data compliance system presented in this report is designed to ensure a safe use of data by identifying and evaluating potential legal issues that may arise from using third-party data as training data for artificial intelligence (AI) models, based on legal standards under current international laws, including the Copyright Law of the United States (Title 17), GDPR[1] and the Federal Trade Commission Act, and it further aims to examine the legal risks associated with the use of AI models developed using data in such manner.

The legality of utilizing specific data for AI training remains a subject of ongoing debate, with numerous disputes and legislative discussions occurring globally. While different countries and institutions may come to different conclusions depending on their jurisdictions and governing laws, the legal issues or risk of dispute that may arise from the use of datasets themselves are partly, if not entirely, of the universal nature. Therefore, the objective of this system is to proactively assess potential legal risks, recognizing them as real concerns with sufficient likelihood of materialization for a wide variety of stakeholders, including (i) developers using data to develop AI models, (ii) users implementing AI models in various applications, and (iii) service providers leveraging AI models for their offerings.

In this system, the standards for determining the category, score, and legal risk of each class in relation to the purpose and scope of data use are only preliminary standards established in consultation with legal professionals and intellectual property experts within the AI industry, considering the potential risks associated with AI models that are supplied and utilized internationally. These standards are neither absolute nor final. They should be used as flexible guidelines, taking into account specific legal standards, current legislative landscape, jurisdiction and governing law, and the type of AI model services in each country.

1 Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data

Class	Score	Category	Legal Risk
A-1	4.90	License/Privacy Risk Free	There is virtually no risk of legal disputes being filed in relation to data by original authors, licensors, data subjects, or related organizations even if the data is disclosed through in-house services or AI model services in public cloud.
A-2	4.57	License/Privacy Low Risk	While license or privacy issues exist with low likelihood of violation, there are no known cases where these issues have escalated to litigation, arbitration, or regulatory interventions.
A-3	4.22		While license or privacy issues exist with some likelihood of violation, there are no known cases where these issues have escalated to litigation, arbitration, or regulatory interventions.
B-1	3.73	License/Privacy Moderate Risk	License or privacy issues exist with high likelihood of violation, and there are a few cases that have escalated to litigation, arbitration, dispute or regulatory interventions. The company faces a slight risk of becoming involved in the dispute.
B-2	3.51		License or privacy issues exist with high likelihood of violation, and there are some number of cases that have escalated to litigation, arbitration, dispute or regulatory interventions. The company faces a slight risk of becoming involved in the dispute.
C-1	3.18	License/Privacy High Risk	License or privacy issues exist with substantially high likelihood of violation, and there are a sizeable number of cases that have escalated to litigation, arbitration, dispute or regulatory interventions. The company faces a risk of becoming involved in the dispute.
C-2	-		License or privacy issues exist with substantially high likelihood of violation, and there are cases that have escalated to litigation, arbitration, dispute or regulatory interventions involving substantial financial stakes. The company faces a notable risk of becoming involved in the dispute.

Risks Related to Data License

A. Assessment Criteria

Criteria	Weight
→ 1.1 The existence of a license to use the data	(Based on class scope)
→ 1.2 Authorization to modify data and produce derivative works	10%
→ 1.3 The potential for dispute arising from the outputs	15%
→ 1.4 The rights to outputs	8%
→ 1.5 The existence of an obligation to notify data usage	3%

* A score in a range from 1 (high risk) to 5 (low risk) is assigned based on the assessment of the risk for each criterion.

B. Scoring Standard for Each Criterion

→ 1.1 The existence of a license to use the data

Standard	Class Scope
Data is usable for commercial purposes without restriction	A-1
Data is explicitly usable for internal research purposes	A-2 – B-2
It is unknown whether the data is licensed	B-1 – C-1
Data is explicitly not usable	C-2

Data is protected under the laws of each country, primarily by copyright laws. When data constitutes a work of authorship, etc. protected by copyright laws, the copyright holder retains exclusive rights to use or authorize others to use the work.[2]

This exclusive right is guaranteed in most countries under the Berne Convention, a global copyright agreement (albeit limited to literary and artistic works).[3]

Therefore, the lawful utilization of copyrighted materials requires obtaining permission from the copyright holder, typically in the form of a license.[4]

2 United States Code Title 17-Copyrights, § 106
Subject to sections 107 through 122, the owner of copyright under this title has the exclusive rights to do and to authorize any of the following:
(1) to reproduce the copyrighted work in copies or phonorecords;
(2) to prepare derivative works based upon the copyrighted work;
(3) to distribute copies or phonorecords of the copyrighted work to the public by sale or other transfer of ownership, or by rental, lease, or lending;
(4) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and motion pictures and other audiovisual works, to perform the copyrighted work publicly;
(5) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audiovisual work, to display the copyrighted work publicly; and
(6) in the case of sound recordings, to perform the copyrighted work publicly by means of a digital audio transmission.

3 Berne Convention for the Protection of Literary and Artistic Works, Article 9(1)
Authors of literary and artistic works protected by this Convention shall have the exclusive right of authorizing the reproduction of these works, in any manner or form.

4 Article 46 of the Copyright Act (Authorization to Use Works)
(1) The holder of author's economic rights may grant another person authorization to use the work.
(2) The person who obtained such authorization pursuant to paragraph (1) shall be entitled to exploit the work in such a manner and within the limit of such conditions so authorized.
(3) The right of exploitation as authorized under paragraph (1) may not be transferred by assignment to the third party without the consent of the holder of author's economic rights.

Nonetheless, there are exceptions to copyright licensing requirements. Firstly, not all data types qualify as "work of authorship" eligible for copyright protection. Copyright eligibility is contingent on meeting specific criteria, with originality being a primary requirement.[5]

Therefore, data that lacks originality is not subject to copyright protection. However, in the US, the "originality" requirement has been described as a "famously low bar,"[6] which essentially requires that the author did not copy the work from another author and that the work has at least a "slight amount" of creativity.[7]

Secondly, individual nations' copyright legislation typically includes provisions for some limited use without the copyright holder's explicit permission. For example, Article 35-5 of the Copyright Act of Korea provides, "where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, such works may be quoted for news report, criticism, education, research, etc."

Copyright law in the US also contains similar provisions regarding "fair use."[8] The fair use doctrine allows limited use of copyrighted works without permission for purposes such as criticism, comment, news reporting, teaching, scholarship, or research.

Courts assess fair use based on four factors:

1. the purpose and character of the use, including whether it is commercial or transformative;
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used; and
4. the effect of the use on the market for the original work. In the digital era, this doctrine has been applied to mediums like e-books, online content, and AI-generated works, with a focus on whether the use "transforms" the original content into something new and of greater public benefit.

There is a history of courts applying the US Copyright Act to digitalized media and databases. For instance, under the precedent established in *Authors Guild v. HathiTrust*[9] and upheld in *Authors Guild v. Google*, the US Court of Appeals for the Second Circuit held that mass digitization of a large volume of protected books in order to distill and reveal new information about the books was a fair use.[10] While these cases did not concern generative AI, they did involve machine learning.

US courts are now hearing pending challenges to ingestion for training generative AI models. For instance, the *Thomson Reuters v. ROSS* Intelligence dispute underscores critical tensions in applying copyright law to AI, as what is likely to be the first substantive AI trial in the US.[11] At the heart of the dispute is whether ROSS's alleged use of Westlaw case summaries to train AI-powered legal research tools constitutes copyright infringement or falls under the protective umbrella of "fair use."

The court's refusal to grant summary judgment on fair use for both parties, followed by the postponement of the trial and request for renewed summary judgment briefing suggests the court's acknowledgment of the evolving nature of AI technologies and their potential ramifications on copyright law.

5 United States Code Title 17-Copyrights, § 102(a)
Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. [...]
Article 2 of the Copyright Act (Definitions)
1. The term "work" means a creative production that expresses human thoughts and emotions.

6 *Gray v. Hudson, 28 F.*4th 87, 97 (9th Cir. 2022).

7 *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.,* 499 U.S. 340, 345 (1991).

8 United States Code Title 17-Copyrights, §107
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. [...].

9 *Authors Guild, Inc. v. HathiTrust,* 755 F.3d 87 (2d Cir. 2014).

10 *Authors Guild v. Google, Inc.,* 804 F.3d 202 (2d Cir. 2015).

11 *Thomson Reuters Enter. Ctr. GmbH v. Ross Intel. Inc.,* 694 F. Supp. 3d 467, 481 (D. Del. 2023).

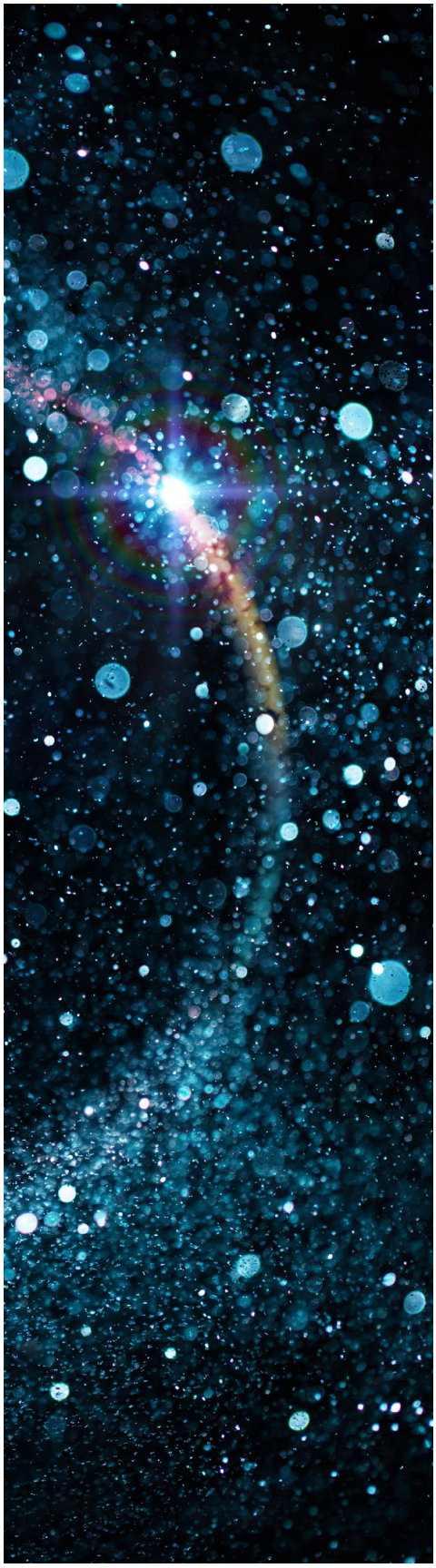
A key implication of this case for future AI copyright disputes is whether AI training constitutes transformative use and how courts weigh the potential public benefits of AI innovation against the rights of copyright holders. As briefing and hearings continue, the outcome could significantly influence how courts nationwide approach fair use defenses in AI copyright disputes, particularly regarding the balance between fostering technological innovation and protecting intellectual property rights.

Furthermore, certain jurisdictions, such as Japan and EU, have text and data mining (TDM) exemption provisions that allow the utilization of works of authorship in certain circumstances.[12] This means that, depending on the country, it may be possible to collect and use data online, such as by relying on TDM or fair use provisions if it is for scientific research or if there is no explicit restriction on the use of works.

However, it is difficult to determine whether certain data qualifies as a work of authorship before using it. That is because whether data is a work of authorship is a matter of normative judgment, and even the copyright register does not provide definitive verification of a work's copyright status as most countries adhere to a no-formality principle, according to which "copyright arises when a work is created and the establishment of copyright does not require undergoing any procedures or formalities."[13]

Further, the applicability of exceptions to copyright laws, such as fair use exceptions, is difficult to ascertain in advance.

In addition, the scope of legal protection for data varies across countries. In the US for instance, while individual facts are not protected by copyright, the creative selection and arrangement of these facts into a database likely qualifies for protection as a compilation.



12 Article 47-5 of the Copyright Act of Japan, and Directive (EU) 2019/790 Article 3, 4.

13 Berne Convention for the Protection of Literary and Artistic Works, Article 5(2)
The enjoyment and the exercise of these rights shall not be subject to any formality.

This principle was established in the precedential Supreme Court case *Feist Publications, Inc. v. Rural Telephone Service Co.,* 499 U.S. 340 (1991), which held that "factual compilations . . . may possess the requisite originality" if the selection and arrangement of data involve a minimal degree of creativity.[14] The ruling limits copyright to the "selection and arrangement" of data but not the content itself.

This distinction weakens database protection relative to most other countries. By denying copyright to purely factual compilations, Feist incentivizes database creators to add innovative features or present data in creative ways to secure intellectual property protection. Companies relying on factual data need to explore other legal protections, such as trade secret laws, contracts (e.g., terms of service), or technological measures to safeguard their databases.

Whereas the EU has much broader protections for databases. Directive 96/9/EC, which serves as a guideline for EU database regulations, establishes database rights and protects them as a copyright. Consequently, even data that are not qualified as a "work of authorship" may still be entitled to legal safeguards, which calls for a need to carefully consider the associated risks.

Moreover, in Korea, the Copyright Act separately provides protection to the producers of "databases."[15] According to the same Act, even the datasets the individual data of which lacks creative nature (i.e. fail to qualify as "works of authorship") or the selection, arrangement or composition of the individual data of which lacks creative nature (i.e., fail to qualify as "compilation works") can be protected as "databases." Additionally, the Unfair Competition Prevention and Trade Secret Protection Act of Korea provides for sanctions against improper use of data.

↓

In light of these considerations, we have established a risk assessment framework based on different risk levels as follows:

1 **Class A-1** where data is explicitly licensed for commercial use or an equivalent license has been granted under major copyright laws by the rights holder (e.g., copyright holders) or where it is not legally protected (such as because it is in the public domain[16]), as there is a minimal risk of legal issue in such cases;

2 **Class A-2** to **B-2** where data use is authorized only for specific purposes, such as internal, research, personal, or non-commercial use, and utilization beyond these specified scopes may potentially infringe upon legal boundaries;

3 **Class B-1** to **C-1** where the licensing status is unknown, and consequently there may still be potential for lawful use based on fair use principles or TDM exemptions; and

4 **Class C-2** where data use is explicitly not permitted, in particular where restrictions are placed on AI learning applications, which leads not only to substantial possibility of unlawful data utilization, but to significantly increased likelihood of legal disputes, barring extremely exceptional circumstances.

14 *Feist Publications, Inc. v. Rural Telephone Service Co.,* 499 U.S. 340 (1991).

15 Article 2, Subparagraph 19 of the Copyright Act of Korea
The term "database" means compilation whose materials are systematically arranged or composed, so that they may be individually accessed or retrieved.

16 Once a set period of time for copyright protection expires, or if the creator failed to comply with the necessary legal requirements at the time of creation or thereafter, the work enters the public domain. This means that the work is freely available for everyone to use without any restriction.

1.2 Authorization to modify data and produce derivative works

Standard	Score
The company is authorized to make all forms of modifications and alterations, including production of derivative works	5
Production of derivative works is prohibited, but modification or alteration to an extent not constituting derivative works is permitted	3
It is unknown whether the company is authorized to modify data or produce derivative works	2
All forms of modifications and alterations, including production of derivative works, are prohibited	1

A derivative work is a creative work produced by means of translation, arrangement, alteration, adaptation, cinematization, etc. of an original work. Such derivative works must be protected as independent works and the right to produce derivative works is vested exclusively in the author of the original work.[17]

The US copyright laws recognize the author's right to integrity, but this right generally applies to fine art categories of "works of visual art": paintings, sculptures, drawings, prints, and still photographs produced for exhibition, and expressly excluding electronic publications.[18]

Thus, US copyright law likely does not protect moral rights for the type of electronic data typically used to train AI models. US trademark law, on the other hand, through Section 43(a) of the Lanham Act, provides additional protections against false endorsement and misrepresentation, which can intersect with generative AI.[19]

Cases like *Waits v. Frito-Lay, Inc.*[20] and *Fifty-Six Hope Rd. Music, Ltd. v. A.V.E.L.A., Inc.* illustrate how creators or their estates can bring claims if AI-generated works misuse identifiable traits, such as an artist's voice, likeness, or brand, in ways that imply false association.[21]

For instance in *Andersen v. Stability AI*, in August 2024, the court ruled on a defendants' motions to dismiss.[22] They allowed the artists to proceed with their claims of direct copyright infringement against Stability AI, Midjourney, and Runway AI, as well as Lanham Act claims against Midjourney. This decision is significant as it permits the artists to continue pursuing their allegations that the AI companies' use of their works for training purposes constitutes false endorsement under the Lanham Act. The case is ongoing, and further developments are anticipated as it progresses through the legal system.

Another potential issue relates to whether generative AI is used to produce content that mimics a celebrity's voice or visual identity without authorization, where the Lanham Act might serve as a legal basis for claims of false endorsement or consumer confusion, especially when such uses commercially exploit the creator's identity.

17 Article 22 of the Copyright Act of Korea
United States Code Title 17-Copyrights, § 106
Subject to sections 107 through 122, the owner of copyright under this title has the exclusive rights to do and to authorize any of the following: (2) to prepare derivative works based upon the copyrighted work;.

18 United States Code Title 17-Copyrights, § 106A
Subject to section 107 and independent of the exclusive rights provided in section 106, the author of a work of visual art [...] shall have the right-(A) to prevent any intentional distortion, mutilation, or other modification of that work which would be prejudicial to his or her honor or reputation, and any intentional distortion, mutilation, or modification of that work is a violation of that right [...].

19 15 USCS § 1125.

20 *Waits v. Frito-Lay, Inc.,* 978 F.2d 1093 (9th Cir. 1992).

21 *Fifty-Six Hope Rd. Music, Ltd. v. A.V.E.L.A., Inc.,* 778 F.3d 1059 (9th Cir. 2015).

22 *Andersen v. Stability AI Ltd.,* No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204 (N.D. Cal. Aug. 12, 2024).

State laws addressing the right of publicity and privacy introduce additional complexities, particularly when AI-generated works simulate personal identifiers such as voices or likenesses. State laws vary widely between jurisdiction (2/3 of states recognize the right of publicity under statutes and/or common law), prompting a push for formalized federal legislation.

However, there has been no major successes shifting these issues away from states. These state laws typically contain carveouts to protect expressive speech under the First Amendment, such as for artistic expression in film and TV.[23]

Few states recognize publicity rights with carveouts allowing creators' heirs to bring claims if generative AI reproduces their likeness in unauthorized ways. Additionally, state privacy laws may apply when AI collects, processes, or reproduces personal data, such as images or biometric identifiers. Overlaps between publicity rights and privacy protections highlight the need for careful navigation when deploying generative AI, especially in jurisdictions with robust privacy statutes like California's Consumer Privacy Act (CCPA).[24]

The existing framework is best described as a patchwork of federal and state laws regulating the intersection of generative AI, copyright, trademark, and data privacy.

Meanwhile, under the Copyright Act of Korea, the author's moral rights include the right to integrity of the content of his or her work.[25] Consequently, even modifications or alterations that do not reach the level of derivative work production may potentially infringe upon the author's integrity right.

In other words, not only the creation of derivative works based on data that are original works but also the modifications or alterations that fall short of the production of derivative works may be subject to legal protection. Therefore, any party other than the original author must obtain an appropriate license to engage in such activities.

Since the scope of a license is contingent upon specific terms delineated in the license agreement,[26] even if a license to use the work is granted, it may not extend to the production of derivative works or other forms of modification and alteration (it should be noted in particular that the production of derivative works is generally not covered by the license unless it is expressly permitted in the license agreement).[27]

For instance, in *Jacobsen v. Katzer*, parties violated a licensing agreement by incorporating the licensed code into their marketed software products. The court held that "it is outside the scope of the Artistic License to modify and distribute the copyrighted materials without copyright notices and a tracking of modifications from the original computer files."[28]

However, for state law claims, claimants must also be wary of federal preemption challenges. For example, in relation to AI, generally US courts have held that AI disputes based on the reproduction of code in output and preparation of derivative works are preempted by US copyright law, whereas a dispute based on unauthorized use of code (e.g., in violation of terms of service) for training purposes is sufficient to avoid preemption.[29]

23 California Civil Code § 3344 ("Any person who knowingly uses another's name, voice, signature, photograph, or likeness, in any manner, on or in products, merchandise, or goods, or for purposes of advertising or selling, or soliciting purchases of, products, merchandise, goods or services, without such person's prior consent, or, in the case of a minor, the prior consent of his parent or legal guardian, shall be liable for any damages sustained by the person or persons injured as a result thereof..") ; Tenn. Code § 47-25-1105; New York Consolidated Laws, CVR § 50.

24 California Civil Code §§ 1798.100–1798.199.100 (West 2024).

25 Article 13(1) of the Copyright Act of Korea.

26 17 U.S.C.S. § 106; Article 46(2) of the Copyright Act of Korea, etc.

27 *Utopia Provider Sys. v. Pro-Med Clinical Sys., L.L.C.*, 596 F.3d 1313, 1316 (11th Cir. 2010).

28 *Jacobsen v. Katzer*, 535 F.3d 1373, 1382 (Fed. Cir. 2008); See also Nimmer on Copyright § 10.15 ("An express (or possibly an implied) condition that a licensee must affix a proper copyright notice to all copies of the work that he causes to be published will render a publication devoid of such notice without authority from the licensor and therefore, an infringing act.").

29 The court originally dismissed the state law claims of interference, unjust enrichment, and unfair competition, with leave to amend in 2023. Later in 2024, the court again granted dismissal of the state law claims, as well as the DMCA Copyright Claim, with no leave to amend. *Doe v. Github, Inc.*, 672 F. Supp. 3d 837, 2023 U.S. Dist. LEXIS 86983 (N.D. Cal. 2023), dismissed, 2024 U.S. Dist. LEXIS 11068 (N.D. Cal. Jan. 3, 2024).

This limits licensing claims in the US relative to other countries by putting an emphasis on the web scraping and training of LLMS, rather than the creation of derivative works in violation of a licensing agreement.

It should also be noted that, some datasets may be released under open-source licenses or Creative Commons Licenses (CCL), which may impose specific restrictions on their utilization. The CCL's No Derivative Works (ND) license, which explicitly prohibits modification or alteration of the original work, is a prime example.

Another is General Public Use Licenses (GNU) which allows the copying, modifying, and distributing of licensed software so long as the license and copyright notices are provided, and any modifications are labeled therein.

A third option is an Open Database License, which licenses the right to share, modify, and use protected databases. Consequently, even in instances where data is not eligible for copyright or other legal protection, the potential for legal risk persists due to contractual terms or standard terms and conditions that proscribe data modification or alteration.

As explained above, data may be under the safeguard of copyrights, contracts, or standard terms and conditions, and the pre-processing phase, which refers to the work of processing data beforehand to facilitate efficient learning by AI models, may potentially constitute production of derivative works, or modification or alteration of the original work.

Furthermore, no clear legal judgement has so far been rendered as to whether the outputs generated by trained AI models are derivative works or modifications or alterations of the original data, which is another legal risk factor.

↓

In light of these considerations,
we assess this item based on different risk levels and allocate scores as follows:

1 **5 points** where the copyright holder has explicitly granted broad modification rights and the likelihood of legal complications is thus minimal;

2 **3 points** where no right to produce derivative works is granted but only modifications or alterations that do not reach the level of derivative works, or where specific conditions for alterations are stipulated (this is because, in such cases, the line between derivative works and modifications or alterations that do not reach the level of derivative works is often blurred and distinguishing between the two is a matter of normative judgment);

3 **2 points** where the authorization status regarding derivative work production or modification rights is unknown, given the possibility that the data may not qualify as copyrightable work or their use may be permissible under fair use doctrine or TDM exemptions; and

4 **1 point** where modifications and alterations are explicitly prohibited, leading to a high likelihood of not only unlawful data utilization but also legal disputes, barring extremely exceptional circumstances.

1.3 The potential for dispute arising from the outputs

Standard	Score
Consent of the original author is obtained or no output is generated	5
Output dissimilar to the original work is generated	4
There is a potential that outputs similar to the original author's work will be generated, but this potential remains low	3
There is a potential that outputs will contain part of the original author's work, irrespective of the potential for outputs similar to the original work being generated	2
There is a high likelihood that outputs similar to the original author's work will be generated	1

Even if AI models are trained using lawfully obtained data, it does not preclude the potential for legal disputes. For example, data for which a license has been acquired may subsequently prove to be a work that infringes someone's copyrights, or the rights holder may turn out to be a different person or entity than was originally known, or the company using the data might have made an incorrect assessment of the legal risk of using it.

As an example of such a dispute, in *Complex Sys., Inc. v. ABN AMRO Bank N.V.*,[30] a dispute arose when ABN AMRO Bank N.V. ("ABN") sold some of its assets -- including subsidiaries LaSalle Bank and ABN AMRO Information Technology Services Company, Inc. ("IT") -- to Bank of America ("BAC") for $21 billion ("the LaSalle Transaction"). IT was the licensee of a software application created and licensed by plaintiff Complex Systems, Inc. ("CSI") called BankTrade. By virtue of its corporate affiliation with ABN, ABN was entitled to use BankTrade through IT and in reliance on IT's license. Following the LaSalle Transaction, IT remained the licensee.

Although ABN did not have a separate license and was not itself a licensee, ABN continued to use BankTrade. The court granted a permanent injunction against ABN, holding that a bank that lost its license to use a software application when it sold its subsidiary in 2007 must stop using the copyrighted software.

In another example, *Warner Chappell Music, Inc. v. Nealy*,[31] a dispute arose approximately ten years after Warner Chappell Music, Inc. ("Warner") obtained a license from Tony Butler, one of the co-authors of the musical work at issue, unbeknownst to the other co-author, Sherman Nealy, while Nealy was serving a prison sentence. When Nealy finished his service and found out that Warner was profiting from his musical work, Nealy sued Warner for copyright infringement.

Although Nealy brought claims for infringing acts occurring approximately ten years before he filed suit, the appellate court and the Supreme Court both held that "the Copyright Act contains no separate time-based limit on monetary recovery," holding in favor of Nealy.

30 *Complex Sys., Inc. v. ABN AMRO Bank N.V.*, 08 Civ. 7497 (KBF) (S.D.N.Y. May 9, 2014).
31 *Warner Chappell Music, Inc. v. Nealy*, 144 S. Ct. 1135 (2024).

Consequently, the training and the use of AI models inherently entail a certain degree of legal risk, even when using data obtained through lawful procedures. The likelihood of this risk may vary depending on the nature of the AI models' output, even if they were trained with the same data.

For example, AI models designed for target identification (e.g., classification AI and recognition AI) provide limited grounds for legal action by the original data rights holders against the AI models because even if the models are trained on specific image data (e.g., dog images), they only perform classification or identification tasks (i.e., classify or identify input images as dogs or not) without releasing the original dog images or those that closely resemble them.

Conversely, AI models generating specific outputs (e.g., generative AI) pose a higher risk of legal challenges because when trained on specific data (e.g., dog images), some of them may produce outputs that closely resemble or potentially replicate the original dog images, and in such cases, the rights holders of the original data may have meaningful grounds for legal action.

In the US, legal disputes based solely on AI's use of copyrighted works to train an AI model has proved largely futile. Several motions to dismiss have been granted over the last decade based on ill-pled theories such as vicarious copyright infringement and various Digital Millenium Copyright Act ("DMCA") claims.

DMCA Section 1202(a)(1) claims fails where the plaintiff is unable to allege an infringing derivative work or output. With respect to a DMCA Section 1202(b) plaintiffs must plausibly allege that the engineers who trained the model intentionally removed the copyright management information from the copyrighted works.

A successful claim based on training data without a showing of similarity in the output could be a significant challenge to the AI industry, as virtually every AI generator could be liable for infringement just by using copyrighted data to train their AI model.

For instance, the court in *Doe v. GitHub, Inc.* affirmed that "courts have held that no DMCA violation exists where the works are not identical."[32]

This includes generated outputs that are "modified," "variants," or "functional equivalents" of the original protected materials. This ruling was again affirmed in *Andersen v. Stability AI Ltd*, where the court dismissed copyright claims related to images generated by Stable Diffusion, since Plaintiffs had failed to allege the produced images were identical to the protected materials.[33]

Moreover, there is the underlying difficulty in establishing Article III standing in bringing certain copyright claims ("CMI") absent output from a generative AI model. Courts have held that use of copyrighted materials for training purposed in insufficient to show injury in fact pursuant to Article III standing.

For instance, in *Doe v. Github, Inc.*, plaintiffs failed to establish Article III standing under the CMI provision of the DMCA where the alleged misuse occurred solely during the training of an artificial intelligence model.[34]

32 *Doe v. GitHub, Inc.*, No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068, at *24 (N.D. Cal. Jan. 3, 2024) (finding that the source code generated by CoPilot, even though it was functionally equivalent and nearly identical to the Plaintiff's source code, was "not sufficient for a Section 1202(b) claim.").

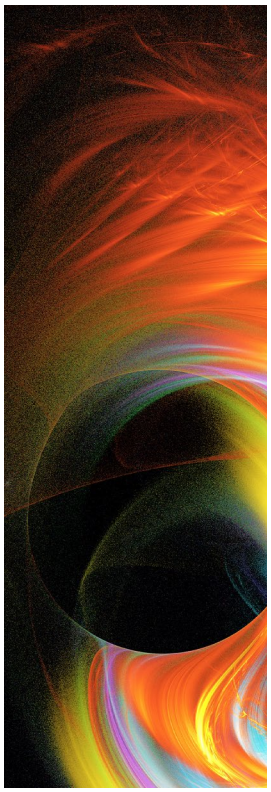
33 *Andersen v. Stability AI Ltd.*, No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204, at *22 (N.D. Cal. Aug. 12, 2024).

34 *Doe v. Github, Inc.*, 672 F. Supp. 3d 837, 850 (N.D. Cal. 2023) ("[W]hile Plaintiffs identify several instances in which Copilot's output matched licensed code written by a Github user, [] none of these instances involve licensed code published to GitHub by Plaintiffs. Because Plaintiffs do not allege that they themselves have suffered the injury they describe, they do not have standing to seek retrospective relief for that injury.").

Similarly, in *Raw Story Media, Inc. v. OpenAI Inc.*, Plaintiffs alleged that their copyrighted works were used as training data and stored.[35]

The court held plaintiffs failed to allege any actual adverse effects stemming from the DMCA violation since only the earlier, rather than current versions of ChatGPT were shown to create reproductions.

Given the above, if it is externally apparent that certain data has been used for AI training, it increases the likelihood of legal dispute; conversely, the likelihood decreases when such resemblance is not readily apparent. For example, data that is difficult to visually recognize, such as audio, numerical data, large tabular formats and chemical molecular structures are less likely to result in litigation than text or images.



↓

In light of these considerations, we assess this item based on different risk levels and allocate scores as follows:

1 **5 points** where explicit consent from the original author has been obtained for derivative work production or modification, or where the relevant AI model is of a type that does not generate specific outputs, because in this case, there is a minimal likelihood of legal complications or claims by the original author or third parties;

2 **4 points** where the original work is of a nature that makes it difficult for generative AI systems to produce outputs that closely resemble it (although some aspects of the original work can be inferred), such as audio data, numerical data, tabular information, or chemical molecular structures, and the production of such outputs is therefore unlikely to pose a risk of legal dispute;

3 **3 points** where outputs similar to the original work can be produced, considering that while there exists a potential for legal disputes, the probability of such outputs being produced is not high;

4 **2 points** where elements that appear to be similar to the original work may be included in the outputs, albeit not to the extent of complete similarity, and there is therefore an increased risk of legal claims by copyright holders or third parties; and

5 **1 point** where there is a high likelihood of the production of outputs that closely resemble the original work, in which case the fact that the original work was used to train the AI will clearly show and the risk of legal disputes will be highest.

35 *Raw Story Media, Inc. v. OpenAI, Inc.*, 2024 U.S. Dist. LEXIS 204101, at *12 (S.D.N.Y. Nov. 7, 2024).

1.4 The rights to outputs

Standard	Score
The company has ownership of or intellectual property rights to the output	5
The company has the right to use the output	4
It is unknown whether the company has the right to use the output	3
The company explicitly does not have the right to use the output	1

There is controversy over who should own the rights to the output of AI models and, in particular, how these rights should be allocated between the owners of training data, AI model users and developers. Issues primarily relate to copyrights, and they revolve around (i) whether the AI model output is a "work of authorship" and (ii) if it is, who the author is. Legal standards to determine these issues vary from country to country: while Korea rarely recognizes AI output as a work of authorship[36], China has lower court precedents where output generated by image-generating AI models is recognized as a work of authorship, and the unauthorized use of such output is therefore considered copyright infringement.[37]

In the US, copyright can protect only material that is the product of human creativity. Most fundamentally, the term "author," which is used in both the Constitution and the Copyright Act, excludes non-humans.[38]

Works that contain AI generated materials (e.g., images or text) may be eligible for copyright protection only if they exhibit significant human authorship.

Specifically, the human contributor must exercise creative control over the selection, arrangement, or modification of AI-generated content to produce a final work that reflects human originality.

This approach ensures that copyright protection is not extended to fully autonomous AI outputs, as such creations lack the necessary human authorship to qualify under US law.

This decision was affirmed in *Thaler v. Perlmutter*, where the court addressed the issue of "how much human input is necessary to qualify the user of an AI system as an 'author' of a generated work."[39]

Plaintiffs that can demonstrate the substantial similarity of AI output to the copyrighted works may be able to raise a viable legal dispute regarding ownership over the output of AI models.

In *Anderson v. Stability AI*, plaintiffs alleged that the AI-generated art produced by defendants was substantially similar to the plaintiffs' original works and attached a 150-page exhibit to the complaint showing exemplary images.[40]

36 "A Guide on Generative AI and Copyright," Ministry of Culture, Sports and Tourism (Dec. 2023).

37 Beijing Internet Court (2023) Beijing 0491 Civil Case, First Instance No. 11279.

38 U.S. Copyright Office, Registration of Works Containing Material Generated by Artificial Intelligence (Mar. 16, 2023) Available at: https://www.copyright.gov/events/ai-application-process/Registration-of-Works-with-AI-Transcript.pdf; See also Guidance on Copyright Registration of Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190 (Mar. 16, 2023) Available at: https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

39 *Thaler v. Perlmutter*, 687 F. Supp. 3d 140, 149 (D.D.C. 2023) (currently on appeal following a decision by the DC District Court (687 F.Supp.3d 140) that copyright law requires human authorship of a work to qualify for copyright protection).

40 *Anderson v. Stability AI Ltd.*, No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204, *33-35 (N.D. Cal. Aug. 18, 2024).

This similarity, according to the plaintiffs, indicated that their copyrighted art had been used without permission to train the AI models, leading to outputs that closely resembled their creations.

The degree of similarity of the copyright work to the output of the AI model may therefore be important in affecting the outcome of US legal disputes.

Whether the output of an AI model is substantially similar, as is alleged in *Anderson*, or less so will factor into infringement analyses and defenses such as "fair use." As data sets grow and AI becomes more advanced it may be more difficult for plaintiffs to map their works to the output of the AI models.

Against this backdrop, it should be noted that in the event that the output of an AI model trained on lawfully obtained data is in turn used to train another AI model, there may still be legal risks associated with using such output or prompt (despite having legal permissions to use the relevant training data) if contractual or legal rights to the model's outputs or prompts have not been secured.

In light of the above,
we assess this item based on different risk levels and allocate scores as follows:

1 **5 points** where the data is not subject to legal protection, or where it is clear that the company directly possesses rights such as ownership or intellectual property rights to the output, because in this case, the legal risk will be minimal;

2 **4 points** where the company has secured the rights for output utilization, if not ownership rights, and thus the likelihood of legal disputes is not too high, although it cannot be ruled out;

3 **3 points** where the company does not know whether it has rights to the output due to lack of clarity in contractual stipulation, since the legal risk is not high as long as the output is not utilized for derivative works; and

4 **1 point** where the company explicitly does not have rights to the output, because in this case, the company is effectively prohibited from utilizing the output for derivative works.

1.5 The existence of an obligation to notify data usage

Standard	Score
The company is able to provide data usage notification at any time (irrespective of the existence of data notification obligations)	5
The company is unable to provide data usage notification and has no notification obligation	4
The company is unable to provide data usage notification and has a notification obligation	2

AI model training frequently involves the use of datasets licensed through open-source agreements or CCLs, which often obligates the user of such datasets to provide disclosure of the use of such data.

For instance, the Apache License 2.0 mandates the disclosure of license content and any modifications to the utilized data, while the MIT License and CCLs require attribution of authorship and source (BY).[41]

In certain jurisdictions, disclosure obligations are imposed on the users of copyrighted works. For example, Article 37(1) of the Copyright Act of Korea stipulates that users making fair use of copyrighted works must specify the source of the work.

Furthermore, in some countries, data collectors, when collecting information that constitutes personal data, are obligated to notify data subjects of the fact that their data is being collected, along with other relevant facts.[42]

For instance, several US states, including California, Colorado, Utah, and Tennessee, have recently passed data and privacy protection laws specifically related to generative AI.[43]

Whereas several US federal AI regulatory initiatives have been proposed are introduced to the House as of 2024. The AI Foundation Model Transparency Act (H.R. 6881)[44] would direct the Federal Trade Commission (FTC) to establish standards for publicly disclosing information about the training data and algorithms used in AI foundation models.

While the Generative AI Copyright Disclosure Act (H.R. 7913)[45] would require notice to be submitted to the copyright owner prior to the release of a new generative AI system with regard to all copyrighted works used in building or altering the training dataset for that system.

The US' adoption of more comprehensive and widespread consumer protection laws aimed at regulating the use of AI is inevitable in next few years, but the current scope of the impending AI regulatory framework remains murky.

Meanwhile, when using data that imposes a disclosure obligation on its user for AI training purposes, there remains uncertainty as to, among others, (i) whether making the disclosure upon the distribution of the AI model will suffice, (ii) whether the disclosure obligation is also imposed in respect of the AI model's output and (iii) whether it can be interpreted that the disclosure obligation applies through a series of processes of using the AI model.

41 Apache License, Version 2.0: Terms and Conditions for Use, Reproduction, and Distribution, January 2004.

42 Article 20 (1) of the Personal Information Protection Act; Regulation (EU) 2016/679 §13, §14; California Civil Code Section 1798.100, 110 (California Consumer Privacy Act), etc.

43 SB-942 California AI Transparency Act ("require a covered provider to offer the user an option to include a manifest disclosure in image, video, or audio content, or content that is any combination thereof, created or altered by the covered provider's generative artificial intelligence (GenAI) system"); Tenessee's Ensuring Likeness Voice and Image Security (ELVIS) Act ("A person is liable to a civil action if the person publishes, performs, distributes, transmits, or otherwise makes available to the public an individual's voice or likeness, with knowledge that use of the voice or likeness was not authorized by the individual"); Utah Senate Bill 149 (SB 149) - Artificial Intelligence Policy Act (the AI Policy Act); Senate Bill 24-205 - Colorado Artificial Intelligence Act (hereinafter the CAIA).

44 H.R.6881 - AI Foundation Model Transparency Act of 2023.

45 H.R.7913 - Generative AI Copyright Disclosure Act of 2024.

For circumstances where such a disclosure obligation is likely to be imposed, we allocate scores for this assessment item as follows:

1 **5 points** where disclosure to data subjects is possible, irrespective of the existence of a disclosure obligation, because in this case, the likelihood of legal complications is minimal;

2 **4 points** where disclosure is currently impossible and, although no obligation currently exists, a disclosure obligation may be imposed due to subsequent legislative enactments or amendments while the relevant data is being stored or controlled; and

3 **2 points** where disclosure is impossible even though a disclosure obligation exists, because in this case, the likelihood of legal disputes becomes high.

Risks Related to Data Use Period and Territory

A. Assessment Criteria

Criteria	Weight
→ 2.1 Restrictions on data use period	7%
→ 2.2 Whether the data license is revocable	3%
→ 2.3 Restrictions on AI model service period	5%
→ 2.4 Restrictions on data use territory	4%

* A score in a range from 1 (high risk) to 5 (low risk) is assigned based on the assessment of the risk for each criterion.

B. Scoring Standard for Each Criterion

→ 2.1 Restrictions on data use period

Standard	Score
Data is usable perpetually.	5
Data is usable for a time period sufficient to avoid operational issues, or no explicit restrictions exist on the data use period.	4
Data use period has restrictions, but whether the period of the AI model service has any restrictions is unknown.	3
Data use period has already expired.	1

The data used by a company is classified as ❶ directly owned by the company, ❷ licensed from the rights holder, or ❸ open source. In the case of data licensed from the rights holder, the data use period is determined pursuant to the provision of license agreement.

The term of such a license agreement is, in most cases, explicitly stated, but sometimes it is not. Sometimes the term of a license agreement is set for a perpetual period, and such a contract may be found valid under the principle of autonomy of contracts unless otherwise specified.[46]

In the US, contracts without express duration language are usually terminable at the will of either party. Courts generally disfavor perpetual contracts and will likely conclude that a contract has a perpetual term only if the contract's unequivocal language necessitates such an interpretation.[47]

46 See Supreme Court Decision, 2023Da209045, decided June 1, 2023 (holding that a lease contract with a perpetual lease term is permissible).

47 See *Zimco Rests., Inc. v. Bartenders & Culinary Workers Union, Local 340*, 165 Cal. App. 2d 235, 238 (1958).

However, if a company is using the data to train an AI model, the question may arise whether the use of the AI model itself, which has been trained using the data, may be considered the "use" of the data.

Since the term "use" is construed relatively broadly, if the AI model is a type of model that generates output that directly copies the data it has learned, such generation of output may be deemed a use of data.

Above all, as the principle of autonomy of contracts requires that contracts be interpreted as intended by the parties to the contract, if the parties explicitly or implicitly intended that the use of the AI model is also the "use" of the data, the AI model may be unusable when the data use period elapses.

↓

In consideration of the above,

1 if a company owns the data or has a perpetual license agreement, the company is unlikely to face any legal risks in using the data, so a score of 5 was given;

2 if a company has set a data use period that is sufficient for the operation of the AI model, or if there is no explicit restriction on the data use period, it is unlikely that legal issues will arise in the near future – however, considering that the operation of the AI model may be prolonged or that legal disputes may arise on the grounds that there is an implied time restriction even if there is no such explicit restriction, a score of 4 was given;

3 if there is a restriction on the data use period but no explicit time restriction for the AI model, a score of 3 was given because even if the data may be used to train an AI model within the restricted time period, there is a possibility that the AI model itself is used after the data use period has elapsed; and

4 if the data use period has already passed, a score of 1 was given because it is equivalent to having no license.

2.2 Whether the data license is revocable

Standard	Score
Data license is irrevocable.	5
Data license's revocability is unknown.	4
Data license is revocable.	3

Continuous contracts are contracts that continue for a set period of time and therefore may stipulate specific reasons for termination. For example, a material breach of the contract, which destroys the trust between the parties that is the basis of the contract to such an extent that the contract cannot be expected to continue, can be a reason for termination.

Similarly, a license agreement, which is a type of continuous contract, may provide for termination or revocation of the license grant. If the agreement expressly states that the license grant is irrevocable, the rights holder of the data will not be able to revoke the license grant at a later date, and the licensee will not have any problems using the data.

However, if the agreement does not expressly state that the license grant is irrevocable, but only states the conditions for revocation, such as the default of one party or the non-compliance with other contractual terms (such as the obligation to pay periodic license fees or the obligation to indicate the rights holder), it may be difficult to determine whether the license grant can be revoked at a later date.

However, these conditions for revocation should not be particularly problematic if the company is in good faith in fulfilling certain contractual obligations.

In addition, if the rights holder of data has expressly indicated that it may revoke the data license at its sole discretion at a later date, it may be difficult to continue using the data, which may make it impossible to update the AI model, and it may be difficult to be certain of a judicial determination as to whether an AI model that has already been created will continue to be legally valid after the data license has been revoked.

In light of the above,

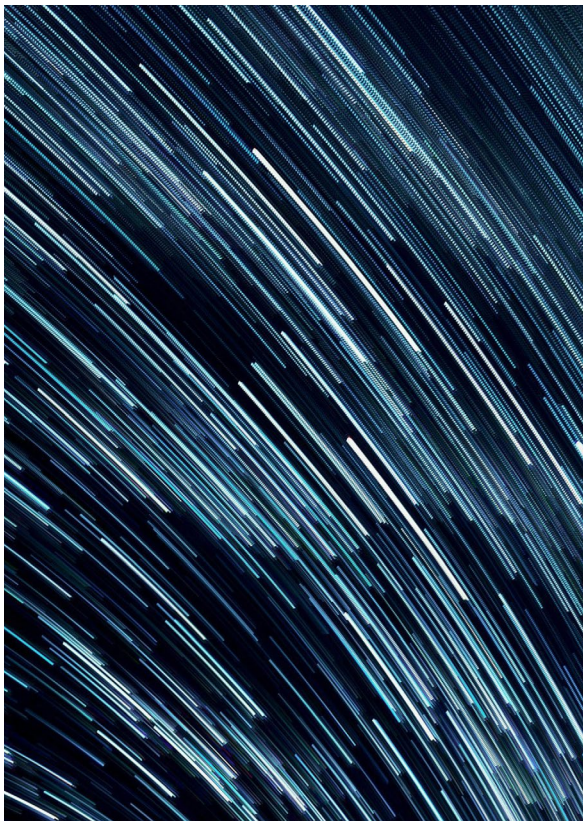
1. if the license grant is not explicitly revocable under the license agreement, the risk of *ex post facto* revocation does not need to be considered as described above, and therefore a score of 5 was given;

2. if the revocability of the license grant is not explicitly stated in the license agreement, a score of 4 was given because there is a possibility that the license grant may be revoked under certain conditions; and

3. if the licensor can arbitrarily withdraw the license grant without certain conditions, a score of 3 was given because it may be unclear whether the AI model can continue to exist effectively.

2.3 Restrictions on AI model service period

Standard	Score
AI model service can be provided perpetually.	5
The permitted duration of AI model service is unknown.	3
AI model service cannot be provided perpetually.	1



As mentioned above, there may be cases where AI model services cannot be provided in perpetuity due to contractual reasons, such as a fixed period for using data and AI models in a license agreement, a fixed period for transforming data, contractual restrictions on AI model services with third parties, or anticipated future enforcement of relevant laws.

Since the ultimate purpose of the AI model developed by a company is to pursue profit through the provision of services, it is an economic risk to have a fixed period of time for which the AI model can be provided.

In consideration of the above,

1. if the data is permanently available and irrevocable, and there are no restrictions on providing the AI model service perpetually, the above risk does not need to be considered, so a score of 5 was given;

2. if it is unclear whether the data is permanently available or irrevocable, or if the permitted duration of the AI model service is not known with certainty due to third-party contracts, such as the existence of termination conditions in contracts with third parties (such as co-developers) that are essential to the AI model service, there is a risk that the AI model service may have to be discontinued at a certain point in time, so a score of 3 was given; and

3. if there are conditions that prevent the AI model service from being provided perpetually, a score of 1 was given.

2.4 Restrictions on data use territory

Standard	Score
Data can be used worldwide.	5
Whether data use is geo-restricted is unknown.	4
Data can be used only in certain regions, including Korea, US, and EU.	2
Data can be used only in certain regions, not including Korea, US, and EU.	1

There may be certain legal or contractual restrictions on where data may be used in certain cases. Legally, certain types of data may be regulated by the laws of individual countries. For example, personal information that is available in Korea may not be available in the EU for reasons such as requiring data subject consent[48] or pseudonymization[49] under the General Data Protection Regulation (GDPR). Also, there may be specially regulated information due to cultural differences in certain countries.

In addition, in the case of copyrighted data, data that is not protected by copyright in one country may not be available in another country due to differences in the copyright laws of each country.

Furthermore, if the license agreement stipulates that the data can only be used in a certain country or region, the data cannot be used in other regions because it is not licensed.

If the data use region is limited, there are certain risks in the service process of AI models. Generally, data use occurs during the AI model training phase, but if the AI model is a type of model that outputs the training data as it is, there is a possibility that the act of generating output to be interpreted as data use, especially if the AI model is a model that is continuously updated with training data even after the service.

In addition, as AI models are often provided globally via the internet, there is a high risk of violating regional restrictions on data availability.

In consideration of the above,

1. if the data is available globally, there is no risk of violating a data use region restriction, so a score of 5 was given;

2. if the data use region restriction is not known with certainty, a score of 4 was given, as violation of the data use restriction may become problematic in certain cases; and

3. even if there is a data use region restriction, if the company will mainly provide AI training and services in Korea, U.S., and the EU, the company may be able to provide services by limiting the learning locations and service areas, so a score of 2 was given; and

4. if data is not available for use in Korea, U.S., or the EU, it will be extremely difficult to provide services unless it is an extremely localized AI model, so a score of 1 was given.

48 Article 6 of the GDPR.
49 Article 89 of the GDPR.

Risks Related to Personal Information and Data Security

A. Assessment Criteria

Criteria	Weight
→ 3.1 Whether personal data is included in AI training data	9%
→ 3.2 Whether data subjects have consented to the use of their data	3%
→ 3.3 Whether pseudonymized data is included in AI training data	3%
→ 3.4 Whether personal data may be entrusted or provided to third parties	5%
→ 3.5 Whether the scope of data users is limited	2%

* A score in a range from 1 (high risk) to 5 (low risk) is assigned based on the assessment of the risk for each criterion.

B. Scoring Standard for Each Criterion

→ 3.1 Whether personal data is included in AI training data

Standard	Score
AI training data includes no personal data or the company has a plan for anonymization or de-identification	5
AI training data includes personal data but the company has a pseudonymization plan, or it is unknown whether personal data is included	4
It is highly likely that AI training data includes personal data	2
AI training data explicitly includes personal data	1

Personal data means data about a living individual that identifies the individual who is the subject of the data, and covers a wide range of data types, from general identity data such as name and gender to more sensitive data about the person's private life, such as physical appearance or voice captured in videos or images, ideas, beliefs, and health.

As personal data fundamentally belongs to its subject, the data subjects should be able to securely manage their data and, in particular, have the right to be fully informed about any third-party handling of their personal data, including the details of such management.

Furthermore, data subjects should have the authority to oversee and influence how their data is managed by these third parties. These rights of data subjects are recognized globally as fundamental individual rights. For example, in the EU, Article 8 of the Charter of Fundamental Rights of the European Union (2010/C 83/02), enacted in the year 2000, provides for the right to protection of personal data.

In Korea, the Constitutional Court has also recognized the right to self-determination of personal data, i.e. the right of the data subject to decide when, to whom, and to what extent their information is to be disclosed and used, as a stand-alone fundamental right.[50]

50 Constitutional Court Decision 99Hunma513 dated May 26, 2005

Whereas the United States still lacks a federal omnibus consumer privacy law similar to the GDPR.[51]

In the absence of consumer-specific federal legislation, several sectoral laws have created a patchwork of privacy protections over the decades, such as the Family Educational Rights and Privacy Act (FERPA), the Children's Online Privacy Protection Act (COPPA), the Health Insurance Portability and Accountability Act (HIPAA), and even the Video Privacy Protection Act (VPPA), to name a few. In this splintered landscape, US states have been passing their own consumer privacy laws.

As of 2023, 12 states have passed consumer privacy regulations, though California's Consumer Privacy Act (CCPA) remains the most far-reaching.

For that reason, we will focus on the CCPA for discussion purposes. Sometimes dubbed California's version of the GDPR, the CCPA—together with its 2022 update, the California Privacy Rights Act (CPRA)—is arguably the most significant state-level effort so far to enact both stringent and broad consumer privacy protections.

A notable difference between California's privacy regime and other states is that California remains the only state to have created an enforcement agency (the California Privacy Protection Agency, or CPPA) with rulemaking authority, rather than delegating this function to the state's attorney general's office, as many such laws do. In practice, this may mean that the CPPA has more in-house expertise than most state attorneys general and latitude to both engage in proactive enforcement via published guidance and tackle complex and emergent issues at the intersection of AI and personal data.

In order to protect the rights of data subjects to their personal data, countries are working to establish a legal framework dedicated to the protection of personal data, including by setting forth the requirements that personal data controllers must comply to lawfully collect and use personal data from individuals and imposing various obligations that personal data controllers must fulfill throughout the course of controlling personal data.

Data used to train AI models includes a wide range of data types and it may also include personal data. If personal data goes into training AI, personal data controllers will need to consider a number of legal compliance issues, including whether the personal data was collected on lawful grounds, whether they have the right to control the personal data, to what extent they can use the personal data for AI learning, how they should securely store, use, and subsequently destroy the personal data, and how they should respond when data subjects request that their data be deleted or exercise other rights they are entitled to.

Particularly worthy of note for personal data controllers is that, even if personal data used in training AI are lawfully collected, it could be deemed an unauthorized divulgence or leakage and therefore a violation of data subjects' rights if such personal data appear in outputs generated by the AI models developed based on such data, or if third party users other than the data subjects become aware of the personal data. Therefore, whether personal data is included in AI training data, how the relevant data was obtained and how it is controlled are important factors in assessing legal risks that could arise in the course of using data.

51 King et al., Rethinking Privacy in the AI Era Policy Provocations for a Data-Centric World (Feb. 2024).

When AI learning data includes personal data, it is advisable to consider pseudonymizing or anonymizing (i.e., de-identifying[52]) the relevant personal data as a way to mitigate legal risks associated with data utilization. Pseudonymization reduces the risk of personal data leakage as pseudonymized data have no use in identifying individuals unless it is combined with additional data. Anonymized data are no longer treated as personal data as anonymization makes it no longer possible to identify individuals (Article 58-2 of the PIPA), which means that if personal data is anonymized before being used for AI learning, it may significantly reduce risks associated with personal data processing.

↓

In light of these considerations, we allocate points for this assessment item (i.e., whether personal data is included in AI training data) based on different risk levels as follows:

1 **5 points** where personal data is not included in AI training data, or even if it is, where the company has a plan for anonymization or de-identification, because in such cases, only anonymized data consequently goes into AI model training and there is little risk of legal issue related to personal data control.

2 **4 points** where personal data is included in AI training data but the company has a pseudonymization plan, or where it is unknown whether personal data is included, because in such cases, there is a relatively low risk of unauthorized leakage, etc. of personal data compared with when personal data that is not pseudonymized is explicitly

↓

included. It should be noted, however, that pseudonymized data, too, is essentially personal data and if rights have not been cleared before it was initially collected and provided to the company, there remains a certain degree of legal risk, given that it may be illegal to use it for AI learning and that we cannot rule out the possibility of the personal data and pseudonymized data that went into AI learning being exposed in AI-generated outputs at a later time (e.g., a prompt injection attack that prompts the AI to generate personal information from its output).

3 **2 points** where it is highly likely that AI training data includes personal data in light of the characteristics and method of collection of the data, because in such a case, there may be issues in respect of the legality of the collection and provision of the relevant data, the legality of its control by the company, the need for pseudonymization or de-identification, and the need for preventing the exposure of such data in outputs, thereby increasing the level of risk in terms of personal data protection.

4 **1 point** where AI training data explicitly includes personal data. In such a case, it is advisable to ascertain whether the personal data has been lawfully collected and provided to the company and whether there exist requirements that the company must satisfy or measures that it must take in order to control the data under lawful authority. If it is found that the personal data used in AI model training has not been lawfully collected (e.g., the company did not have the lawful authority to collect the data in the first place, or even if it did, the authority did not extend to the controlling of such data for AI training purposes), or that controlling the personal data involves excessive risks, the company may have to delete all personal data through filtering techniques and proceed with AI training with the remaining data.

52 "De-identification" is a broad concept that includes both pseudonymization and anonymization. However, in order to distinguish between pseudonymization and anonymization, this Tech Report assumes that de-identification is only limited to anonymization, which removes or modifies personally identifiable information.

3.2 Whether data subjects have consented to the use of their data

Standard	Score
Personal data is not included in AI training data, or it is included with consent from data subjects	5
Personal data is included in AI training data without consent from data subjects	1

Most countries' personal data protection laws require data subjects' consent for the controlling of personal data in principle, allowing data controllers to collect and use personal data only if the data subject has given explicit consent to such collection and use.

For example, the EU GDPR, while requiring the consent of data subjects in principle, allows the controlling of personal information without such consent where (i) control is necessary for the performance of a contract to which the data subject is a party or it is requested by the data subject for entering into a contract; (ii) control is necessary for compliance with a legal obligation to which the controller is subject; (iii) control is necessary in order to protect the vital interests of the data subject or another natural person; (vi) control is necessary for the performance of a task carried out in the public interest; and (v) control is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, where such interests override the interests or fundamental rights and freedoms of the data subject.[53]

Similar exception provisions exist in Korea's PIPA, allowing personal data control if there are reasonable and valid reasons, such as contractual necessity, legitimate interests of the personal information controller overriding the interests of the data subject, and use within a reasonable scope of the purpose of collection.[54]

It may also be possible to use AI training data that includes personal data lawfully without data subjects' consent by relying on the exception provisions provided in the law.

However, it may not be safe to expect that the events triggering such exceptions, such as the necessity for performing contracts with individual data subjects and the controller's pursuit of legitimate interests overriding the interests of data subjects, will be easily recognized, given that the means of data collection, such as crawling, may result in the unintended collection of large amounts of unspecified personal data and it will therefore be difficult to ascertain or guarantee on a case-by-case basis that all personal data is lawfully collected.

Given the foregoing, "whether data subjects' consent has been obtained," which is the clearest indicator of the legality of personal data control, may be a significant factor in determining whether personal data contained in AI training data is lawfully collected and whether it can be controlled and utilized.

53 Article 6 of the GDPR

54 Article 15 (Collection and Use of Personal Information)
(1) A personal information controller may collect personal information in any of the following cases, and use it within the scope of the purpose of collection:
1. Where consent is obtained from a data subject;
2. Where special provisions exist in other statutes or it is unavoidable due to obligations under statutes or regulations;
3. Where it is unavoidable for a public institution's performance of work under its jurisdiction as prescribed by statutes or regulations, etc.;
4. Where it is necessary to take measures at the request of a data subject in the course of performing a contract concluded with the data subject or concluding a contract;
5. Where it is deemed manifestly necessary for the protection, from imminent danger, of life, bodily and property interests of a data subject or a third party;
6. Where it is necessary to attain the legitimate interests of a personal information controller, which such interest is manifestly superior to the rights of the data subject. In such cases, processing shall be allowed only to the extent the processing is substantially related to the legitimate interests of the personal information controller and does not go beyond a reasonable scope.
7. Where it is urgently necessary for the public safety and security, public health, etc.

In light of these considerations,
we assess risks as follows based on whether personal data is included and whether data subjects' consent has been obtained:

1 **5 points** where personal data is not included in AI training data, or it is included but data subjects' consent has been obtained. If AI training data does not contain personal data, there is no risk related to personal data. Furthermore, even if personal data is included, using it to train AI models can be considered to be based on the explicit will of the data subject, as long as the personal data was collected with the data subject's lawful consent. In these circumstances, we believe that the likelihood of a legal issue relating to the control of personal data is very low.

2 **1 point** where personal data is included in AI training data, but data subjects' consent has not been obtained for its collection and control. In this case, using such personal data to train AI models is highly likely to be considered controlling personal data without lawful authority, unless it is recognized as an exception to the prohibition of personal data control without the data subjects' consent in accordance with laws of the relevant country.

(3) A personal information controller may use personal information without the consent of a data subject within the scope reasonably related to the initial purpose of the collection as prescribed by Presidential Decree, in consideration whether disadvantages have been caused to the data subject and whether necessary measures to ensure safety such as encryption have been taken.

3.3 Whether pseudonymized data is included in AI training data

Standard	Score
Pseudonymized data is not included, or it is included with consent from data subjects	5
Personal data is included but it will be pseudonymized	3
Pseudonymized data is included	1

Pseudonymization means deleting parts of personal data or replacing all or parts of the original datasets with an alias or pseudonym so that the personal data can no longer be attributed to the relevant individuals without additional data.

Thus, it is not possible to identify the data subject only by using pseudonymized data. In order to identify the data subject using pseudonymized data, one must undergo a recovery process, such as recombining the pseudonymized data with the key value used for pseudonymization, such as encryption and hashing processes. It should be noted that pseudonymized data is still personal data as it can be linked to specific individuals when combined with additional data, and therefore the separation of pseudonymized data from any additional data and the strict control on the possibility of such combination must be carried out with the same level of care as in the management of personal data.

The PIPA includes pseudonymized data in the scope of personal data, imposing the same protection obligation that applies to personal data in general with respect to the management of pseudonymized data.[55] The EU GDPR also treats pseudonymized data as one form of personal data, by defining pseudonymization as the "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to [...] natural persons."[56]

That, however, does not mean that pseudonymized data is not valuable as AI training data. Even though pseudonymized data is one kind of personal data, it can be useful if the need for data utilization and the need to protect personal data are reconciled appropriately, given that pseudonymized data cannot be traced back to specific individuals unless combined with additional data and pseudonymization reduces the risk associated with the control of personal data.

This view is reflected in the provisions of Korea's PIPA, which provides that "a personal data controller may control pseudonymized information without the consent of data subjects for statistical purposes, scientific research purposes, and archiving purposes in the public interest, etc.,"[57] relaxing the strict consent requirements that it imposes on personal data controllers to some extent.

In contrast, under the GDPR, there is no regulation concerning the data subject's consent for pseudonymization.

However, the GDPR considers that "further processing of personal data for purposes of public interest archiving, scientific or historical research, or statistical purposes can be deemed compatible with the original purpose."[58]

55 Article 2, Subparagraph 1, Item C of the PIPA.
56 Article 4 of the GDPR.
For the purposes of this Regulation:
(5) 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.
57 Article 28-2(1) of the PIPA.
58 Article 5(1)(b) of the GDPR.

Therefore, further processing of personal data for public interest archiving, scientific research, or statistical purposes is regarded as compatible with the original purpose without requiring additional consent from the data subject. This can be interpreted to mean that pseudonymization can be performed without separate consent from the data subject.

In light of these considerations,
we assess the risks associated with controlling pseudonymized data and allocate points as follows, considering both the fundamental nature of pseudonymized data as personal data and the fact that controlling pseudonymized data involves a somewhat lower risk of violation of the rights of data subjects compared with controlling personal data in general:

1 **5 points** where pseudonymized data is not included in AI training data, because in such a case, there cannot be any issue concerning the controlling of pseudonymized data. Also 5 points where pseudonymized data is included in AI training data but with data subjects' consent, because it is highly likely that it can be used lawfully based on the explicit will of the data subject as long as the consent has been obtained.

2 **3 points** where AI training data includes lawfully collected personal data, and the company plans to take further steps to pseudonymize it. In this case, pseudonymization will make the control of this data even safer and not create any additional risk when the personal data may already be lawfully utilized. However, it will still be necessary to ensure that pseudonymization is strictly applied and that the use of pseudonymized data is safe and lawful.

3 **1 point** where pseudonymized data is explicitly included in AI training data. In such case, it is only possible to use pseudonymized data for training purposes if it is ascertained that the data was collected lawfully and the company secures lawful basis to control the data while complying with all obligations to ensure that the control will be safe.

3.4 Whether personal data may be entrusted or provided to third parties

Standard	Score
The company has a right of entrustment or provision of personal data to third parties	5
It is unknown whether entrustment or provision of personal data to third parties is permitted or restricted	3
Entrustment or provision of personal data to third parties is explicitly prohibited	1

Providing personal data to third parties for processing by such third parties, not by the ones who collected it in the first place, creates an additional risk in the context of personal data protection.

Therefore, personal data protection laws in various countries categorize such behavior (i.e., providing personal data to third parties) into different types and distinguish them from "collection." In Korea, the PIPA divides the act of sharing personal data with third parties into (i) provision to third parties[59] and (ii) entrustment of processing work.[60] In the former case, the provision is made for the benefit of the recipient and, as a result, control and management of the personal data is transferred to a third party; in the latter case, the processing of personal data for business purposes is entrusted to a third party for the benefit of the trustor, within the scope of the trustor's business.

While the GDPR does not clearly distinguish between provision in general and entrustment, it categorizes a personal data handler into two types, "controller" and "processor," and defines "processor" as a person or entity that processes personal data on behalf of the controller, conceptualizing an act equivalent to the entrustment of personal data.

Generally, in the case of third-party provision, the responsibility to protect and manage personal data rests with the recipient once it has been provided, based on the fulfilment of the legal requirements for the provision (including obtaining the consent of the data subjects).[61]

On the other hand, in the case of the entrustment of processing work, the controller remains responsible for overall supervision and management of the entrusted personal data.

We assess the risks associated with the provision and entrustment of personal data to third parties as set forth in (1) – (3) below, taking into account the different legal responsibilities and consequent risks borne by personal data controllers depending on how personal data is transferred to and processed by third parties.

It should be noted that we assessed these risks based on the assumption that personal data can always be included in AI training data. If it is certain that no personal data is included in the training data and there is no possibility of such inclusion, entrustment or provision of data to third parties will not give rise to any personal data control risks.

Even that, however, does not completely eliminate risks concerning (i) whether the company is entitled to entrust or provide personal data under relevant contracts; (ii) whether the third parties will lawfully use and safely manage such data and to what extent the company may be held responsible for the supervision and management of the third parties' activities; and (iii) whether the means of transferring and transmitting for the entrustment or provision of data are appropriate.

59 Article 17 of the PIPA.
60 Article 26 of the PIPA.
61 For reference, the GDPR provides that rights and responsibilities may be distributed between the provider and the recipient as co-controllers.

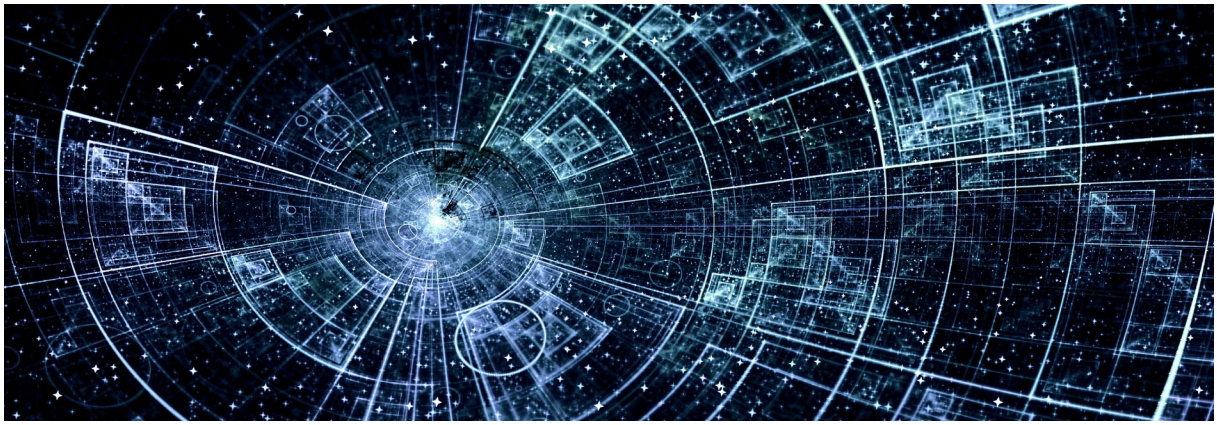
With this in mind, we believe that the following risk assessment system applies both when personal data is included and when it is not.

1 **5 points** where the company has been granted the right to entrust or provide personal data to third parties in the course of acquiring such data, or where such entrustment or provision is not restricted, because in this case, the company is highly likely to be able to lawfully entrust or provide the data based on the above right. In particular, in the case of personal data already in the public domain, it may be presumed that the data subject has at the time of disclosure given a certain level of consent to the controlling of their personal data, including its collection and provision to a third party, and therefore, it is not necessary to obtain separate consent from the data subject for the provision and entrustment of the data to a third party.[62]

2 **3 points** where it is unknown whether entrustment or provision of personal data to third parties is permitted or restricted. In this case, the company is advised to ascertain that it is in fact permitted to control the relevant data in such a manner, and on the basis of which provisions of the applicable laws it may entrust or provide data to third parties.

3 **1 point** where entrustment or provision of personal data to third parties is explicitly prohibited, because in this case, only the company is entitled to control the relevant data in principle, and any involvement of third parties may be considered illegal or violation of contract, creating a high legal risk.

62 Supreme Court Decision 2014Da235080 dated August 17, 2016.



3.5 Whether the scope of data users is limited

Standard	Score
No limitation is placed on the scope of users authorized to use data	5
Limitation is placed on the scope of users authorized to use data	3

If only certain users are authorized to use AI training data, access and control of such data by unauthorized users may be considered a breach of relevant contracts. It is also advisable to consider the possibility that the scope of the authority granted depends on the agreed purpose of control. In this case, if the scope of the agreed purpose effectively limits the use of the data for AI learning or places a specific limit on the scope of use, the company may have limits on using the relevant data in AI model training.

If personal data is contained in AI training data, it further increases the possibility that the use of relevant data will be restricted as well as the need to place such restriction. Granting different levels of authorization to access and use personal data in the course of its control, as well as restricting access or use by unauthorized persons is an essential means of ensuring the security of controlling personal data, and some countries have statutory provisions regulating such restriction and control.

Therefore, when personal data is included in AI training data, it may be important for the company to ascertain whether there is any limitation on the scope of users with the authority to control such data under the applicable law, even if there is no explicit contractual limitation.

The GDPR provides that the processing of personal data shall be adequate, relevant and limited to what is necessary[63] and that personal data shall be processed in a manner that ensures appropriate security of the personal data,[64] implying the need to limit the scope of controlling personal data and the scope of controllers. In Korea, the PIPA sets forth the minimum necessary principles that should be observed in controlling personal data[65] and specifically requires data controllers to restrict access authority to personal data and effectively manage access to personal data, as part of measures to ensure the security of personal data.[66]

63 Article 5(1)(c) of the GDPR.
64 Article 5(1)(f) of the GDPR.
Personal data shall be:
(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')
(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').
65 Article 5 of the PIPA.
66 Article 29 of the PIPA; Article 30(1), Subparagraphs 2 and 3 of the Enforcement Decree to the PIPA.

In light of the above,
in order to use personal data for training AI models, it is necessary to look into whether the rights to use such data are limited to certain users to ensure that it is accessed and used within the lawful scope provided in the relevant contracts and laws.

Therefore, we assess the risks of using personal data based on whether the scope of authorized users is limited, as follows:

1 **5 points** where the scope of authorized users is not limited, because in this case, there will be no need to restrict access authority and users will be able to freely access and control the relevant data within the scope of due purpose on a need-to-access basis.

2 **3 points** where the scope of authorized users is limited under the relevant contracts or law, because in this case, the company will be required to grant different levels of access authority in accordance with the details of such limitation, and also take measures to effectively restrict access to data.

Additional Legal Risk

A. Assessment Criteria

Criteria	Weight
→ 4.1 Risks in the data collection process	8%
→ 4.2 Known disputes involving the use of same dataset in AI models	10%
→ 4.3 Other contract risks associated with licenses	5%
→ 4.4 Type of license terms	-

* A score in a range from 1 (high risk) to 5 (low risk) is assigned based on the assessment of the risk for each criterion.

B. Scoring Standard for Each Criterion

→ 4.1 Risks in the data collection process

Standard	Score
There is no particular issue in data collection process.	5
Data was collected through web crawling, etc.	3
Method of data collection is unknown.	2
Data was collected through bypassing robots.txt or questionable methods.	1

Data collection involves a process of acquiring publicly available information, both online and offline. As AI training requires large datasets, it is common to gather information from publicly available online sources, often using automated methods like web crawling.

In some cases, however, the legality of data collection may be at issue. While it may not be an issue for data where no rights holder exists (e.g., weather data or national statistics), where rights holders do exist or appear to exist, there may be certain restrictions on data collection, even if the data is fully publicly available. For example, the Korean Personal Information Protection Act ("PIPA") sets forth the specific circumstances under which personal information can be collected, such as with the consent of the data subject or as otherwise provided by the law,[67] and even then, only the minimum amount of personal information necessary for the purpose may be collected.[68] The EU GDPR also requires that the collection of personal data must have a clear purpose, be minimal in scope,[69] require the consent of the data subject in principle,[70] provide certain information to the data subject,[71] and prohibit the processing of certain sensitive data.[72]

Also, as mentioned above, web crawling is often used for data collection. Web crawling involves the extensive and systematic collection of data from a website using automated programs. While web crawling may set conditions on the scope of the collection or the type of data collected, the indiscriminate nature of web crawling makes it difficult to assess the legality of each individual piece of data. This leaves open the possibility that the data collected may be protected under various laws as intellectual property, such as copyrighted work or database.

67 PIPA, art. 15(1).
68 PIPA, art. 16(1).
69 GDPR, art. 5(1).
70 GDPR, art. 6(1)(a).
71 GDPR, art. 13.
72 GDPR, art. 9.

Furthermore, there are cases where the act of web crawling may be found illegal. For example, in South Korea, web crawling has been found to constitute infringement of the rights of database creators in one case,[73] and unfair competition under the Unfair Competition Act in another case where civil liability was imposed.[74]

The legal challenges to web crawling in the US often rely on Computer Fraud and Abuse Act (CFAA), which makes it illegal to intentionally access a protected computer without authorization or by exceeding authorized access. In *HiQ v. LinkedIn*,[75] the Ninth Circuit ruled that HiQ did not violate the CFAA by scraping publicly available information from LinkedIn. However, the court made it clear that its ruling was limited to the CFAA and did not address potential legal claims under other laws, such as trespass to chattels, copyright infringement, misappropriation, unjust enrichment, conversion, breach of contract, or privacy violations.

Courts in the U.S. have largely dismissed claims such as misappropriation, unjust enrichment, and privacy violations when the rights asserted are equivalent to those protected under copyright law.[76] This reflects a broader trend of narrowing the legal remedies available against web scraping to avoid duplicative or inconsistent rulings that might undermine the federal copyright framework. By doing so, courts signal that disputes over web crawling or AI training with scraped content must largely be resolved within the contours of copyright law, leaving limited room for state-law or non-copyright federal claims.

In the EU, web crawling can violate GDPR, contractual terms, or database rights. In *Ryanair Ltd. v. PR Aviation BV*,[77] PR Aviation scraped flight data from Ryanair's website to provide ticket comparison services. The Court of Justice of the European Union ruled that Ryanair could not rely on copyright or database rights to prevent the scraping of flight data because the website was not protected by the EU Database Directive. However, as Ryanair's terms of use prohibited scraping, the court left open the possibility of enforcing anti-scraping policies through contract law.

Meanwhile, to prevent issues with web crawling, websites often specify what they will and will not allow to be crawled in the form of a text file called robots.txt. Although robots.txt is not legally binding, it serves as an explicit opt-out by the website administrator from crawling. If a web crawler collects legally protected data despite robots.txt restrictions, the fact that the data was gathered against this stated restriction could be used against the collecting company in negotiations with the rights holder or in court proceedings. Additionally, the company may face moral criticism for disregarding these guidelines.

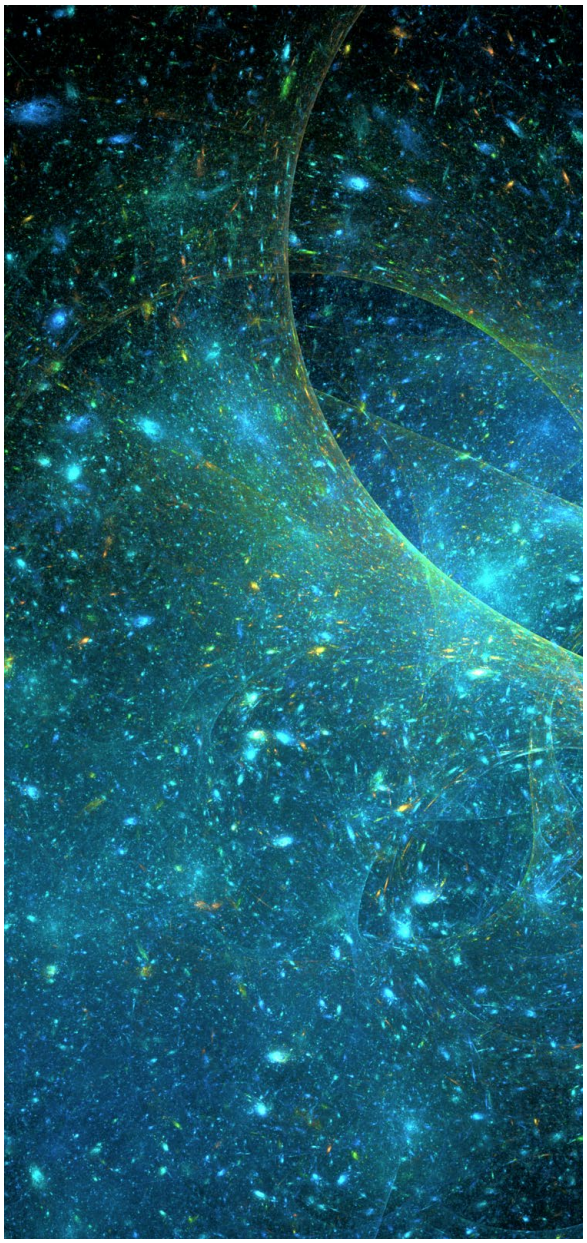
73 Seoul High Court, Case No. 2016Na2019365, decided April 6, 2017

74 Seoul High Court, Case No. 2021NA2034740, decided August 25, 2022.

75 *HiQ Labs, Inc. v. Linkedin Corp.*, 938 F.3d 985 (9th Cir. 2019).

76 17 U.S. Code § 301 - Preemption with respect to other laws ("no person is entitled to any such right or equivalent right in any work of visual art under the common law or statutes of any State"); *Andersen v. Stability AI Ltd.*, No. 23-cv-00201-WHO, 2024 U.S. Dist. LEXIS 143204 (N.D. Cal. Aug. 12, 2024); *Doe v. GitHub, Inc.*, No. 22-cv-06823-JST, 2024 U.S. Dist. LEXIS 11068 (N.D. Cal. Jan. 3, 2024); *J.L. v. Alphabet Inc.*, No. 23-cv-03440-AMO, 2024 U.S. Dist. LEXIS 118100 (N.D. Cal. June 6, 2024).

77 *Ryanair Ltd v. PR Aviation BV* (C-30/14) EU:C:2015:10 (15 January 2015).



In consideration of the above,

1 if the data was collected legally, such as by obtaining permission to use the data or by generating raw data directly, the legality of the data collection process is unlikely to be challenged and thus a score of 5 was given;

2 if the data was collected through web crawling, a score of 3 was given in view of the risks inherent in web crawling, such as the difficulty in fully reviewing the legality of individual data;

3 if the method of data collection is unknown, a score of 2 was given, considering the high risk that legal disputes may arise if the original collector of the data gathered it unlawfully; and

4 if robots.txt was bypassed, a score of 1 was given, considering that it would be difficult to escape not only moral criticism to start with, but furthermore potential for legal disputes. Similarly, a score of 1 was given if the data was acquired through improper methods such as from a shadow library, since it is likely to be subject to legal disputes unless an *ex post facto* agreement can be reached with the rights holder who has sufficiently expressed his/her objection.

4.2 Known disputes involving the use of the same dataset in AI models

Standard	Score
There is no known dispute involving the use of same dataset in AI models.	5
Data disputes do exist, but they are minor disputes.	4
Disputes exist where a considerable level of damages is being sought.	3
Disputes exist where an even greater level of damages is being sought.	1

Even if data was acquired upon sufficient review of its legality, disputes may have arisen over AI models trained using the same data or database or data acquired from the same licensor.

Such disputes can arise if a company misjudges the legality of the data, misidentifies the true rights holder, or if the data exists in a legal gray area due to the absence of relevant laws or precedents on AI models. In some cases, disputes may also occur because the disputant is particularly aggressive in pursuing claims.

In any case, if the same data or database is used for a new AI model, there is a risk of triggering similar disputes. Even if the data is legally acquired, the possibility of a dispute itself poses a risk to the service of an AI model.

In light of the above,

1. if there is no known data dispute, there is no such risk that should be considered in advance, so a score of 5 was given;

2. even if there are known data disputes, if such disputes are minor ones involving individual plaintiffs, the risk exposure due to a loss in a similar dispute is small given also the prior legal clearance review of the data; therefore, a score of 4 was given in such cases;

3. if a dispute suggests a considerable claim, it likely reflects that the dataset in question is substantial in volume, holds significant importance, or that the disputing party is assertive in pursuing compensation through litigation; therefore, a score of 3 was given; and

4. if a dispute suggests an even more significant claim, a score of 1 was given, as the potential for extensive disputes is high for similar reasons outlined above.

4.3 Other contract risks associated with license agreements

Standard	Score
There is no known additional contract risk.	5
The contract sets forth stringent obligations regarding data management (e.g., data security, confidentiality).	4
Unlimited liability, possibility of unrestricted audits, and other similar contract risks exist.	3

License agreements may impose various obligations on the licensee in addition to the payment of royalties. For example, the licensee may be obligated to maintain confidentiality regarding the entire agreement, or may be subject to data security obligations such as data storage or post-use destruction procedures.

The more complex and demanding the license agreement, the more likely it is that (i) unintended non-performance of contractual obligations may occur due to inadequate management of human and material resources, (ii) the data may not be available for use at the desired time and in the desired form, and (iii) even if all obligations are complied with, there is a higher probability of the licensor bring some legal claim.

Above all, if the liability for data use is entirely borne by the licensee (e.g., if the licensee must indemnify the licensor against any third party claims of unlawful use of the data), there is a risk that the licensee may assume liability even for those faults attributable to the licensor.

In addition, if the licensor has the unrestricted right to conduct audits on licensee's data use and management in general, the risk of legal disputes is high, as even minor deficiencies in the implementation of data management procedure can be easily exposed during the audit process.

↓

Given the above,

(1) if there was no known additional contractual risk, a score of 5 was given;

(2) if the contract required stringent data management structure or otherwise set forth complex and demanding obligations, there is the possibility of unintended non-performance of obligations due to mismanagement as discussed above, and thus a score of 4 was given;

(3) if the licensee assumed unlimited liability or the licensor had the unrestricted right for audits, the risk of legal dispute increases, along with a higher likelihood of losing the dispute or facing significant damages in the event the case is lost, and thus, a score of 3 was given.

4.4 Types of license terms

Cat.	Standard
1	Data can be freely **distributed, used, modified, combined, used to create derivative works, etc.** without any restrictions
2	Data can be **distributed, used, modified, combined, used to create derivative works, etc.** only under certain conditions
	① Obligation to attribute authorship, source, copyright, license, etc.
	② Obligation to notify modifications
	③ Requirement to obtain author's permission when creating larger work
	④ Obligation to apply same license terms to all derivative works, or obligation not to impose additional terms when distributing derivative works
	⑤ Sub-licensable (but not distributable) dataset
3	Data can be used but cannot be **distributed, used, modified, combined, used to create derivative works, etc.** (*This includes cases where the condition for using data is feasible to comply with, but the conditions for modifying, creating larger work, or distributing, combining, creating derivative works, etc. are impossible or practically impossible to comply with.)

* This is not the criteria for assigning compliance risk assessment scores, but rather for categorizing the type of license terms for datasets.

(1) Background of License Type Classification

The assessment criteria presented in the foregoing sections are based on the assumption that assessment will be conducted on a single dataset or a small class of dataset; however, unlike such theoretical cases, most datasets collected, stored, and used in practice are multi-layered datasets which are created by combining multiple datasets to form a larger dataset (horizontal combining) that is repeatedly increased in size (vertical combining) over several times. The final large dataset formed from such a process is called a mother set.

There are various reasons for forming a mother set, such as (i) to collect data that is suitable for the purpose or performance goals of the AI model, (ii) to prevent data bias, and (iii) to supplement the data collected by individual licensors as they cannot build a large enough dataset by themselves.

However, in this process, the licenses applicable to the individual datasets ("**subsets**") comprised in the mother set may differ from each other, and the license applicable to the mother set formed by combining the subsets may also differ from the licenses of the subsets. For example, subset A may be subject to license X, subset B may be subject to license Y, and mother set C may be subject to license Z.

In such a situation, (i) if the terms and conditions are stricter or the scope of use is narrower under a particular subset's license, the mother set may be required to meet those stricter terms and conditions or the narrower scope of use, regardless of the terms and conditions or the scope of use under the licenses of other subsets, and (ii) if, among other things, the terms and under each subset's license are different and cannot be met simultaneously, resulting in a "conflict" between the licenses, the use of the mother set itself may become problematic.

Therefore, by classifying in advance the licenses for each "**unit dataset**,"[78] which is a unique subset that is not combined with other subsets comprised in the mother set, we aim to determine the terms and conditions and the scope of use applicable for the mother set, and furthermore, to identify cases where conflicts between licenses occur, thereby preventing issues that may arise in using the mother set. In the following license type classification criteria, the "dataset" covered by the license refers to the "unit dataset" even if not expressly specified.

78 A unit dataset refers to a subset of a single unit that is either created directly by AI Research Institute, provided by a third party, or collected externally, regardless of the size of the dataset. For example, if subset A and subset B are combined to form subset X, subset X and subset C are combined to form subset Y, and subset Y and subset D are combined to form the mother set, then subsets A·B·C·D each correspond to a unit dataset. However, as a unit dataset is a subset of a single unit, a single license will be applied to each unit dataset.

The reason for creating a classification system based on this unit dataset is that there are cases where a new mother set is constructed by utilizing existing unit datasets, and as such, the risk of conflict or the scope of use in the mother set can be assessed based on the classification results of the unit datasets.

(2) Classification of License Types

The licenses for a dataset can be categorized into three main types.

Type 1 First, the license is unrestricted, which means that there are no conditions or requirements for using the dataset under the license. Distribution, modification, combination, creation of derivative works, etc. (hereinafter referred to as "**distribution, etc.**") of the dataset are all permitted in addition to use.

Type 2 Next, there is a type of license that allows the use, distribution, etc. of the dataset, but requires certain conditions to be met in the process (e.g., attribution obligation). In this case, the conditions for use, distribution, etc. required by each license (hereinafter referred to as "**license conditions**") may be different, so to determine whether there is a possible conflict between the licenses and whether the mother set meets the license conditions, the license conditions under each license must be assessed individually. It is therefore necessary to further categorize these Type 2 licenses again by grouping them according to similar license conditions, as discussed further below.

Type 3 The third type of license allows the use of dataset but not its distribution, etc. To effectively utilize a dataset, it is often necessary to provide it to a partner, modify some parts, or, most importantly, use the dataset to create a higher-level dataset (i.e., a larger work). However, some licenses prohibit all such activities, including distribution, etc., and are therefore classified separately as Type 3 licenses. In addition, licenses that nominally permit distribution but have conditions that are impractical to fulfill are also classified as Type 3 licenses.

(3) Detailed License Terms (Sub-categories of Type 2 licenses)

Type 2 licenses are further classified into sub-categories based on similar license conditions. In general, a license will often contain multiple license conditions. In such cases, the sub-categories associated with all such license conditions will be shown in the notation. For example, a dataset licensed with both a Type 2-① "Attribution Obligation" and a Type Type 2-② "Change Notification Obligation" license conditions would be classified as a Type 2-①,②. This notation system is chosen because the sub-categories are in parallel relationship with one another, with none encompassing the other; and thus, it is important to show all sub-categories in the notation to ensure that the license conditions for individual datasets are fully evaluated without omission when assessing potential conflict between licenses or when using the mother set thereafter.

Note that some of the open source licenses described below may explicitly state that the license applies only to "source code" in principle (e.g., GPL). However, given the licensor's intent in applying the license to the dataset, it is reasonable to interpret the dataset as being equivalent to, or substituting, the "source code" in the original license, at least in the context of the use and distribution, etc. of the dataset. In other words, it would be reasonable to interpret the dataset as covered by the license, even if the original license provides that works other than the dataset are covered by the license. The official GNU website appears to adopt a similar interpretation.[79]

79 "You can apply the GPL to any kind of work, as long as it is clear what constitutes the 'source code' for the work. The GPL defines this as the preferred form of the work for making changes in it." Free Software Foundation, *Frequently Asked Questions About the GNU Licenses*, https://www.gnu.org/licenses/gpl-faq.en.html#GPLOtherThanSoftware.

Type 2-① Attribution Obligation

Type 2-① relates to license terms that require the licensee to credit the original author, link the source, include copyright notices or the text of license, etc. (hereinafter "**attribution**") when distributing the dataset.

As effectively all open source licenses include such attribution obligation, all open source licenses can be considered to fall under Type 2-①. For example, the MIT license, which is considered to be the least restrictive of the open source licenses, requires a licensee to include copyright notices and a copy of the license in any distribution. The BSD license requires a licensee to credit the author and provide license texts in any distribution, and the CC-BY license likewise requires a licensee to attribute authorship.

From a company's perspective, this type of license condition is relatively easy to comply with unless there are special circumstances. To comply with the terms of this type of license, for instance, it is sufficient to provide attribution when distributing the dataset, and there is no obligation to notify anyone if the dataset is used internally. Specifically, none of the popular open source licenses imposes a separate notice and attribution obligation for the use of dataset. Even if a commercial license did impose such an obligation, the license terms would be classified as either Other or Type 2-④. Therefore, Type 2-① would include only licenses that impose an attribution obligation "for distribution, etc. of the original dataset." In other words, if there is no distribution, etc., the license conditions of Type 2-① do not apply, and in case of any distribution, etc., the license conditions can be easily complied with, so Type 2-① is not particularly problematic for a company.

* Examples of Type 2-① licenses: MIT, BSD, CC-BY (except SA), and most other open source licenses

Type 2-② Modification Notification Obligation

Type 2-② relates to license terms that require licensees to disclose the fact that the dataset has been modified when distributing the modified dataset. In some cases, Type 2-② licenses may also require licensees to disclose, in addition to the fact that the dataset has been modified, the date and time the modification was made, as well as the content of the modification.

In addition, licenses of this type often include a license term requiring that the modified source code be provided or distributed with an undertaking to make it available upon request. In the case of software, the "obligation to disclose modified source code" makes sense because there are ways to distribute modified software without the disclosure of the modified source code. In the case of datasets, however, the "obligation to disclose modified datasets" is not so meaningful because there is no way to distribute modified datasets without disclosing the modified data in the dataset. Therefore, the obligation to disclose modified dataset, which is most commonly included in Type 2-② licenses, is not particularly considered for this type.

If a company modifies a dataset covered by a Type 2-② license, it is relatively easy to comply with the license terms because the company is only required to disclose the fact that the modification was made, the date and time of the modification, and the content of such modification, when distributing the dataset. As there is no separate obligation to notify or seek permission from the author, etc. to use the modified dataset only for internal purposes, it is relatively easy to comply with the license terms of this type.

However, if a company collects or receives a licensed dataset under a Type 2-② license from an external source, it may be difficult to determine whether the license condition has been already breached by failing to provide adequate notice of the modification for the dataset that had been modified. In the case of Type 2-①, as most licenses fall into this category, if a particular

dataset does not have a notice of authorship, etc., it may be reasonable to immediately suspect a breach of the license condition; however, in the case of Type 2-②, as a notice is required only when the dataset has been modified from the original dataset, the absence of a notice regarding modification in a particular dataset does not immediately indicate a breach of the license condition. Nevertheless, as it is often difficult for a company to check the full history of externally provided or collected datasets, it is worth considering that there is always an inherent risk of violating the license condition when using datasets under a Type 2-② license.

* Examples of Type 2-② license: Apache-2.0, GPLv2.0, GPLv3.0, AGPLv3.0

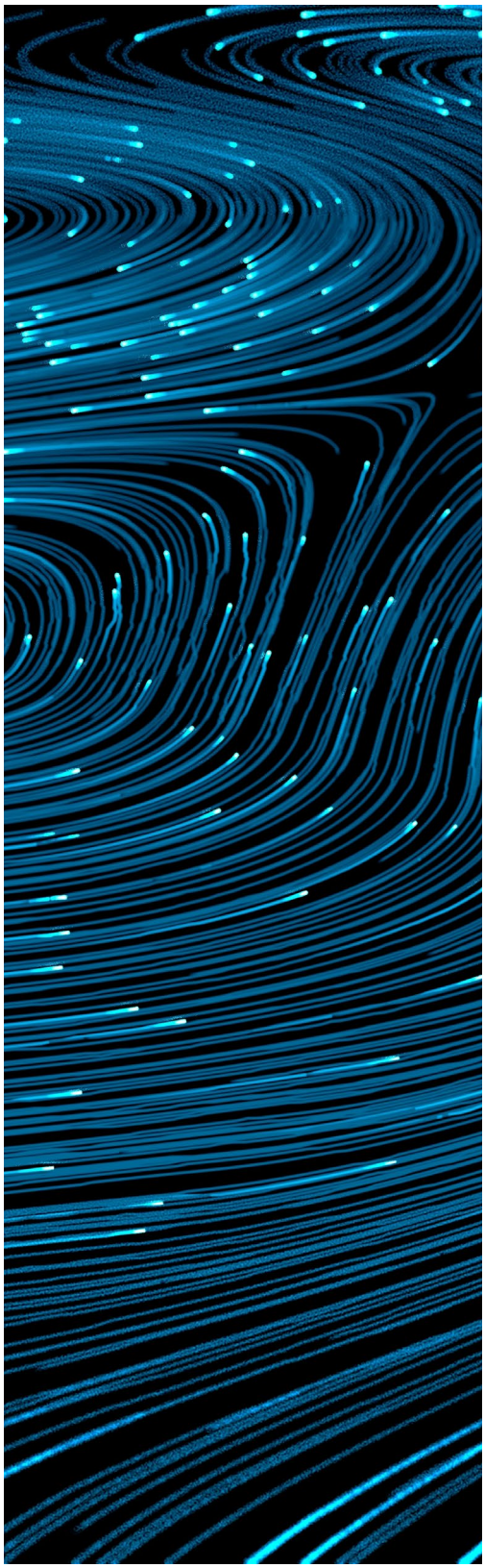
Type 2-③ Requirement for Author's Permission in Creating Larger Work

Larger work is a concept first used in the Mozilla Public License (MPL) 1.1, which refers to the result of combining a licensed work with an unlicensed work. Larger work is used in the MPL to refer to a combination of source code, but in this report it refers to a larger dataset created by combining differently licensed datasets.

Type 2-③ requires permission from the author of the original dataset to create a larger work (i.e., the mother set) by combining a dataset under that type of license with a dataset that is not under such type of licenses.

However, since (i) the author's identity or contact information may be difficult to obtain, and (ii) the author may not give permission to create the larger work or may impose unreasonable conditions as a prerequisite for such permission, a dataset with this type of license may be of limited usefulness in forming the mother set.

* Examples of Type 2-③ licenses: GPLv2.0, LGPLv2.1



Type 2-④ Obligation to Apply Same License Terms to All Derivative Works

Derivative work,[80] a concept that first appeared in GPLv2.0 and evolved into the term "covered work"[81] in GPLv3.0, refers to a program made the original source code under a license or the work derived from such a program. In this report, the term "derivative work" in the context of datasets means a dataset created (by modification, combination, or otherwise) based on a licensed dataset. Derivative work may be a concept broader than larger work.

Type 2-④ means that, when creating a derivative work of a dataset covered by a license of this type, a license that is the same as, or at least no more onerous than,[82] the original data set's license must be applied. This includes when making modifications to the original dataset, or when creating a larger work that combines the original dataset with other datasets not covered by that license.

In general, it is possible for a mother set to comply with the license terms of different datasets at the same time, as well as for one attribute of a mother set (e.g., unlimited distribution right) to comply with the license terms of different datasets simultaneously.

Therefore, although a license condition that requires all licenses applied to a mother set to satisfy a specific condition is usually not a problem, a license condition under this Type 2-④ can be problematic because it is likely to cause conflicts between licenses applied to datasets.

For example, if a dataset Z is created by combining a dataset X under one license and a dataset Y under another license, and there is no license condition of this Type 2-④, the dataset Z can be split into datasets X and Y and the license conditions applicable to each dataset will be complied with.

However, suppose the dataset Z is licensed under a GPLv2.0 license of this Type 2-④, and the dataset Y is licensed under an Apache-2.0 license, then the GPLv2.0 must apply to the entire dataset Z, resulting in the application of both GPLv2.0 and Apache-2.0 licenses for the dataset Y portion of the dataset Z.

However, because GPLv2.0 does not allow the addition of restrictions while Apache-2.0 does, and because there are conflicting terms between the two licenses (e.g., whether or not there are restrictions relating to patents), the two licenses are incompatible and cannot coexist. In other words, having a Type 2-④ license may result in a definitive conflict between licenses.

Therefore, if you have a dataset that is licensed under this Type 2-④ license, it is important to more thoroughly review the potential for conflict between the licenses.

* Examples of Type 2-④ licenses: GPLv2.0, GPLv3.0, CC-BY-SA

80 "TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION ... a 'work based on the Program' means either the Program or any derivative work under copyright law... ." GPLv2.0, *available at* https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html.

81 "0 .Definitions. ... A 'covered work' means either the unmodified Program or a work based on the Program." GPLv3.0, *available at* https://www.gnu.org/licenses/gpl-3.0.html.

82 In other words, it is a license that does not impose any additional obligations on the licensee beyond those specified in the original dataset's license, and does not arbitrarily restrict or remove the rights stated in the original license.

Type 2-⑤ Mere Sub-licensing Right

There may be cases where a company may need to provide datasets to third parties (e.g., partnering company, affiliates), in addition to using them internally. However, if provision of the dataset to third parties is prohibited, or permitted only for certain third parties (such as data processors), or subject to certain conditions or obligations (hereinafter simply referred to as "**right to distribute**"), the terms of license may be violated if the company overlooks these conditions and provides the dataset to third parties.

To determine the terms and conditions applicable for distribution of a mother set, it may not be sufficient to simply review the text of the mother set's license. It is important to review the licenses of all the individual datasets comprised in the mother set to determine the actual scope of rights permitted for distribution of the mother set.

In practice, however, it is often difficult to understand the scope of right by simply looking at the language of the license. For instance, it appears that the terms "distributable" and "sub-licensable" are used interchangeably to refer to the concept of "distribution." According to the Open Source Initiative's (OSI) definition of open source, the term "distributable" or "distribution" appears to refer to the copyright law concept of distribution, which allows for unrestricted and perpetual redistribution.[83]

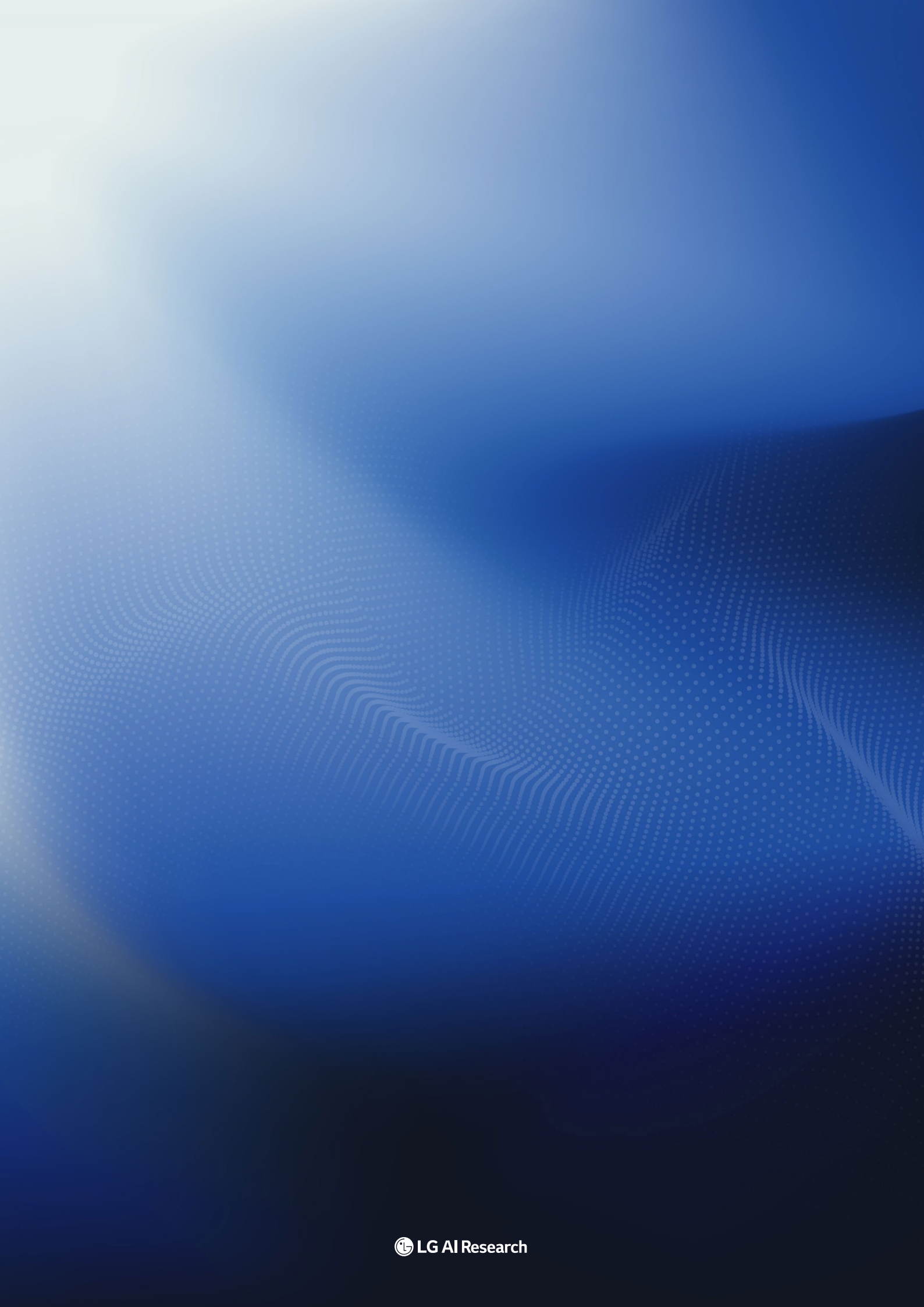
In contrast, it is not always clear whether the term "sub-licensable" permits any distribution and/or redistribution. In this report, we make a distinction between the terms "distributable" and "sub-licensable," where "sub-licensable" means that the dataset can only be provided to the sub-licensee without permitting the sub-licensee to provide the dataset in the downstream through multiple tiers (i.e., dataset cannot be distributed or redistributed by the sub-licensee).

In any case, the license language should not be the sole determinant in deciding whether the licensee is granted with the right to distribute the dataset; instead, the actual intended meaning of the license term should be reasonably determined to decide whether the dataset may be distributed and/or redistributed.

As open source licenses must, by their very nature, allow unrestricted distributions and redistributions to qualify as open source, it is unlikely that any open source license will fall into this Type 2-⑤. However, there may exist commercial licenses that are only "sub-licensable" and not "distributable."

Most importantly, the reason Type 2-⑤ has been assigned a category of its own is because distribution rights are more likely to create license violation risks or conflicts between licenses. For example, even if most of the constituent datasets in a mother set are "distributable," if any one dataset is merely "sub-licensable," then distributing the entire mother set may constitute a license breach. Also, an individual dataset may come with non-sublicensable distribution rights, but the license for the mother set may permit unrestricted distribution. In such a case, sub-licensing to a third party based on the license terms of the mother set may inadvertently create the risk of breaching the license for the individual dataset.

83 The distribution terms of open source software must comply with the following criteria:
1. Free Redistribution
The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale...
3. Derived Works
The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software...
7. Distribution of License
The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties. *Opensource.org, The Open Source Definition*, https://opensource.org/osd.


LG AI Research