Copyright Infringement in Generative AI Input Data Acquisition

Ruyu Yan
Faculty of Law, Macau University of Science and Technology, Macao, China

Keywords: Generative Artificial Intelligence (GenAI), Copyright Infringement, Data Acquisition, Fair Use.

Abstract: The rapid development of generative artificial intelligence (GenAI) increases the risk of copyright infringement during data acquisition and use. This study examines infringement risks at GenAI's input stage, focusing on the legal conflicts in data collection, processing, and output. It highlights substantial violations of economic rights, such as reproduction and adaptation. Under China's Copyright Law, statutory licensing is inapplicable due to non-compliant subject qualifications and behavioral discrepancies. Fair use defenses fail because of commercial intent and excessive scope. Tests and analyses, including the three-step test, four-factor analysis, and transformative use doctrine, consistently show non-exemption. To address liability asymmetries, algorithmic opacity requires a fault presumption mechanism with a reversed burden of proof. To counter enforcement deficiencies, the study proposes novel remedies like dynamic compensation models and algorithmic injunctions. It concludes with institutional recommendations: enforcing enhanced robots.txt compliance, creating open-licensed data repositories, and developing international compliance frameworks to balance technological innovation with copyright protection.

1 INTRODUCTION

The accelerated evolution of generative artificial intelligence (GenAI) exerts profound societal impacts. While catalyzing transformative innovation in the Fourth Industrial Revolution, its unregulated deployment within incomplete legal frameworks has triggered pervasive infringement litigation.

In 2022, software engineers instituted proceedings against GitHub for unauthorized code exploitation (DOE, 2022). March 2023 witnessed artists filing claims against Stability AI for scraping copyrighted images to train models and generate derivatives (Andersen et al, 2023). Q4 2023 saw The New York Times litigate against OpenAI and Microsoft for training on millions of unlicensed articles, seeking data expunction and statutory damages (Li, 2023). Concurrently, generative art features on NetEase's LOFTER and Xiaohongshu platforms precipitated mass user attrition due to unauthorized training practices. These developments critically erode creator incentives and stifle innovation ecosystems.

Scholarly consensus regarding copyright infringement in GenAI input data acquisition remains elusive. While predominant academic opinion endorses fair use exemptions to foster AI advancement, a significant minority advocates robust creator rights protection to ensure innovation quality (Xu & Yang, 2019 … Jiao, 2022). This paper synthesizes these divergences to examine infringement liabilities, fair use controversies, and regulatory paradigms for GenAI systems.

2 ANALYSIS OF COPYRIGHT INFRINGEMENT PATHWAYS

2.1 Data Acquisition as Functional Exploitation

GenAI's purported "creativity" derives from computational architectures of large language models (LLMs) and corpus ingestion during training. LLM construction and Transformer algorithm optimization require massive datasets for pre-training/refinement. Critically, training data quality dictates GenAI output fidelity. Google's text models ingested >1.5 trillion tokens during training, while ChatGPT-3 (released June 2020) utilized multi-terabyte pre-training corpora (Ye, 2025).

As acquired data materially constitutes generative capability through model training, such acquisition,

Yan, R.
Copyright Infringement in Generative AI Input Data Acquisition.
DOI: 10.5220/0014390300004859

In *Proceedings of the 1st International Conference on Politics, Law, and Social Science (ICPLSS 2025)*, pages 577-582
ISBN: 978-989-758-785-6

though superficially informational, constitutes functional exploitation. This is demonstrated by its direct utilization in: (i) Model parameterization processes;(ii) Enhancement of content-generation efficacy.

2.2 Does Data Utilization Constitute Infringement?

Article 52 of the Copyright Law of the People's Republic of China enumerates eleven infringement liabilities. Provisions I, II, VI, VII, and VIII expressly stipulate that unauthorized use constitutes infringement. Regarding GenAI training data acquisition, scholars contend that securing mass-scale licensing from numerous rights holders is prohibitively costly and impractical for service providers (Jiao, 2022).

Unauthorized data acquisition violates Article VII, which prohibits "using copyrighted works without remuneration." Such use constitutes actionable infringement requiring civil liability. Nota bene: While this analysis addresses input-stage infringement, the concealed nature of such acts creates evidential barriers: Infringement processes lack traceability; Direct evidence is largely inaccessible; Determination must rely on output-stage "access + substantial similarity" tests.

2.3 Rights Infringed in Data Acquisition

GenAI entails phase-specific copyright infringement risks throughout its data processing lifecycle: During input phase, unauthorized reproduction and storage of works in training media may directly violate reproduction rights; In processing phase, deconstruction, reorganization, or adaptation of original works through translation, annotation, or compilation for model optimization may infringe derivative rights, including translation, compilation, and adaptation rights; when it comes to output phase, dissemination of generated content bearing substantial similarity to source works in expression or core creative elements may trigger communication to the public rights infringements.

Collectively, these full-process chain activities—from data collection and processing to content generation—create direct conflicts with copyright law. Core legal controversies center on: Whether unauthorized reproduction/derivation qualifies as fair use; The applicable standard for determining legally cognizable similarity between outputs and source works.

2.4 Analysis of Infringement Liability Exemptions

2.4.1 Statutory Licensing

China's Copyright Law establishes five statutory licensing regimes. Besides Article 25 (textbook compilation), the other four are Article 23 (periodical reprinting), Article 35 (phonogram production), Article 46 (broadcasting of published works), Article 50 (digital reproduction by public institutions).

A systematic review confirms none of the licensing regimes mentioned above can be applied to GenAI data acquisition.

In the first place, there is a discrepancy regarding the eligibility of the subject entities. Statutory licenses are strictly limited to specific entities, such as textbook compilation organizations, periodical publishing units, and producers of sound recordings. However, the entities involved in artificial intelligence (AI) research and development are predominantly commercial companies, which do not meet the qualifications of the legally stipulated entities.

Secondly, the criteria for the required actions are not met. Each statutory license mandates specific modes of use, such as textbook compilation and production of sound recordings. In contrast, the method of obtaining data for AI involves complex technological processes such as data scraping, storage, and analysis via information networks. This method significantly differs from the specific modes of use stipulated in statutory licenses.

Finally, there is an insufficiency in procedural requirements. Except for the clauses related to textbooks, other statutory licenses retain the rights for copyright holders to prohibit use. In the absence of explicit permission from copyright holders and without an effective mechanism for such declarations, the acquisition of data for AI does not meet the procedural requirements of statutory licensing (Zeng, 2019).

2.4.2 Fair Use

• Statutory Analysis

Article 24 of China's Copyright Law adopts an exhaustive list with open-ended clause structure. None of its twelve specific exceptions encompass GenAI data acquisition. Scholar Jiao Heping conducts a detailed analysis of several contested points that are relatively relevant (Jiao, 2022).

He elaborates the reasons why AI-generated creations do not comply with the first clause on

"individual learning and research": Firstly, from the perspective of the subject, the "individual" in "individual learning and research" typically refers to a natural person. However, in the context of AI-generated creations, the entity using the data is the AI system rather than a natural person. Secondly, concerning the purpose requirement, individual use must be based on the non-commercial purposes of "learning and research." Currently, AI-generated creations are predominantly controlled and executed by large commercial internet companies, which are unlikely to meet the non-commercial purpose requirement of personal use. Thus, this requirement is not satisfied.

The reasons why AI-generated creations do not comply with the second clause on "proper citation" are also explained: Firstly, they do not meet the purpose requirement, as the use of data in AI-generated creations aims to produce new works rather than to "introduce or comment on a specific work" or "clarify a particular issue." Secondly, the use of data works in AI-generated creations clearly exceeds the requirements of propriety. Therefore, this requirement is not fulfilled.

The attention is also drawn to why AI-generated creations do not comply with the sixth clause on "scientific research": Firstly, the type of fair use for scientific research specified by copyright law involves restrictions on copyright aiming at public interest. Therefore, under this provision, research institutions and activities should "only apply to state-established educational and research public institutions," which does not include commercial internet companies. Secondly, this type of fair use imposes limitations on the number of copies made, while AI-generated creations often involve the full-text replication of works, failing to meet the "limited quantity" requirement. Thirdly, the entities using AI data are not "restricted to use by researchers only." As such, this requirement is not met.

Consequently, statutory interpretation alone precludes fair use exemptions for generative AI, establishing prima facie infringement liability. However, Article 8 of the Supreme People's Court's 2011 Opinions on Promoting Socialist Cultural Development through Intellectual Property Adjudication introduced a hybrid standard expanding fair use boundaries, stipulating that courts may recognize fair use in exceptional circumstances necessitated by technological innovation or commercial development after evaluating: (i) purpose and character of use; (ii) nature of copyrighted work; (iii) substantiality of portion used; and (iv) market impact—provided such use neither conflicts with normal exploitation nor unreasonably prejudices rightsholders' legitimate interests. This establishes a multifactor framework requiring demonstrated necessity, exceptional circumstances, satisfaction of the four-factor analysis, and compliance with the three-step test. For comprehensive rigor, this analysis incorporates the Berne Convention's three-step test (Art. 9(2)), U.S. Copyright Act's four-factor standard (§107), and the contemporary transformative use doctrine.

- Combining the " three-step test "

When applying the three-step test—which assesses fair use through sequential criteria: (1) limitation to certain special cases; (2) non-conflict with normal exploitation; and (3) non-prejudice to legitimate interests—GenAI data acquisition fails all requirements (Xiong, 2018). First, it satisfies no statutorily recognized "special case." Second, market substitution effect analysis confirms GenAI's output capabilities displace human creators in multiple domains, violating criteria (2) and (3) by unreasonably prejudicing economic interests. Finally, unauthorized data ingestion lacks normative legitimacy under fair use doctrine. Scholar CHANG Ye consequently contends that unlicensed GenAI training remains non-exempt under China's copyright framework (Ye, 2025). While the author concurs with this assessment, the three-step test's inherent ambiguity necessitates judicial refinement before full adoption in Chinese jurisprudence.

- Combining the "Four-Factor Test"

The "Four-Factor Test" originates from Section 107 of the United States Copyright Act, which enumerates some typical scenarios of fair use and sets flexible and open general provisions. Judges can comprehensively consider the following four factors to determine whether it constitutes fair use: ① the purpose and character of the use; ② the nature of the copyrighted work; ③ the amount and substantiality of the portion used in relation to the copyrighted work as a whole; ④ the effect of the use upon the potential market for or value of the copyrighted work (Shen, 2020). It mainly explains "fair quotation". However, the data acquisition of generative artificial intelligence is for creation, and the research and development companies are for commercial purposes. Moreover, the quotations are basically full copies, and the learning and imitation of the creator's unique style will undoubtedly affect the market position of the original author and intensify market competition. Therefore, the author believes that the "Four-Factor Test" cannot be used to defend fair use.

- Combining the "transformative use"

The "transformative use" standard was first proposed by Judge Leval in the United States in the judgment criteria for fair use of copyright and has been permitted for application in American judicial practice, with its connotation becoming increasingly clear. In the application of the "transformative use" standard, the key lies in determining whether the new work adds new content with different characteristics, using new expressions, meanings or information. The more transformative the new work is, the more likely it is to constitute fair use (Gu & Fang, 2023).

Subsequently, this standard was introduced into China's judicial practice and academic discussions, providing an important reference for the fair use system and effectively compensating for the closed nature of the Copyright Law. For instance, in cases heard by courts in Beijing, Shanghai, Guangdong and other places, the content of "transformative use" has appeared in over 30 judgments (Han, 2023).

Although "transformative use" has been frequently cited in judicial decisions, it is still limited to specific contexts such as data generation for educational and training purposes. Even when courts refer to this standard, they often impose additional thresholds such as "commercial purpose" and "market impact". However, the author holds a contrary view and does not recommend using "transformative use" as a defense element for fair use. Essentially, "transformative use" still involves deep learning, imitation, and utilization of the original work. The works output by generative AI after processing through the "algorithm black box" do indeed present "new expressions" due to the absorption of a large amount of work data and the integration of various styles, but in essence, they are still pieced together from original works.

Netizens jokingly refer to AI as a "sewing monster". If such pieced-together creations can be exempted from legal liability, it would be akin to feeding AI on the "bones" of original authors, which not only chills and terrifies them but also greatly dampens their creative enthusiasm, making it difficult to foster a healthy and positive creative environment.

In conclusion, whether based on legal provisions or various analytical methods, the analysis points to one result: the infringement of obtaining input data for generative AI cannot be exempted and should bear the liability for infringement.

3 DILEMMAS OF ATTRIBUTION AND PUNISHMENT

3.1 Dilemma of Attribution

The principle of attribution is the standard and principle for determining the civil liability that different types of tortious acts should bear, which decides the elements of liability for a certain tortious act, the burden of proof, the conditions for exemption, the principles and methods of damages compensation, etc (Wang, 2010).

However, there are significant difficulties in its practical application. Even if it can be determined that generative artificial intelligence has used unauthorized works in the "feeding" of data for large model training, there are still new problems brought by new technologies in confirming the responsible party. The output content of generative artificial intelligence, on the one hand, relies on the training of massive data, and on the other hand, is based on interaction with users.

Therefore, the subjects of data "feeding" may involve both users and generative artificial intelligence, and the possible infringing party is naturally not unique. The cause of this attribution dilemma also lies in the complexity of the explainability of algorithms from the input end to the output end of generative artificial intelligence.

3.2 Dilemma of Punishment

Making the infringer bear responsibility is an important means of punishing infringement. According to Article 52 of China's Copyright Law, if the data "feeding" of generative AI models does indeed involve infringement, the infringer may be required to assume responsibilities such as ceasing the infringement, eliminating the influence, making an apology, and compensating for losses (Li, 2003).

However, the issue of determining the amount of compensation for losses has always been a difficult problem in the field of intellectual property research. What is even more challenging is that responsibilities such as ceasing the infringement and eliminating the influence may become unenforceable or difficult to verify when applied to generative AI, a special object.

Currently, the training of generative AI in practice is generally unidirectional and progressive, and cannot be reversed. Many studies have shown that the "contribution" of previous infringing "feeding" training may continue to have an impact on the subsequent content generation of generative AI, and

the infringement of the copyright holder's rights by generative AI may persist (Wang, 2003).

Therefore, for generative AI, ceasing the infringement and eliminating the influence may essentially become unenforceable or difficult to verify. This dilemma still has a lot of room for research.

4 FRAMEWORK CONSTRUCTION FOR REASONABLE REGULATION

4.1 Establishing a Technical Authorization Mechanism

Professor Chang Ye proposed regulating the use of generative AI data by implementing the Robot Exclusion Protocol and introducing mandatory "machine forgetting" norms (Ye, 2025). The author fully agrees. As a general rule for web crawlers, the Robot Exclusion Protocol has been incorporated into China's "Self-discipline Convention for Internet Search Engine Services," but its current legal effect is limited. It is suggested to "strengthen its binding force through administrative regulations or legislation" to enhance its actual enforcement. With a certain legal foundation, specific measures can be more easily implemented.

First, establish a technical rule for authorizing the "feeding" of online works, setting whether it is allowed to be used for generative AI training as a necessary rule.

Second, require network service providers to transfer the rule-setting authority to the users who upload the works, and set the default to prohibit the use of user works by generative AI, and prohibit obtaining authorization through user agreements. This move can substantially ensure that the works uploaded by users are not crawled by AI, and better protect the copyright rights of users.

In addition, I suggest that to improve the quality of data obtained by AI and encourage users to grant open licenses, a reward mechanism for users who grant open licenses can be further improved. For example, users who upload a certain number and quality of works and have them adopted by AI can be awarded certificates or monetary rewards. This will further promote the creation of high-quality works and the development of AI on the basis of protecting the intellectual property rights of original authors.

4.2 Establishing a Prosecutorial Management Mechanism

Professor Gao Yang proposed the establishment of an open licensing mechanism for copyrighted works and a dynamic review mechanism for infringing content (Gao, 2024). The author strongly agrees. In the open licensing mechanism, after the copyright management department reviews the application of the licensor, it publicly announces the information of the data collection of copyrighted works and the licensing conditions.

When potential licensees fulfill their obligations, an open license is formed. This mechanism innovates the traditional one-on-one negotiation model between copyright holders and AI enterprises for licensing. It adopts a new form where copyright holders voluntarily license to the public, set licensing fees and payment methods, and licensees can obtain the data collection upon accepting the conditions.

This move not only benefits AI enterprises in obtaining training data, breaking down data barriers, and helping small and medium-sized AI enterprises access high-quality data, but also enhances the utilization efficiency of the data collection of copyrighted works, promoting mutual benefit and win-win situations between the copyright industry and AI enterprises.

5 CONCLUSION

This study has not exhaustively addressed compensation mechanisms for input-stage infringement risks in GenAI. A comprehensive analysis of damages quantification standards and liability forms requires further empirical investigation into dispute resolution practices. GenAI regulation constitutes a global regulatory challenge. As GenAI evolves into a productivity tool, transborder data flows become inevitable. Consequently, governance of training data acquisition necessitates international cooperation and regulatory harmonization. A critical imperative remains: establishing harmonized regulatory frameworks to mitigate copyright infringement risks during GenAI training data acquisition through collective international efforts.

REFERENCES

Andersen et al. v. 2023. Stability AI et al. 3:23-cv-00201, N.D. Cal.

C. Ye. 2025. Regulation of Copyright Infringement behavior in the “Feeding” of Generative AI Data. *JOSAL*, 2.

DOE 1 v. 2022. GitHub, Inc. 4:22-cv-06823, N.D. Cal.

H. Jiao. 2022. Copyright risks and resolution paths in data acquisition and utilization in AI creation. *CPJ*, 4.

H. Li. 2003. Reversal of the burden of proof: theoretical analysis and problem research." *SILAB*, 4, 87-94.

L. Wang. 2010. Characteristics of the accountability principles system in China's tort liability law. *Legal Forum*, 25, 7-10.

L. Wang. 2023. On some Issues of reversal of the burden of proof." *GDSS*, 1, 150-158.

N.F. Gu, and Z. Fang. Reasonable boundaries and Infringement Regulations for the use of works by generative AI like ChatGPT." *DLF*, 7.

Q. Xiong. 2018. Clarification of judicial standards for fair use of copyright. *OJLS*, 1, 144-157 .

S. Liu. 2024. Exploration of legal pathways for fair use of copyright in AI data training. *JNJ*, 46.

T. Zeng. 2019. Research on copyright infringement issues in AI Creation. *HBLJ*, 37 .

W. Han. 2023. Research on risk avoidance pathways for copyright infringement liability in digital library construction: from the perspective of introducing transformative use rules in judicial trials of copyright cases." *LST*, 3, 64-69

W. Shen. 2020. Study on transformative use from the perspective of the public domain in copyright law." *SALJ*, 5 29-30.

X. Xu, and Y. Yang. 2019. On the fair use of copyright in AI deep learning." SJTU Law Review, 3.

X. Zhao. 2024. Issues of fair use in generative AI and machine learning." *JNJ*, 3.

Y. Gao. 2024. Regulation of Copyright Infringement in AI training data. *CPJ*, 15.

Y. Li. 2023. The urgency of AI governance. *CBJ*, 3.