Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38155

Full-Process Protection Technology Path for Copyright Risks in Generative AI Training Data

Zewen Zhang

Baoding Heyang Senior High School, Baoding, China
wenchoas@outlook.com

Abstract. The rapid iterative development of generative AI technology is highly dependent on the supply of massive training data. The industry generally uses crawlers to grab Internet copyright works on a large scale for pre-training, which has led to global copyright disputes. The risk of copyright infringement in the acquisition and use of training data has become the core challenge restricting the development of industrial compliance. This article focuses on the complete life cycle of training data and explores three cutting-edge technological paths: pre-filtering of training data, in-process training constraints, and post-detection and evidence collection. Analyze the underlying operating principles of various technologies, conducts horizontal quantitative comparisons of mainstream technologies from four dimensions: detection accuracy, computing power cost, black-box adaptability, and judicial evidence collection effectiveness, and clarifies the practical bottlenecks of the existing protection system. The multi-layer technology defense and intellectual property system collaborative integrated protection framework is constructed in this article. It can achieve full process risk management of training data, reduce the probability of copyright infringement of large models, improve the efficiency of infringement tracing and evidence collection, and balance the protection of creator rights and the needs of AI technology innovation and development. This study can provide a complete theoretical basis for the construction of data compliance pipelines for large model enterprises, the development of industry compliance technical standards, and the improvement of digital supervision of intellectual property rights.

Keywords: Generative AI, Copyright risk, Digital watermark, Member inference attack, Quasi-statutory license

1. Introduction

In recent years, generative artificial intelligence has entered the stage of large-scale implementation. The mainstream large-scale models represented by GPT -4 and LLaMA 3 require Petabyte (PB) scale training datasets for pretraining, with training tokens often reaching the trillions of levels. The training process consumes a large amount of GPU computing power and capital costs, and the authorization mode of traditional copyright systems is difficult to match this type of large-scale data usage scenario [1]. The reality of a large number of unauthorized bulk uses of copyrighted works in

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38155

the AI industry has given rise to large-scale copyright infringement, and related lawsuits have occurred frequently worldwide [2].

From a legal perspective, existing copyright systems in various countries are difficult to adapt to AI training scenarios. The scope of fair use in China's copyright law is narrow, and its exhaustive legislative model cannot cover the commercial and large-scale needs of AI training, leading to difficulties in legal application. Although the European Union's (EU's) Text and Data Mining (TDM) mechanism has established exception rules, the use of the rights-holder opt-out mechanism increases the compliance costs for AI developers, and the prevention and control capabilities are weak [3]. Relying solely on post-litigation accountability cannot solve infringement problems at the source, and it is necessary to rely on technological means to achieve pre-risk control.

From the current research status, academic research is relatively scattered. Legal research focuses on exploring models such as rational use expansion, statutory licensing, and collective authorization, but these studies generally lack technical implementation support. Technical research often focuses on optimizing a single algorithm; for example, Hans et al. studied the detection method of copyrighted content in language model training data based on the member inference attack paradigm and proposed the DE COP method [3]. Shumailov et al. studied the output distribution degradation problem of AI models under recursive data generation training based on the phenomenon of model collapse, and proposed an analytical approach using long-tail information loss to indirectly trace the source of training data [4]. However, this method has not yet formed a complete compliance pipeline covering the entire lifecycle of data, and lacks a systematic framework for copyright protection research, resulting in significant research gaps.

In response to the above issues, this study takes the complete process of training data as the logical mainline, integrates multiple cutting-edge technologies, establishes a more structured AI training data copyright protection system, fills the gap of fragmented research in the intersection of law and artificial intelligence, and constructs an integrated protection framework that combines multi-level technical defense and intellectual property system coordination. It combines the design of quasi-statutory licensing, collective management, and other systems with the technical defense system, providing a solution that combines theoretical depth and practical feasibility for the healthy development of the generative AI industry.

2. Training the full process copyright risk and technical protection system

2.1. Pre-filtering

2.1.1. Access control and rights declaration protocol

The access control and rights declaration protocol constitutes the most fundamental filtering layer in the data input process. Its core concept is to respect and enforce the rights declaration set by the content publisher through technical means, and regulate the data acquisition behavior from the source. This level of protection mainly relies on three progressive mechanisms.

Firstly, there is the Robots. txt protocol, which is a widely accepted web crawler access standard [4]. AI developers should follow the instructions in the robots. txt file in the website root directory when crawling data, which can clearly specify which directories or pages are prohibited from being crawled. The code of conduct of the EU AI Act clearly states that failure to comply with access restrictions such as robots. txt not only violates industry practices for web crawlers, but also constitutes a disregard for the copyright retention intentions of website owners [3]. Secondly, there is a machine-readable rights retention mechanism. Article 4 of the EU Digital Single Market

DOI: 10.54254/2755-2721/2026.38155

Copyright Directive innovatively introduces an opt-out mechanism, allowing rights holders to retain their right not to use their works for text and data mining through machine-readable means, such as declaring in webpage metadata or website service terms [4, 5]. This provides a proactive and clear technical means for rights holders to express their authorization intentions, and AI developers have a legal obligation to identify and respect such statements. Finally, there is the emerging commercial access control model of pay-per-use crawling. This model fully returns control of data access to content owners, allowing them to set prices for AI crawlers' access. This not only grants creators commercial control over their data assets, but also opens up legal commercial channels for AI companies to obtain high-quality, authorized data, transforming potential infringement into compliant data transactions.

2.1.2. Content fingerprint and perceptual hash

Content fingerprinting and perceptual hashing techniques are used to identify known or highly similar copyrighted content and are the core means of multimedia data filtering. The underlying principle is that, unlike traditional encryption hash algorithms such as MD5 and SHA, perceptual hash algorithms such as pHash generate a robust fingerprint feature vector for content such as images, audio, and video [6]. Even with slight modifications such as compression, cropping, and color grading, the fingerprint remains highly similar, achieving approximate matching. At the application level, AI companies can establish or use existing copyright content fingerprint databases. In the data preprocessing stage, the massive multimedia data captured is compared with the fingerprints in the database. If the similarity exceeds the preset threshold, it is determined to be infringing content and filtered. The Content ID system of YouTube is a successful application example of this technology, which has achieved automated copyright management of massive video content worldwide [6].

2.1.3. Machine learning classifier

For content such as text that is difficult to match accurately with hashing, machine learning models can be used for intelligent classification and filtering. The underlying principle is to build a binary classification model and train it using a large amount of annotated copyright text and public domain text. The model learns to recognize the language style, structural features, specific entities, and other patterns of copyright text, in order to score the infringement risk of unknown text [6]. At the application level, companies such as Anthropic have developed classifiers for filtering dangerous content, and their technical framework can be transferred to the field of copyright filtering [6]. In the data cleaning stage, this classifier is used to perform batch scanning and classification of text data, automatically removing high-risk content. The accuracy of this method is highly dependent on the quality of training data and the accuracy of labeling, and it is difficult to precisely define the boundaries of ideas and expressions. In real industrial datasets, the false positive rate (FPR) of this type of copyright text filtering classifier can reach 0.05% -68.6%, and the false negative rate (FNR) can reach 0.3% -99.6%, indicating a prominent issue of misjudgment [7].

DOI: 10.54254/2755-2721/2026.38155

2.2. In-process training constraints

2.2.1. Digital watermarking and neural network steganography

Digital watermarking technology embeds copyright information invisibly into the output or internal parameters of the model, achieving copyright identification and traceability. The underlying principle is to draw on digital watermarking technology in the multimedia field and add an additional encoder-decoder network to the deep neural network model to be protected. The encoder embeds watermark images containing user identity, copyright identification, and other information into each output image of the model in a way that is imperceptible to the human eye. Simultaneously train a corresponding decoder network to extract watermarks from the output to verify copyright ownership [8]. At the application level, model developers can embed different watermarks for different users when distributing models. Once a copyright dispute or model abuse occurs, the watermark information extracted from the output content can be traced back to the specific model user, achieving copyright protection and traceability in multi-user scenarios [8].

2.2.2. Copyright confirmation and circulation based on blockchain

The use of blockchain technology to provide tamper-proof ownership for training data or models has been a highly anticipated direction in recent years. The underlying principle is that non-fungible tokens have unique, indivisible, and tamper-proof characteristics, and can serve as identity cards for digital assets. Casting the metadata of the dataset or AI model, such as creator, version, authorization scope, etc., into non-fungible tokens and recording them on the blockchain to form on-chain ownership certificates [9]. At the application level, data providers can authorize their high-quality datasets to AI companies in the form of non-fungible tokens for use. Every use, circulation, and profit distribution is recorded on the chain, forming a transparent and trustworthy copyright trading and traceability chain. This provides technical endorsement for the legitimate source of training data, but its effectiveness depends on the authenticity of off-chain data when it is on-chain [9].

2.3. Post-event testing and evidence collection

2.3.1. Member inference attack

Member inference attack is currently the most cutting-edge detection paradigm, aimed at determining whether a specific piece of data belongs to the training set of the model [10]. The underlying principle is that the model exhibits different memory features on the data it has been trained on and the data it has not seen before. For example, models usually have higher confidence in predicting member data, or there are differences in the probability distribution of word elements used during generation. The DE COP method is an innovative member inference attack method. It does not directly query the model, but constructs a multiple-choice question with the stem of a suspicious copyrighted text, and the options include the original text of the text and multiple AI-generated interpretations. If the model can select the original text significantly higher than the random probability, it proves that the text is likely to be in its training data. This method achieved a detection accuracy of about 72% on the black box model, far exceeding the traditional method's accuracy of about 4% [10].

DOI: 10.54254/2755-2721/2026.38155

2.3.2. Model collapse as indirect evidence

Model collapse itself is a negative phenomenon, but its mechanism of occurrence can provide clues for data traceability. The underlying principle is that when an AI model is trained using data generated by previous generations of AI models, the model crashes, causing the model to gradually forget the long-tail information of the original data distribution, that is, low-probability and diverse data, and the output tends to be single and repetitive [5]. At the application level, by comparing and analyzing the output distribution of a model with the distribution of a known human-created dataset, if it is found that its long-tail information is significantly missing and the output content is highly similar to that of a specific AI model, it can indirectly infer that a large amount of its training data comes from that AI model, and thus trace the legitimacy of its upstream data source [5].

2.4. Full process protection logic flowchart

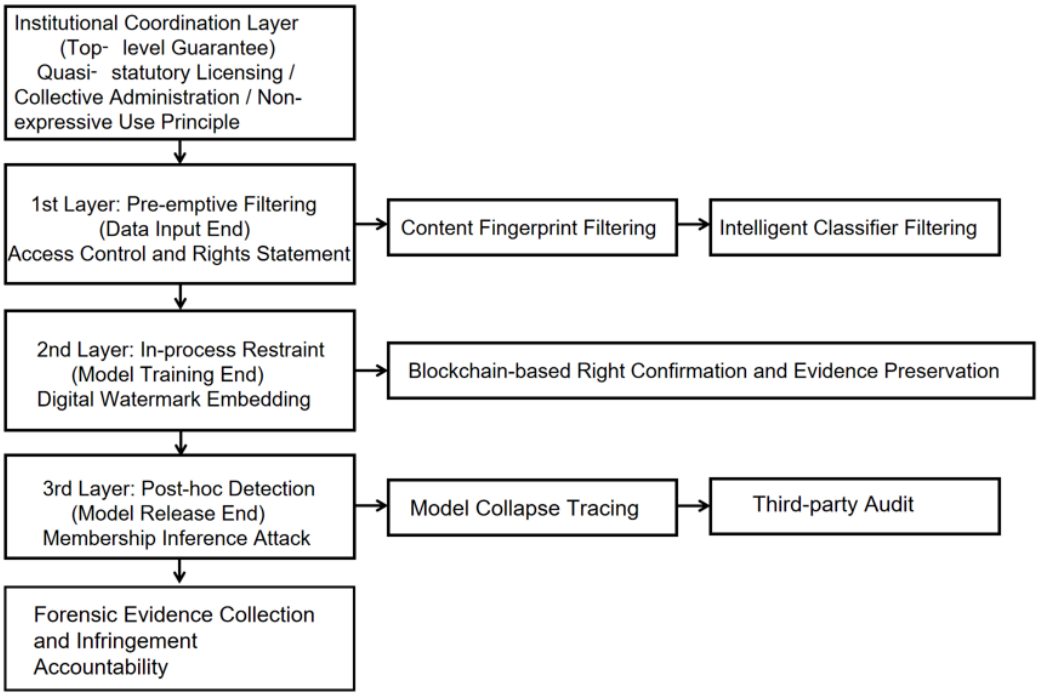

Figure 1. Full process protection logic flowchart (photo/picture credit: original)

Figure 1 presents a four-layer architecture consisting of institutional collaboration, pre-filtering, in-process constraints, and post-detection. Each layer forms a top-down protection chain to achieve full life-cycle copyright risk control.

3. Horizontal comparison of mainstream technical solutions

Table 1 presents a horizontal comparison of mainstream technology solutions.

DOI: 10.54254/2755-2721/2026.38155

Table 1. Comparison of the four key dimensions for the mainstream technology solutions mentioned above

Technical Proposal	Detection Accuracy	Computing Power Cost	Black Box Adaptability	Judicial Evidence Collection Effectiveness
Access Control	Low	Extremely low	High	Low, only preliminary evidence
Perceived Hash	High	Medium, need to compare massive fingerprints	High	Medium, need to prove the authority of the fingerprint database
Machine Learning Classifier	Medium, relying on the quality of training data	High, high training and reasoning costs	High	Medium, there is a possibility of misjudgment
Digital watermarking	High, under ideal conditions	Low to medium, parallel lightweight computing in the inference stage	A black box model that adapts to all callable output interfaces	Medium, it can only prove the similarity of a single output and cannot trace the training source
Copyright Confirmation Based on Blockchain	Extremely high, based on immutability on the chain	Medium, on-chain transaction costs	Universal across all scenarios, independent of model architecture	Highest, chain data cannot be tampered with, complete evidence chain
Member inference attack	Medium to high, about 72% under black box	Extremely high, requiring a large number of queries	High	Medium to high, provides probabilistic evidence
Model collapse	Low, indirect evidence	High, requires comparative analysis	High	Low, only for auxiliary clues

4. The triple reality bottleneck

4.1. Technical bottlenecks

The training data scale of generative AI reaches trillions of dollars. Taking cutting-edge models such as GPT -3 and GPT -4 as examples, GPT -3 requires about 355 GPU-years of computing power for pre-training, while GPT -4 requires about 25000 A100 GPUs to run continuously for 90 to 100 days to complete pre-training. The industry estimates that the training capital cost can reach tens of millions of dollars. The computational and time costs of real-time, fully covered copyright filtering for such a massive amount of data are currently unaffordable with current technology. Copyright law protects expression rather than ideas, and technology is difficult to make such precise legal judgments. Automated systems struggle to distinguish between protected original expressions, reasonable citations, and knowledge that has entered the public domain, resulting in both accidental and missed kills. In addition, there is currently a lack of standardized audit tools to verify whether

DOI: 10.54254/2755-2721/2026.38155

AI companies have truly implemented effective filtering, and regulators and rights holders can only rely on the company's self-reporting, creating a verification paradox [4].

The inherent limitations at the technical level will be transmitted to legal practice. Filtering technology cannot achieve 100% accuracy, and post-detection mostly only outputs probabilistic results, unable to provide conclusive evidence of infringement. The inherent inadequacy of evidence attributes directly leads to difficulties in obtaining evidence and determining facts at the judicial level, further causing practical difficulties in the division of tort liability and the calculation of damages.

4.2. Legal bottlenecks

The definition of whether AI training constitutes fair use or text and data mining exceptions varies among countries. The United States tends to use a broad range of transformative measures, while the European Union has established an opt-out mechanism. There are still legal gaps in China, which makes it difficult for companies to comply with global regulations. There is no legal consensus on whether the responsibility should be borne by the model developer, data provider, or end user when infringement occurs. The post-event detection technology provides mostly probabilistic evidence, and its probative value in court needs to be tested. At the same time, there is a lack of clear standards for calculating compensation amounts for massive data infringement.

4.3. Industrial bottlenecks

Building a full process protection system requires significant investment in technology and manpower, which may become a barrier to entry for startups and exacerbate the market landscape dominated by a few giants monopolizing data and models. Strict copyright protection will increase the cost of data acquisition for developers, which may inhibit innovation, while being too lenient will harm the interests of creators and undermine their creative enthusiasm. The chain of data collection, cleaning, labeling, and trading is extremely opaque, with a large amount of data being transferred multiple times and ownership relationships being chaotic, making it difficult to trace.

5. Building an integrated framework for technology and institutional protection

To overcome the above bottlenecks, it is necessary to go beyond a single technical or legal perspective and build a comprehensive protection framework that combines technology and institutional coordination.

5.1. Multi-layer technical defense system

This system emphasizes defense in depth, combining different technologies to form complementarity. The first layer is the entrance layer, which enforces access control protocols and combines pay-per-use crawling mode to ensure the legitimacy of data acquisition from the source. The second layer is the cleaning layer, which deploys a hybrid filtering engine of perceptual hashing and machine learning classifiers to perform hierarchical classification processing on massive data and efficiently remove high-risk content. The third layer is the kernel layer, which promotes digital watermarking and blockchain authentication technology, assigns trusted labels to high-quality, authorized data and models themselves, and achieves full lifecycle traceability. The fourth layer is the audit layer, which establishes an independent third-party audit mechanism and uses post-

DOI: 10.54254/2755-2721/2026.38155

detection techniques such as membership inference attacks to conduct random checks and compliance verification on published models, forming a deterrent.

5.2. Collaboration of intellectual property systems

Technology requires institutional guidance and guarantees to maximize its effectiveness. Firstly, a quasi-statutory licensing system should be introduced, drawing on the proposals put forward by Chinese scholars, to establish a quasi-statutory licensing mechanism that includes announcement, objection, and payment. AI companies make a public announcement before using publicly available works, and the rights holder may choose to withdraw during the objection period. Failure to raise objections is considered an implied license, but AI companies are required to pay reasonable compensation. This not only reduces the transaction cost of massive authorizations, but also guarantees the remuneration rights of rights holders. Secondly, a collective management and compensation distribution mechanism should be established to encourage the establishment of AI training data copyright collective management organizations, representing a large and dispersed number of rights holders to negotiate and distribute comprehensive authorization and compensation with AI companies, in order to solve the problem of unrealistic one-on-one authorization. Finally, the principle of non-expressive use should be clarified, and legislation or judicial interpretations should specify that if the use of works in AI training is only for extracting non-expressive information such as statistical patterns and language patterns, it can be included in the scope of reasonable use, leaving room for technological innovation.

6. Conclusion

The copyright risk of generative AI training data is a complex systemic problem, and relying solely on technical means or legal systems is insufficient to resolve conflicts. This article focuses on the ability boundaries and shortcomings of existing technologies in the complete lifecycle of pre-filtering, in-process constraints, and post detection. It proposes a four layer integrated protection architecture for institutional collaboration in pre filtering, in-process constraints, and post detection. This architecture fills the gap in systematic research in the intersection of law and artificial intelligence, as previous studies either focused solely on legal systems or single algorithms, bridging the gap between technological tools and intellectual property systems.

It is necessary to objectively understand the boundaries of existing technological protection. Currently, various detection and filtering methods are mainly suitable for text and image content. With the rapid development of multimodal large models, infringement detection and traceability of video and real-time speech data will become a key research direction in the future. Industrial practice requires collaboration among technology developers, legislative bodies, content creators, and AI companies to maintain a dynamic balance between protecting the legitimate rights and interests of creators and encouraging technological innovation, and to promote the healthy and sustainable development of the generative AI industry.

References

[1] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783

[2] Zhao, Y. (2026). Regulatory pathways for copyright infringement in generative AI training data. *Law Science, 14*(5), 128–137. https://doi.org/10.12677/ojls.2026.145142

DOI: 10.54254/2755-2721/2026.38155

[3] Duarte, A. V., Zhao, X. D., Oliveira, A. L., et al. (2024). DE-COP: Detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 11152-11165). PMLR. https://arxiv.org/abs/2402.09910

[4] Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(7557), 755-759. https://doi.org/10.1038/s41586-024-07566-y

[5] Zhang, H. (2024). Reflections and prospects on intellectual property protection for AI-generated content: From the "copyright path" to the "patent model". *Journal of Wuhan Transportation Vocational College, 26*(2), 43–50. https://doi.org/10.3969/j.issn.1672-9846.2024.02.006

[6] Kryrychenko, M., Mudryi, M., & Chaklosh, M. (2026). Copyright in AI pre-training data filtering: Regulatory landscape and mitigation strategies. *arXiv preprint arXiv:2512.02047*. https://arxiv.org/abs/2512.02047

[7] Layton, S., Medeiros, B. B. P., Butler, K., et al. (2026). AI wrote my paper and all I got was this false negative: Measuring the efficacy of commercial AI text detectors. In *Proceedings of the 2026 IEEE Symposium on Security and Privacy* (pp. 1652-1671). https://doi.org/10.1109/SP63933.2026.00088

[8] Liu, Y., He, H., Chen, F., et al. (2023). A traceable DNN model protection method based on watermarking neural networks. *Journal of Applied Sciences, 41*(2), 183–196. https://doi.org/10.3969/j.issn.0255-8297.2023.02.001

[9] Liu, C., Wang, X., & Jin, J. (2021). Innovative practices of blockchain technology in copyright management and trusted circulation. *New Technology Applications*, (S1), 67–68. https://doi.org/10.16045/j.cnki.rti.2021.s1.018

[10] Sun, D. (2025). A study on the system of fair use of generative artificial intelligence. *Asia-Pacific Research Forum, 1*(10), 32–52. https://www.hanspub.org/journal/apr