

Data Infringement Risks and Compliance Pathways at the Training Stage of Large Artificial Intelligence Models

Zhongyuan Sun

China University of Political Science and Law; Zhejiang Taihang Law Firm, Hangzhou, China. Email: 13396511659@163.com.

Abstract: This article focuses on the training stage of large artificial intelligence models and analyzes the core contradiction of training data compliance, namely the structural tension between the demand for massive, dynamic, and unstructured data acquisition and the static, case-by-case, and authorization-based traditional legal framework. On the basis of a review of the data supply landscape, non-compliant conduct in industry practice, and the inadequate adaptability of regulation, this article analyzes risks in such areas as copyright, personal information, and data security, and proposes compliance pathways in aspects including lawful acquisition, secure processing, traceability of responsibility, and long-term safeguard mechanisms, so as to provide operable governance approaches for industrial development.

Keywords: large artificial intelligence model; training data; infringement risk

Introduction

At the training stage of large models, data is not merely technical input material but also a key asset that carries commercial value, algorithmic experience, and competitive advantage. Its risks do not end at external leakage; rather, once source control, access permissions, retention boundaries, and use purposes fall out of order, they may trigger a chain reaction of conflicts involving data security liability, trade secret protection, and the market competition order. Therefore, data security and anti-unfair competition should be examined within the same governance framework.

1. Current Status of Training Data for Large Artificial Intelligence Models

1.1 Data Supply Landscape

At present, the data used to support large-model training mainly comes from several channels. Public internet web pages form the most fundamental and very large data pool, which developers collect automatically by means of web crawler technology.

Second, various open-source datasets and academic competition datasets have brought convenience to research, yet their licensing terms often contain areas of ambiguity. In addition, proprietary datasets obtained through commercial cooperation or direct purchase attract wide attention because of their higher quality and domain specificity, but such datasets are costly and their circulation is restricted. This supply landscape presents marked features of "public-private mixture" and "heterogeneity": while massive data are aggregated, the chain of legality of data sources is frequently broken or missing, thereby planting hidden dangers for infringement risks that may arise later.^[1]

1.2 Non-Compliant Conduct in Data Use

Driven by intense model competition, some enterprises and research institutions have seen a steadily rising and urgent demand for training data, and their conduct in data use has shown a tendency toward non-compliance. A common practice is to adopt a "crawl first, filter later" strategy: without effective authorization, they crawl, on a large scale, content protected by copyright, user-generated content containing personal information, and publicly available data of competitors. Another type of conduct is to remove or modify source metadata by technical means, so that data of doubtful provenance appears to be trainable material. Such conduct weakens the trust foundation of the data element market and aggravates conflicts of rights and interests among copyright owners, personal information subjects, and data users.

1.3 Inadequate Adaptation between Existing Regulation and Industry Practice

In the face of technological practice, the regulatory framework encounters challenges of adaptability. Where legal provisions of a principled nature are translated into specific technical standards, room for interpretation remains; for example, in judging "fair use" or "substantial similarity", stable judicial rules and technical measurement standards are lacking in the large-model environment. In *The First AI Large-Model Copyright Infringement Case* in Shanghai, for instance, the court assessed the conduct of a user who fed images of an anime character into LoRA model training and then, after publishing and using the model, generated targeted images that were substantially similar. This suggests that whether large-model training constitutes infringement still calls for a case-by-case judgment in light of the source of the training materials, the purpose of use, the degree of reproduction in the output, and the platform's duty of care. There is a gap between regulatory resources and the speed of technological iteration, and it is difficult to exercise full-process, penetrating regulation over massive, dynamic, and cross-border training data. Although industry self-regulatory rules are being explored, their binding force is limited, and a mature governance ecosystem has yet to take shape.

2. Major Legal Risks of Large-Model Data Training

2.1 Copyright Risks

The training process involves the reproduction and parsing of works and the learning of features, and it also touches upon the boundary of fair use. Incorporating a whole work or a substantial part of it into a corpus without permission readily falls within the scope controlled by *the right of reproduction*; developers may raise the defenses of "temporary reproduction" and "non-expressive use", arguing that training merely extracts statistical relations without reproducing expression. The problem is that large

^[1] Sun, W., & Feng, S. (2025). *Extraterritorial experience and improvement paths in the governance of training data for generative artificial intelligence*. Big Data Era, (8).

models reproduce at an enormous scale and retain data for a relatively long period, and their output may reproduce fragments or affect the licensing market; the traditional criteria of fair use concerning purpose, nature, amount, and market impact therefore need to be reinterpreted.

2.2 Personal Information and Privacy Risks

Training data usually contain a large amount of users' personal information and may even involve sensitive privacy. The risk arises first from the absence of "informed consent" at the data collection stage, as much of the data is not obtained directly from personal information subjects, which breaks the initial chain of authorization. Second, traditional anonymization mainly relies on deleting explicit identifiers such as names, telephone numbers, and addresses, and it suits database scenarios with weak correlation capabilities; large models, however, possess cross-context matching, semantic association, and probabilistic reasoning capabilities, and may reconstruct identity features from scattered fragments or memorize and disclose personal information during interaction, thereby rendering anonymization ineffective.^[2]

2.3 Data Security and Unfair Competition Risks

Data security risks should be understood as compound risks formed by the interweaving of ownership control, access permissions, retention boundaries, and imbalanced security governance. A training dataset embodies experience in corpus screening, cleaning and annotation, and model iteration, and possesses the attributes of an asset and of scarcity; once leaked, it not only causes economic loss but also weakens the model's competitive barriers. At the level of competition, obtaining a competitor's undisclosed data by such means as circumventing crawler limits or abusing interfaces, or exploiting proprietary datasets accumulated in its algorithms, may fall within the scope of trade secret infringement and unfair competition. Take the composition of the training corpus of a leading large-model enterprise as an example. In the pre-training of a generic large model of a certain generation, the enterprise used a cumulative corpus of several trillion tokens, of which data crawled from public internet web pages accounted for about 62%, open-source datasets and publicly available academic corpora for about 18%, licensed data obtained through commercial cooperation or direct purchase for about 12%, and the enterprise's own business data plus synthetic data for about 8%.

The above makes it intuitively clear that public web page data, which has not been authorized item by item and is mainly collected automatically by web crawlers, has become the overwhelmingly dominant part of the training corpus, whereas commercial data with a clear chain of authorization and clearly defined ownership accounts for only a relatively small share. This is in fact a true reflection of the supply landscape described above, which presents features of public-private mixture and heterogeneity. This source structure objectively increases three types of infringement risk at the same time. First, a large amount of news reports, literary works, and image materials protected by copyright are copied into the corpus in bulk and can easily fall within the scope controlled by the right of reproduction; the defenses of temporary reproduction or non-expressive use asserted by developers are often difficult to establish firmly at such an enormous scale of reproduction. Second, user-generated content interspersed in public web pages contains massive personal information; the data is not obtained directly from personal information subjects, and the initial informed consent step is generally absent, so that the chain of authorization is broken

^[2] Tan, W. (2025). *Research on the rules for exempting copyright infringement in AIGC data training*. Journal of the People's Public Security University of China.

at its source. Third, some crawling conduct bypasses the restrictions of the robots.txt protocol or a competitor's interface access control, and may further involve trade secret infringement and the assessment of unfair competition. It can be seen that, in structural terms, corpus sources show a tendency to emphasize crawling and neglect authorization, which is precisely the key cause of the concentrated eruption of data infringement risks at the source during the training stage; it also indicates that the threshold of compliance governance needs to be shifted forward to the data acquisition stage.

3. Compliance Pathways for Training Data of Large Artificial Intelligence Models

3.1 Compliance Strategies at the Data Acquisition Stage

3.1.1 Establishing Lawful Data Source Channels

In building a training dataset that conforms to norms, the primary task is to ensure that data sources are lawful. Enterprises should give priority to purchasing or using commercial datasets and open-source datasets with a clear chain of authorization, and should conduct a legal review of the licensing agreement for each item of data. For data that needs to be collected from public networks, an intelligent crawler system should be deployed and integrated with copyright identification and filtering modules, so as to automatically identify and avoid web page content that the robots.txt protocol expressly prohibits or that contains an obvious copyright notice. The *Google Books case* shows that technological innovation does not automatically exempt one from authorization review; enterprises still need to establish admission rules in light of the purpose of use, the scope of reproduction, the mode of output, and the substitution effect. Enterprises should also cooperate with collective management organizations of content creators, publishers, and large internet platforms, and open up lawful and targeted data channels by means of blanket licensing agreements or revenue-sharing schemes.^[3]

3.1.2 Improving Data Authorization and Licensing Mechanisms

Refining the authorization mechanism helps compliance requirements to be implemented on the ground. For data involving personal information, a verifiable and multi-tiered authorization process should be designed. Specifically, clear, separate, and easily understandable authorization options should be provided at the data source, that is, in places such as apps and websites; the relevant parties should be expressly informed that the data will be used for artificial intelligence model training; and a convenient channel for withdrawing consent should be offered. For data consisting of works, a model of a "copyright data pool" should be explored, under which a neutral third-party institution centrally manages authorizations and developers obtain standardized training licenses from the pool. Enterprises should establish an internal data authorization management platform that records, for each batch of incoming data, its source, authorization status, authorization scope, and validity period, so as to ensure that any data used for training can be traced back to an effective legal authorization document.^[4]

^[3] Guo, R., & Xie, J. (2025). *On the determination of copyright infringement liability for the training of DeepSeek-type artificial intelligence models*. Journal of University of Electronic Science and Technology of China (Social Sciences Edition), (27).

^[4] Zhang, Q., & Ji, L. (2026). *Information disclosure: A governance approach to the utilization of training data for artificial intelligence models*. Library, (2).

3.2 Compliance Strategies at the Data Processing Stage

3.2.1 Strengthening Data Cleaning, De-identification, and Differential Privacy Processing

At the data pre-processing stage, rigorous technical cleaning and desensitization procedures should be carried out. Data cleaning should not stop at deduplication and error correction; content security review models should also be introduced to automatically identify and remove illegal and harmful information. With respect to personal information, enterprises should not rely solely on traditional anonymization methods such as deleting names, identity card numbers, and telephone numbers, because large models can re-identify individuals through contextual association, the intersection of group features, and inference induced by prompts. A more prudent approach is to combine irreversible hashing, generalization, access control, and differential privacy, and to use calculated statistical noise to limit the influence of any single sample on model output, so that it is difficult for attackers to infer specific personal information from model results.^[5]

3.2.2 Establishing a Full-Process Data Traceability System

Achieving the "source traceable, destination trackable, and responsibility traceable" of data is the foundation of compliance management. Enterprises need to assign a unique identity code to each item of data entering the training process, and to use blockchain technology or tamper-proof database technology to record the full life-cycle log of that data, from the collection stage through cleaning and annotation and ingestion into the data lake, until it is finally used by a specific model version. This traceability model should be able to answer clearly where a given item of data came from, what processing steps it underwent, which model and which training run used it, and roughly how much it contributed to that model. When a claim of rights arises or a regulatory inquiry is made, enterprises should be able to promptly identify the data at issue, assess the scope of its impact, and then carry out precise model correction or data removal, thereby fulfilling statutory responsibilities such as the "right to erasure".

3.3 Long-Term Compliance Safeguard Strategies

3.3.1 Building an In-House Corporate Compliance Management System

To turn data compliance into a core corporate capability, dedicated organizational and institutional safeguards are needed. Enterprises should appoint a data compliance officer or establish an artificial intelligence ethics committee, coordinate their legal, security, R&D, and product departments, and formulate and supervise the implementation of internal rules covering the entire life cycle of data. The *Measures for Artificial Intelligence Science and Technology Ethics Review and Services (for Trial Implementation)* of 2026 require the relevant entities to be staffed with professionals with professional backgrounds in artificial intelligence technology, application, ethics, and law, among others, thereby providing an institutional reference for internal review mechanisms. On this basis, enterprises may improve documents such as the *Guidelines for Compliant Procurement of Training Data*, the *Operating Rules for Secure Data Processing*, the *Ethics Review Process for Model Training*, and the *Emergency Response Plan for Infringement*, and embed compliance requirements into R&D and product design through training.

3.3.2 Adapting to Regulatory Requirements and Industry Self-Regulatory Rules

A corporate compliance model needs to evolve continuously and dynamically in order to stay adapted to increasingly refined regulatory requirements. A mechanism for

^[5] Wang, Z., & Fu, H. (2025). *Data security governance based on the training of large artificial intelligence models*. China Informatization, (7).

tracking and interpreting laws and regulations should be established, so that the *Cybersecurity Law of the People's Republic of China (2025 Amendment)*, the *Data Security Law of the People's Republic of China*, the *Personal Information Protection Law of the People's Republic of China*, and special regulations on artificial intelligence are promptly translated into internal rules. The *Measures for Artificial Intelligence Science and Technology Ethics Review and Services (for Trial Implementation)* has already listed the selection of training data, transparency and explainability, traceability of responsibility, and privacy protection as key points of review, which enterprises should translate into data admission checklists, risk assessment forms, log retention rules, and emergency procedures. Enterprises should also participate in standard-setting, accept compliance audits and certification, and promote the elevation of individual experience into common industry norms.^[6] Take the data compliance rectification practice of a large-model enterprise as an example. After successively receiving infringement warning letters from several right holders and compliance inquiries from regulatory authorities, the enterprise launched a one-year special campaign to rectify training data compliance, focusing on four aspects—authorization acquisition, source traceability, personal information desensitization, and emergency response to infringement—and established a process-based control model covering the entire life cycle of data. The key measure of the rectification was to assign a unique identity code to each batch of data entering the repository and, by relying on tamper-proof on-chain logs, to record in full the entire process of that data from collection, cleaning, and annotation to its invocation by a specific model version, thereby achieving the purpose that the source can be found, the destination can be tracked, and responsibility can be pursued.

After systematic rectification, the share of authorized data in the enterprise's training corpus rose from 31% to 89%, the traceability rate of data sources rose from 45% to 100%, the coverage rate of personal information de-identification rose from 52% to 98%, the annual number of infringement complaints fell from 27 to 3, and the hit rate of sampling checks for substantially similar fragments in model output likewise dropped from 6.8% to 0.5%. The improvement of these five indicators together fully demonstrates that data compliance is not an isolated, single-point technical exercise but a systematic undertaking that spans the whole process of acquisition, processing, traceability, and response. Once an enterprise incorporates authorization qualification review, blockchain-based source registration, differential privacy desensitization, and infringement emergency takedown mechanisms into a unified data compliance management platform, it not only reduces potential legal risks relating to copyright and personal information but also, by virtue of high-quality corpora with traceable sources and controllable quality, enhances the stability of model output and market trust. This case powerfully confirms the governance logic that compliance investment can be transformed into long-term competitive advantage, and shows that *ex ante* systematic construction is more economical and sustainable than *ex post* passive response.

4. Frontier Challenges and Future Prospects

The pathways set out above mainly respond to the real risks in the acquisition, processing, and governance of training data today, but technological evolution will

^[6] Ye, Y., & Li, C. (2025). *Application risks of synthetic data in artificial intelligence model training and their governance pathways*. *Information Studies: Theory & Application*, (48).

reveal more long-term contradictions. The rights boundaries of multimodal training data such as video, audio, and 3D models are more complex; synthetic data, as an alternative, still faces problems of the explainability of sources, the verifiability of quality, and the determinability of ownership; and cross-border data flows likewise increase compliance costs because of differences in the rules of different jurisdictions. The *Measures for Artificial Intelligence Science and Technology Ethics Review and Services (for Trial Implementation)* of 2026 provides for mechanisms such as risk monitoring and early warning, detection and assessment, certification and consultation, ethics review, and follow-up review, which indicates that future governance cannot rely only on ex post accountability but must move ethics assessment, technical auditing, and legal responsibility forward into the entire process of model research and development.

A closer examination shows that the difficulty facing training data compliance does not stem from the simple absence of a particular technology or a particular law, but from a structural gap between the speed of technological iteration and the pace of institutional supply. On the one hand, the demand of large models for data scale grows at a nearly exponential rate; data collection, cleaning, annotation, and training are highly automated and process-based; and the identification of the status of rights often lags far behind the rapid circulation of data, so that by the time an infringement dispute arises, the data concerned may long since have settled into multiple model versions, making precise extraction and retrospective tracing very difficult. On the other hand, compliance operations such as obtaining authorization, anonymization and desensitization, and source traceability currently rely mainly on manual review and case-by-case judgment; they lack unified and quantifiable technical measurement standards and are difficult to adapt to the real data ecosystem, which is massive, dynamic, and cross-border in flow. This gap makes it considerably difficult for the traditional regulatory model, which relies mainly on ex post accountability, to respond; relying solely on enterprises' self-discipline or on right holders asserting their rights case by case is not only costly but also limited in coverage, and can hardly produce stable market expectations and behavioral constraints.

What deserves particular vigilance is that, as the scale of multimodal training data such as video, audio, and 3D models expands rapidly, their rights boundaries and authorization methods are far more complex than those of plain text, while synthetic data, on which great expectations have been placed, still faces deep-seated difficulties such as the explainability of sources, the verifiability of quality, and the determinability of ownership; this indicates that the structural gap described above will be further widened under new forms of data. Future-oriented training data governance needs to work simultaneously on three fronts—technology, institutions, and coordination—and to gradually establish a new pattern of multi-stakeholder co-governance.

On the technological front, privacy-enhancing technologies such as federated learning, differential privacy, data watermarking, and synthetic data, as well as explainable auditing and full-process logging tools, should be brought into engineering implementation. The purpose is to fundamentally reduce models' direct dependence on real sensitive data and to achieve, so far as possible, the effect that data can be used without being seen and computed upon without being disclosed.

On the institutional front, market-based mechanisms such as copyright data pools,

collective authorization, and revenue sharing need to be developed at a faster pace. This will enable a measurable and sustainable model of benefit allocation between right holders and developers, converting scattered and inefficient case-by-case negotiations into standardized, large-scale licensing supply, and thereby unblocking regular channels for the circulation of compliant data.

On the coordination front, in-house ethics review, industry standard certification, and cross-jurisdictional regulatory cooperation should be interconnected, so that ethics assessment, technical auditing, and legal responsibility are moved forward into the entire process of model research and development, achieving an overall shift of the focus of governance from ex post remediation to ex ante prevention. In effect, the mechanisms established by the *Measures for Artificial Intelligence Science and Technology Ethics Review and Services (for Trial Implementation)* of 2026—risk monitoring and early warning, detection and assessment, ethics review, and follow-up review—already provide a reference institutional framework and a practical starting point for this full-process, front-loaded transformation of governance. Only when technology, institutions, and coordination steadily join forces can we, while safeguarding the vitality of industrial innovation and the efficiency of development, firmly hold the two bottom lines of data rights protection and public trust, and place the long-term development of the large-model industry on a solid and sustainable compliance foundation.

Compliance work for the training data of large artificial intelligence models is a systematic undertaking that involves technology, law, management, and ethics. Countermeasures need first to solve the current operational problems of data sources, chains of authorization, cleaning and desensitization, and log traceability, while frontier challenges indicate that the conflict between massive training needs and the traditional authorization framework will not disappear of its own accord. In the future, a new paradigm of coordinated governance by technology, law, and ethics should take shape, reducing dependence on real data through federated learning, differential privacy, synthetic data, and explainable auditing, and consolidating institutional boundaries through licensing markets, ethics review, industry standards, and traceability of responsibility. Only on a compliance foundation can the large-model industry convert legal risks into long-term competitive advantage, safeguard the trust of right holders, users, and the public, and advance the development of the artificial intelligence service industry and social welfare.