A Study on the Graded and Categorized Regulatory Path for Copyright of Generative AI Training Data

Jie Ni
College of Intellectual Property, China Jiliang University, Hangzhou, China

Abstract: With the rapid advancement of emerging technologies and continuous evolution of the artificial intelligence industry, generative AI models such as Grok and DeepSeek have emerged en masse and quickly captured market share. During the data input phase of model training, practices involving web scraping and data processing—using copyrighted works without the permission of their original rights holders—have triggered a series of copyright infringement disputes. Based on principles of interest balancing and proportionality, this paper argues that generative AI data training indeed poses real risks of copyright infringement. Traditional regulatory approaches such as licensing, fair use, and statutory licensing each have limitations, necessitating the development of a tiered and categorized regulatory framework. Drawing on relevant case law, this paper classifies data training activities according to the purpose and characteristics of the training process, as well as the legal status and nature of the trained works, and introduces the EU's Text and Data Mining (TDM) exception to categorize different types of generative AI data training. During the transitional period of implementing related systems, it proposes adopting and refining the U.S. four-factor test to comprehensively determine whether an exemption from liability applies. Through this analytical framework, the paper aims to establish a structured, tiered governance approach to better address copyright regulation challenges in generative AI data training.

Keywords: Generative AI; Data Training; Tiered Classification; TDM Exceptions

1. Introduction

Currently, artificial intelligence has gradually evolved into a driving force behind digital economy through continuous iteration, transitioning from a neutral tool to a collaborative partner [1,2]. Driven by research objectives in machine learning, connectionism is increasingly supplanting symbolism. This article focuses on the generative AI data training practices based on connectionism—specifically, human-brain-like neural network connectivity and parallel computation—and examines the copyright infringement disputes arising during its training process.

During the data input phase of model training, a significant portion of high-quality data obtained through technical implementation—such as web scraping and data processing—using copyrighted works without the permission of the original rights holders comes from works still under copyright protection [3,4]. Ross Intelligence, the U.S. Court of Appeals for the Third Circuit applied the four-factor test and concluded that fair use did not apply. In short, at the judicial level, neither domestic nor international courts have established a consistent standard, making traditional copyright theories inadequate in effectively addressing issues surrounding the use of AI training data.

2. The Practical Dilemmas of Copyright Regulation on Generative AI Training Data Usage

Current artificial intelligence systems based on large language models often rely on human feedback regarding generated text to iteratively refine and optimize their performance, aiming to better align with "human-like" evaluation metrics. This data training practice in generative AI differs significantly from traditional uses of works, continuously challenging and undermining existing copyright legal frameworks. At present, regulating the data training practices of generative AI under copyright law faces several significant practical challenges [5].

2.1 Ambiguity in the Theoretical Characterization of Behavior

 Copyright @ STEMM Institute Press

The data training practices of generative artificial intelligence are characterized by socialization and publicness, making it difficult to directly define them as acts of reproduction or adaptation under copyright law [6,7].

Before discussing the qualitative aspects, it is necessary to first clarify the specific process of data training in generative artificial intelligence. The data training end can be divided into the data input end and the initial data output end. Here, the initial data output end differs from the secondary data output end and refers specifically to the data output end of the AI's foundational model itself. The data input end encompasses three interrelated and sequentially connected stages: data collection, data cleaning, and data usage, while the initial data output end corresponds to the stage of data output [8].

Based on the four time phases outlined above, this study categorizes the data training activities of generative artificial intelligence into three types: first, data replication behavior that spans the entire lifecycle—from data collection and cleaning to data usage and output—specifically referring to the act of mining, inputting, annotating, and storing works protected by copyright in model training servers without the permission of the original rights holder; second, data organization behavior during the data cleaning phase, which involves extracting data features from original works and compiling them into underlying code or integrating weights into model parameters; third, data output behavior during the output phase, where a base model processes user prompts, runs internal algorithms, and ultimately generates outputs such as text, images, or music. It should be emphasized that, to reflect the completeness of the process, the initial data output end is also included within the data training process. However, this article focuses exclusively on data input-side training behaviors in generative AI and does not address data output behaviors [9].

Regarding the nature of data curation activities, based on the provisions concerning derivative rights under Article 10(14) to (16) of the Copyright Law of the People's Republic of China, this study holds that such activities do not fall within the scope of derivative works as defined by copyright law and therefore do not constitute infringement of derivative rights. A representative view argues that data curation involves selection, modification, and arrangement of copyrighted works, and its outcome may produce new works with originality, thus posing a risk of derivative rights infringement. However, in the context of base model training, data curation neither substantially preserves the original work's fundamental expression nor creates a new derivative work with originality; it fails to meet the elements required for derivative rights infringement. On one hand, the entire process of data curation merely involves a purely technical decomposition of raw data into minimal units recognizable by algorithms, which are then aggregated into a dataset. This process includes only non-creative acts of extraction and segmentation, without any alteration or arrangement of the original work. On the other hand, the resulting dataset is highly homogenized, lacks perceptible expressive form, and does not qualify as a work under copyright law. Therefore, data curation activities do not constitute derivative works within the meaning of copyright law and do not amount to derivative rights infringement [10].

Regarding the nature of data copying, mainstream academic views can be summarized into three categories based on Article 10(5) of the Copyright Law of the People's Republic of China, which pertains to the right of reproduction. The first view holds that data copying does not fall within the scope of copyright-protected reproduction and therefore does not constitute infringement of the reproduction right. Supporters cite the idea-expression dichotomy, arguing that data copying extracts abstract ideas—classified as "non-authorial value-based foundational knowledge" in the public domain—rather than protected original expressions. The second view asserts that data copying is indeed a form of reproduction covered by copyright law and thus constitutes infringement of the reproduction right. Advocates of this position disregard the technical purpose of copying and the nature of intermediate copying, maintaining that any substantial replication without the original rights holder's permission amounts to infringement. The third view extends the second, proposing that data copying should be exempted from liability through doctrines such as fair use or statutory licensing. Proponents argue that data copying occurs throughout the entire lifecycle of data collection, cleaning, usage, and output, forming the most fundamental and critical activity in training generative artificial

Copyright @ STEMM Institute Press

intelligence models. To prevent the emergence of a "tragedy of the anticommons," data copying should be defined as a fair use act serving significant public interest and legitimate purposes. While each of these viewpoints has considerable support, perspectives and degrees of application vary significantly across different countries and legal systems, making it difficult to establish a universally applicable standard [11].

2.2 Failure of Institutional Regulations at the Legislative Level

The use of works in the data training process of generative artificial intelligence represents a novel mode of work utilization [12,13]. This failure of institutional regulation at the legislative level has led to a lack of uniform standards for determining copyright infringement related to data training activities in generative AI.

Legislative models across countries have not negated the significance and economic value of training data for generative artificial intelligence; instead, they all seek to establish new copyright exceptions or limitations to preserve technological development space for such training activities [14,15]. While these models offer practical experience for China, the divergences and value conflicts among them make the establishment of uniform rules increasingly ambiguous and challenging.

Copyright classification of data training practices varies across jurisdictions, leading to divergent judicial outcomes [16,17]. This inconsistency creates uncertainty for industry participants and hinders the development of predictable regulatory frameworks.

3. Pathways and Critiques of Copyright Regulation on Generative AI Training Data Practices

The current academic proposals for copyright regulation of data training practices in generative artificial intelligence can be categorized into three approaches. First, the use of copyrighted works in training generative AI is deemed an infringing act of reproduction, thus not subject to any exemption rules. Second, while acknowledging that such training constitutes infringement, certain models such as fair use or statutory licensing are applied to provide exemptions, thereby expanding or redefining the boundaries of existing exceptions. Third, the non-expressive use doctrine and temporary copying theory exclude the use of works during the AI training phase from copyright protection altogether, eliminating the need for further analysis under copyright exceptions. However, these regulatory pathways struggle to address the complex scenarios involving multiple stakeholders, diverse rights interests, and varied conduct patterns inherent in data training contexts, resulting in a misalignment between institutional supply and practical demands. This study will raise critical questions on this matter below [18].

3.1 Critique of the Regulatory Approach to "Fair Use"

Under the closed legislative model of Article 24 of the Copyright Law of the People's Republic of China, it is explicitly stipulated that only acts falling within the 12 specific scenarios listed in the provision or those otherwise specified by laws and administrative regulations can constitute fair use [19,20]. This works fine for traditional activities, but it struggles to handle new technologies like AI training, which did not exist when the statute was written. Courts have to stretch the existing categories to cover AI training, which leads to inconsistent and sometimes strained decisions.

At the level of legal provisions, only two scenarios listed in the statutes—“personal study, research, or appreciation” and “scientific research”—are relevant to data training activities for generative artificial intelligence. The subject of “personal study, research, or appreciation” is strictly limited to natural persons, whereas the primary subjects involved in generative AI data training are machine models developed by enterprises, research institutions, and similar entities. Although some scholars argue that the process of data training in generative AI closely resembles human learning, thus drawing analogies between machine learning and human cognition, it is evident that similarity in process does not equate identity in subject; machines cannot be equated with natural persons. Furthermore, “scientific research” is subject to restrictions such as use “for teaching or scientific purposes” and “translation, adaptation, compilation, performance, or limited reproduction.” In contrast, data training typically involves large-scale, profit-driven data extraction and utilization, which clearly falls outside these

 Copyright @ STEMM Institute Press

limitations. If data training were exempted under a doctrine of fair use, it would not only contradict the original intent of fair use principles but also violate the principles of equitable balance and proportionality, thereby failing to protect the legitimate rights and interests of copyright holders [21].

3.2 Questions on the Regulatory Approach to "Authorization Licensing"

The licensing approach requires artificial intelligence companies to obtain explicit authorization from every rights holder before using works for data training, which is impractical in reality and essentially an idealized institutional design detached from actual conditions [22,23].

First, the cost of obtaining authorization for massive data exceeds the capacity of most market participants. Current large language models require training on billions or even trillions of tokens, covering nearly all types of works; the transaction costs involved in securing individual permissions would effectively stifle the artificial intelligence industry before it even gets off the ground. Second, technical challenges in tracing ownership make licensing practically unfeasible in most cases. Web crawlers typically strip away copyright notices, author attributions, and other metadata during data collection to improve processing efficiency, resulting in severe metadata deficiencies across mainstream training datasets. Under these circumstances, users cannot determine which works require authorization, let alone contact the respective rights holders. Third, the licensing process inevitably exacerbates data monopolization and undermines fair market competition. Major companies, leveraging their substantial financial resources, can secure exclusive or preferential bulk licensing agreements with large content platforms and collective copyright management organizations, thereby monopolizing high-quality training data. In contrast, small and medium-sized enterprises and startups are unable to afford the high licensing fees, effectively excluding them from the competitive landscape. This leads to a "rich get richer, poor get poorer" Matthew effect, severely hindering innovation and development in the artificial intelligence industry [24].

3.3 Critique of the Regulatory Approach to "Statutory Licensing"

Statutory licensing is a copyright infringement exemption pathway that allows certain acts controlled by exclusive rights to be carried out without the copyright holder's permission, provided that remuneration is paid [25,26].

On one hand, statutory licensing essentially restricts the copyright holder's right to authorize use, and its scope should be strictly limited to specific non-profit scenarios such as textbook compilation and newspaper reprinting. However, the significant economic value generated by data training is primarily captured by artificial intelligence companies. Including such activities within the scope of statutory licensing would, in essence, compel private creative works to be transferred for commercial use through legal force, lacking a legitimate foundation. On the other hand, the collection, management, and distribution of statutory licensing fees depend on the unified operation of collective copyright management organizations, whose actual implementation outcomes at the end stage are difficult to predict—making it an ideal but time-tested solution. In China, the standards for determining statutory licensing remuneration mainly include royalty-based payment, fixed-rate payment, and usage-proportional payment. Applying these remuneration standards to data training practices would not only excessively burden artificial intelligence enterprises but also increase the costs associated with interest allocation and procedural design by public authorities. To avoid these issues, this paper proposes introducing an extended collective copyright management organization that follows a tiered and categorized approach. This model combines broad upfront authorization with an exit mechanism for non-member rights holders, thereby maximizing authorization efficiency. At the same time, a composite fee structure is applied for specific remuneration standards, significantly lower than those under China’s current statutory licensing framework, aiming to reduce the financial burden on AI companies and properly balance data utilization in science, culture, and art with copyright protection [27].

4. Reflections on the Pathway for Grading and Categorizing Generative AI Data Training Practices

Building on the critiques of existing regulatory models, this chapter elaborates on the necessity and feasibility of adopting a tiered and categorized approach to generative AI data

training. The following sections address both the theoretical justification for this framework and its practical implementability.

4.1 Justification for the Necessity of a Graded and Categorized Approach

Specifically, the tiered and categorized regulatory framework for generative AI data training can be outlined as differentiated regulation at the legislative level and transitional rules at the judicial level [28,29]. Second, the classification logic based on objects—categorizing trained works according to their rights status and nature into public domain works and non-public domain works. Non-public domain works can be further classified into orphan works, individual works, and organizational works. The higher the commercial value embodied in these works, the greater the potential market impact when used for data training, thus requiring correspondingly stricter rules for acquisition and use. Third, the differentiated regulatory logic based on conduct—designing distinct institutional frameworks according to the purpose and characteristics of the data training entity and the rights status and nature of the trained works, thereby achieving a balanced interest structure among all parties involved.

To ensure the predictability and stability of legal frameworks, the formulation of legal norms requires careful and lengthy procedures [30,31]. four-factor test: the purpose and character of the generative AI data training activity, the rights status and nature of the works used, the proportion and extent of content utilized, and the impact of the training activity on the potential market or value of the original works. Furthermore, to support this transitional rule, a preliminary review mechanism for legitimate sources should be established according to the origin and acquisition method of the training data. Drawing on the risk-based classification regulatory framework introduced by the EU Artificial Intelligence Act, which came into effect on August 1, 2024, differentiated regulation should be applied accordingly.

4.2 Feasibility Justification for a Tiered Classification Framework of Generative AI Data Training Practices

The proposed tiered classification framework is feasible across institutional, practical, and technological dimensions. The following subsections examine these three aspects in detail. Generative AI data training practices serve as the fuel that powers artificial intelligence engines and are a key element in fostering new forms of productive capacity [32,33]. The this study will elaborate on this in detail below.

5. Rule Construction for the Grading and Categorization Path of Generative AI Data Training Behaviors

At the theoretical level, this study has elaborated in detail on the necessity, feasibility, and legitimacy of adopting a tiered and categorized approach, thereby laying a solid theoretical foundation for establishing its rules. Building upon this, this study will further develop specific rules for the tiered and categorized framework, moving beyond existing analytical models by formulating a more suitable solution tailored to China's national conditions and public sentiment through three dimensions: legislative standards, judicial practice, and supporting institutional mechanisms [34].

5.1 Differential Regulation of Legislative Norms

Legislative reform represents the primary avenue for establishing a systematic framework for generative AI data training. Drawing on the tiered classification logic outlined above, this section proposes specific legislative measures centered on the introduction and refinement of TDM exception rules.

5.1.1 Theoretical justification for introducing TDM exceptions

First, based on the underlying logic of tiered and categorized governance aligned with the classification of "scientific research purposes" and "text and data mining purposes," this paper categorizes the subjects of data training activities into for-profit and non-profit entities, and classifies the objects of such activities into orphan works, individual works, organizational works, and public domain works. Among them, for-profit entities can be further subdivided into those conducting pre-training for general foundational models and those performing commercial fine-tuning within specialized domains. This typology of subjects and objects stems from the TDM exception and serves as a standardized framework derived therefrom; the TDM exception not only provides the existing regulatory blueprint in the DSM Directive but also functions as a concrete reference standard

 Copyright @ STEMM Institute Press

for tiered and categorized governance [35,36].

5.1.2 Selecting an appropriate TDM exception model

On one hand, this study argues that the TDM exception should not be confined solely to fair use, nor should fair use be expansively interpreted merely to encompass the TDM exception [37,38]. Applying the three-step test to determine whether data training constitutes fair use not only increases the difficulty for judges in reasoning and exercising discretion but may also lead to inconsistent rulings in similar cases. Currently, there are no precedents in China that have clarified the legal status of data training activities, precisely due to these reasons. As China is a civil law country rather than a common law jurisdiction, there are neither established regulations nor uniform standards governing data training practices at present. Relying exclusively on judicial practice to define the TDM exception lacks practical feasibility.

5.1.3 Establish specific TDM exception rules

Given the newly introduced model for TDM exceptions discussed above, this paper will establish specific TDM exception rules within the framework of existing legal norms, based on a foundational logic of subject categorization and object classification. It should be emphasized that these rules are not entirely new solutions to emerging issues, but rather refinements to current copyright regulations, developed under careful consideration of both the feasibility of legislative amendments and the potential for interpretive clarification. Furthermore, from the perspective of object classification, data training involving works in the public domain generally does not raise copyright infringement concerns, and thus will not be further elaborated here. (1) Application of the TDM Exception Rule for Non-Profit Entity Data Training Behavior [39]

As discussed above, a new TDM exception clause could be established for data training activities conducted by non-profit entities, with strict conditions governing its application. First, such data training activities must be grounded in legality as a primary requirement; there is no room for exemption rules where data has been obtained unlawfully, let alone any discussion of protecting or limiting the rights holders' interests. In the case of Bartz v. Anthropic, the UK court held that subsequent use of works obtained from pirated websites constitutes an irreparable copyright infringement and cannot be exempted under fair use principles. Second, the post-training operational phase should impose restrictions on non-profit entities: without the consent of the rights holder, these entities must not provide the trained datasets or models to commercial parties in any form. Although this restriction may appear stringent, it is actually intended to ensure the purity of the non-profit purpose and prevent the entity from falling into the trap of "data laundering." Third, the specific definition of non-profit entities should be clarified. Rather than focusing solely on the legal status of the entity, emphasis should be placed on the non-commercial intent. A non-profit entity can be defined as "an entity that uses published works for artificial intelligence model training for purposes such as learning, research, appreciation, classroom instruction, or other non-profit objectives." [40,41]

Based on the above conditions, it is proposed that an exception for text and data mining (TDM) related to data training be added to the Implementing Regulations of the Copyright Law, specifically stated as: "Non-profit entities may use works created by natural persons, legal persons, or non-legal entities for model training without the permission of the rights holder and without payment of remuneration"; "Works used under the preceding provision must not have their resulting datasets or models provided in any form to commercial entities, nor used for commercial purposes." This can also be indirectly supported by referencing the clause in Article 24 of the Copyright Law of the People's Republic of China regarding "other circumstances stipulated by laws and administrative regulations" for fair use. As for the requirement of legality, the provision in Article 24(1) of the Copyright Law—"which shall not unreasonably infringe upon the legitimate rights and interests of the rights holder"—is sufficient to cover this aspect. For post-deployment operational requirements, additional explicit supplementary provisions will further regulate such activities [42].

Compared to non-profit entities, for-profit entities are not motivated purely by public interest when acquiring data; instead, they seek to expand model application scenarios and maximize profits through data exploitation [43,44].

General-purpose foundational model training involves using vast, cross-domain datasets to

Copyright @ STEMM Institute Press

build base models with broad semantic understanding and reasoning capabilities through self-supervised learning. In contrast, vertical domain model training refers to fine-tuning these base models using high-quality, industry-specific data tailored to the business needs of particular sectors. The use of works in general-purpose foundational model training is characterized by fragmented feature extraction that does not preserve complete expressions, leaning toward non-expressive usage. On the other hand, vertical domain model training involves targeted extraction and learning, which more closely resembles expressive use. Therefore, compared to vertical domain model training, general-purpose foundational model training has a relatively smaller market impact on original works, and appropriately relaxing restrictions on it can help promote industrial development and technological innovation [45].

For commercial entities using corporate or individual works to train vertical domain models, the "TDM exception plus opt-out" rule may apply. Compared to the "opt-out" mechanism under the EU's TDM exception framework, this paper interprets "opt-out" more broadly as "rights holders explicitly indicating that their works should not be accessed or processed." Rights holders may prohibit others from acquiring and using their works in model training through contractual agreements, written notices, public declarations, or technical measures such as anti-scraping mechanisms that prevent machine access. Although vertical domain model training is highly commercial in nature, allowing rights holders to opt out preserves their rights while balancing their legitimate interests with the practical needs of artificial intelligence industry development [46].

For profit-driven entities using orphan works in model training, the rule of "TDM exception plus deposit of compensation" can be applied. Orphan works refer to works whose copyright protection period has not yet expired but for which the rights holder is unknown or cannot be located despite diligent efforts; they are commonly described as "works with unidentified or unreachable rights holders." This study limits the obligation to pay compensation to publicly listed companies in the artificial intelligence sector, with an extended copyright collective management organization responsible for drafting and regularly updating a list of such enterprises, which will be made public. On one hand, compared to individuals and small- and medium-sized enterprises, the training datasets of listed companies contain hundreds of millions of works. This large-scale data collection inevitably involves acquiring, copying, and utilizing vast numbers of orphan works, thereby significantly impacting the market. On the other hand, listed companies generate substantial profits and possess mature risk tolerance; thus, charging reasonable compensation would not threaten their survival or development. Furthermore, this study argues that not all orphan works should be subject to the compensation deposit system—only those with significant economic or artistic value, or those irreplaceable in terms of cultural heritage preservation and societal continuity, warrant such protection. However, given the difficulty in defining "significant value" and "irreplaceability," it is impractical to calculate compensation based on the quantity, category, or market value of orphan works. Instead, a fixed fee schedule or a percentage of the listed company's annual revenue could serve as the basis for determining compensation deposits [47].

5.2 Transitional Rules in Judicial Practice

First, the source and method of obtaining training data should be considered as a preliminary step in assessing legality. There are primarily three ways to obtain training data: first, legally authorized acquisition—where parties voluntarily reach an agreement through licensing agreements or content collaborations, thereby securing explicit authorization from rights holders; such acts do not constitute infringement and thus are unrelated to issues of exemption. Second, unauthorized acquisition—obtaining works by bypassing paywalls, hacking private databases, or downloading from pirated websites constitutes infringement and cannot benefit from fair use exemptions. Third, access to publicly available works without technical restrictions—works already disseminated on the internet without access barriers require a comprehensive evaluation based on the U.S. four-factor test for determining fair use [48].

Second, this study draws on and refine the U.S. four-factor test for fair use—purpose of use, nature of the work, amount used, and market effect—. Regarding purpose of use, data training activities are categorized into commercial and non-commercial purposes based on their

 Copyright @ STEMM Institute Press

objectives and characteristics. Commercial uses are further subdivided into training for general foundational models and training for vertical domain-specific models. The commercial intensity of these three categories increases progressively; although commerciality does not necessarily preclude fair use, the stronger the commercial aspect, the higher the threshold for transformative use becomes. Concerning the nature of the work, works are classified as either public domain or non-public domain, with non-public domain works further divided into corporate works, individual works, and orphan works. For the amount used, the extent and scope of content utilization are categorized into fragmented feature extraction, full-text replication limited to internal training, and targeted reproduction of specific works or authors. As for market effect, the impact of training activities on the potential market or value of original works is classified into substantial substitution, loss of business opportunities, market value dilution, and no significant market impact. Drawing from judicial opinions in relevant foreign cases, market effect is treated as the core criterion among the four factors, while integrating the other elements to comprehensively assess whether data training activities qualify for copyright exemption. Copyright law in the age of artificial intelligence must neither hinder technological innovation nor undermine creators' rights. Through these transitional rules in judicial practice, we aim to achieve a balanced interest between technological advancement and creators' rights in case-by-case adjudication, thereby promoting the healthy and orderly development of generative AI industries [49].

5.3 Refinement and Improvement of Supporting Systems

Beyond core legislative and judicial rules, a comprehensive governance framework requires supporting institutional mechanisms. This section proposes several complementary systems, including extended collective management, compensation mechanisms, opt-out provisions, and orphan works deposit systems.

5.3.1 Extended copyright collective management organization

Extended collective management focuses on reducing transaction costs and aligns with the characteristics of data training processes, serving as an institutional solution to address the challenges of licensing vast numbers of works [50,51]. Collective management organizations already exist in many countries for music, text, and visual works. Extending them to AI training data makes practical sense. These organizations can collect licensing fees from AI companies and distribute them to creators based on usage data. This is much more efficient than trying to track down every individual creator and negotiate separate deals.

5.3.2 Composite fee-based copyright compensation system

The legitimacy of the copyright compensation system is grounded in the "market harm compensation theory [52,53]. It employs a hybrid fee structure combining a base rate with adjustable floating components, requiring providers of general-purpose foundational models to pay compensation within a reasonable range, with specific rates adjusted flexibly according to each company's development status and operational scale. The extended collective management organization must be subject to public oversight and is required to regularly disclose to the public the total amount collected, detailed distribution records, operating expenses, and fund utilization from the previous year. Rights holders have the right to inquire about the usage of their works and the corresponding details of compensation allocations.

5.3.3 Limited scope opt-out system

Rights holders shall submit withdrawal statements through the national unified AI copyright declaration platform established by the National Copyright Administration [54,55]. However, general "all rights reserved" statements, standard terms in website user agreements, or withdrawal statements not publicly disclosed on the national unified AI copyright declaration platform shall not produce legal effect for opting out.

The withdrawal statement shall take effect 30 days after being publicly announced on the platform. Artificial intelligence enterprises must remove the relevant works from their training datasets within 90 days of the effective date and cease using them in subsequent model training and iteration processes. It should be emphasized that the withdrawal statement is not retroactive and only applies to training activities conducted after its effective date. For training activities completed prior to the effective date, artificial

Copyright @ STEMM Institute Press

intelligence enterprises shall not bear liability for infringement; however, they must pay compensation to rights holders according to statutory standards from the effective date of the statement until the actual removal of the works.

5.3.4 The deposit system for remuneration of orphan works

As mentioned earlier, this study limits the obligation to pay remuneration to listed companies in the field of artificial intelligence. An extended copyright collective management organization shall be responsible for compiling and regularly updating the list of enterprises, making it publicly available. Additionally, a dedicated escrow account for orphan works remuneration shall be established and entrusted to a qualified third-party financial institution for custody. Funds in the escrow account must be used exclusively for their intended purpose and may not be diverted. The third-party financial institution is required to disclose annually to the public the inflows and outflows of funds held in the escrow account [56].

6. Conclusion

Copyright disputes arising from the use of generative artificial intelligence data for training have become a major issue requiring urgent attention in copyright law during the digital economy era.S. four-factor test, we establish case-by-case discretionary rules for the transitional phase of the system, complemented by coordinated safeguard mechanisms, offering practical solutions to current regulatory challenges. Going forward, legislative, judicial, and industry self-regulatory efforts must work together to ensure creators' rights are effectively protected while fostering a sound legal environment that supports the healthy and orderly development of the generative AI industry.

References

[1] T. Teubner, et al. Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models. Business and Information Systems Engineering, 2023, 65(1): 96.

[2] Thomson Reuters Enter. Ctr. GmbH v. Ross Intelligence Inc., No. 1:20-cv-11195-613-SB. D. Del. 2025.

[3] L. Y. Gu. Research on the legal basis for copyright of AI machine learning content: From the perspective of text and data mining exception rules. Publishing Wide Angle, 2024, (15): 54-59.

[4] Y. F. Liu. Copyright dilemmas and solutions of generative AI data training. Jianghai Academic Journal, 2025, (4): 182-189+256.

[5] H. Huang, P. W. Zhai. Copyright characterization of big model data use from the public domain perspective. Intellectual Property, 2025, (1): 94-110.

[6] J. J. Gao. Research on copyright regulation of generative AI data training. China Publishing Journal, 2025, (24): 52-57.

[7] J. G. Shang. Copyright law analysis of data training. Tribune of Law, 2025, 40(2): 68-78.

[8] New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195. S.D.N.Y. 2023.

[9] J. Y. Zhang, S. F. Wang. Research on copyright fair use in big model data training. ECUPL Journal, 2024, 27(4): 20-33.

[10] M. S. Huang. Copyright risks and regulatory pathways for AI big model training data. Journal of Hubei University, 2025, 52(5): 185-193.

[11] W. B. Liu. Reproduction behavior in generative AI training: Copyright law evaluation from a market perspective. Journal of Gansu University of Political Science and Law, 1-19[2026-04-30].

[12] P. Zhang. Institutional dilemmas and solutions for copyright legality of AI-generated content. Science of Law, 2024, 42(3): 18-31.

[13] X. Chu, Y. L. Zhou. Improving copyright fair use rules adapted for generative AI data training. China Publishing Journal, 2025, (6): 3-7.

[14] C. Liu. Copyright protection and international experience of generative AI data training. Chongqing Social Sciences, 2025, (7): 6-17.

[15] Hamburg Regional Court, Germany. Robert Kneschke v. LAION e.V., Case No. 310 O 227/23. 2024.

[16] Concord Music Group, Inc. v. Anthropic PBC, No. 3:23-cv-01092. 2023.

[17] Q. Tao. Copyright issues in foundation model training: Theoretical clarification and rule application. Tribune of Political Science and Law, 2024, 42(5): 152-164.

[18] W. Y. Shao, Z. Y. Huang. Analysis of fair use of copyright in generative AI model training data. Journal of Wuhan University

 Copyright @ STEMM Institute Press

of Technology (Social Sciences Edition), 1-8[2026-03-17].
[19] X. C. Liu. Non-work use in generative AI data training and its legitimacy justification. Tribune of Law, 2024, 39(3): 67-78.
[20] J. M. Sun. From fair use to fair learning: Copyright dilemmas and normative reconstruction of AI training data. China Law Review, 2025, (5): 52-65.
[21] Q. Li, L. S. Shen. Open licensing rules for copyright in data training and their implementation pathways. Journal of Information Security Research, 2026, 12(1): 68-74.
[22] K. X. Ruan, X. Y. Huang. Research on copyright issues in generative AI data training. China Copyright, 2024, (5): 61-72.
[23] Q. Wang. Intellectual Property Law Course. Beijing: China Renmin University Press, 2021: 312.
[24] Z. Q. Zou, J. Z. Yu. Construction of statutory licensing system for generative AI training data. Artificial Intelligence, 2026, (2): 98-107.
[25] Y. H. Liu, Y. S. Wei. Copyright infringement issues in machine learning and their solutions. ECUPL Journal, 2019, 22(2): 68-79.
[26] S. W. Mei, Y. Gong. Constructing copyright incentive mechanisms for hierarchical classification of data training use. China Publishing Journal, 2025, (20): 68-72.
[27] Y. G. Liang. On hierarchical regulation of AI big model training data risks. Journal of Zhengzhou University, 2025, 58(3): 61-67+144.
[28] State Council. Interim Provisions on Remuneration Standards for Statutory Licensing of Sound Recordings. 1993.
[29] X. Z. Gou. Legal incentives for inclusive governance in the digital economy. Modern Law Science, 2024, 46(1): 92-106.
[30] Q. Q. Feng. Research on interest balance theory in copyright law. Journal of Hunan University, 2008, 22(6): 113-120.
[31] Y. Tan. Copyright obstacles and antitrust approaches to generative AI training data acquisition. Science Technology and Law, 2025, (3): 81-92.
[32] H. D. Wu. Questions of copyright law regarding AI-generated works. Peking University Law Journal, 2020, 32(3): 537-555.
[33] X. H. Xie, Y. Zhao. Accelerating construction of AI Chinese corpora. Study Times, 2025-07-02(6).
[34] A. Strowel. Reconstructing the reproduction and communication to the public rights//Hugenholtz P B. Copyright Reconstructed. Alphen aan den Rijn: Kluwer Law International, 2018.
[35] R. Watt. The Efficiencies of Aggregation: An Economic Theory Perspective on Collective Management of Copyright. Review of Economic Research on Copyright Issues, 2015, 12(1): 26.
[36] X. Y. Yang. Dilemmas and solutions of copyright fair use system in the digital age//Wisdom Rule of Law Collected Papers. 2024: 98-104.
[37] Z. J. Guo, M. L. Li, Y. M. Zhou, et al. Research progress on digital watermarking technology for AI-generated content models. Journal of Cyberspace Security, 2024, 2(1): 13-39.
[38] R. M. Yang, M. Y. Fan. Intelligent ESG audit model construction: A multimodal system study based on RegTech, blockchain and AI. Business Accounting, 2026, (4): 24-30.
[39] U.S. Copyright Office. Copyright and Artificial Intelligence, Part 3: Generative AI Training. (2025-07-10) [2026-05-29]. https://www.copyright.gov/ai/.
[40] H. X. Fan. Construction of text and data mining copyright exception rules in the digital environment. Library Work and Study, 2024, (9): 26-35.
[41] F. M. Bator. The anatomy of market failure. The Quarterly Journal of Economics, 1958, 72(3): 351-379.
[42] L. Zhu. On extended copyright collective management system for generative AI training data. Publishing Wide Angle, 2026, (1): 69-75.
[43] M. Zhang. Interpretive theory on applying copyright fair use to text and data mining. Publishing Journal, 2023, (2): 51-56.
[44] W. B. Liu. Reproduction behavior in generative AI training: Copyright law evaluation from a market perspective. Journal of Gansu University of Political Science and Law, 1-19[2026-04-30].
[45] C. Y. Guan. Justification and rule improvement for including commercial use of data training in fair use. Journal of Hainan University, 1-11[2026-05-30].

Copyright @ STEMM Institute Press

[46] W. Du. Analysis of copyright extended collective management system. Intellectual Property, 2013, (1): 62-65+74.
[47] S. W. Mei, Y. Gong. Constructing copyright incentive mechanisms for hierarchical classification of data training use. China Publishing Journal, 2025, (20): 68-72.
[48] H. J. Lu. On market-oriented copyright collective management. Electronic Intellectual Property, 2007, (3): 16-20.
[49] S. J. Bao, D. M. Xiao. Copyright law responses to generative AI training data: EU exception rules and implications for China. Library Tribune, 2025, 45(6): 123-132.
[50] J. X. Zhang. Research on interest-sharing mechanisms for work data source providers in AI training. Intellectual Property, 2025, (5): 111-126.
[51] K. X. Ruan, X. Y. Huang. Research on copyright issues in generative AI data training. China Copyright, 2024, (5): 61-72.
[52] G. Z. Wang. Institutional analysis and legislative construction of copyright opt-out implied license. Contemporary Law Review, 2015, 29(3): 106-112.
[53] W. Du. Analysis of copyright extended collective management system. Intellectual Property, 2013, (1): 62-65+74.
[54] S. W. Mei, Y. Gong. Constructing copyright incentive mechanisms for hierarchical classification of data training use. China Publishing Journal, 2025, (20): 68-72.
[55] W. J. Zhang. On copyright regulation pathways for data used in big model training. East China Law Review, 2025, (2): 79-92.
[56] J. X. Zhang. Research on interest-sharing mechanisms for work data source providers in AI training. Intellectual Property, 2025, (5): 111-126.

 Copyright @ STEMM Institute Press