

# Stratified evaluation and rights allocation of work utilization in generative AI training

*Xiaoru Chen*

Hunan Yanjing Law Firm, Hengyang, China

116978119@qq.com

---

**Abstract.** The large-scale utilization of works in the training of generative artificial intelligence models is exerting a structural impact on the institutional framework of copyright law. The functional positioning of works has shifted from end products for human appreciation to underlying production factors that drive the generation of model capabilities. This paper proposes the concept of "parametric utilization" to describe the new mode of utilization in which works are integrated into model capabilities through numerical representation and parameter optimization, and reveals its normative uniqueness through comparative analysis with reproduction, text and data mining, and transformative use. On this basis, taking the "degree of expression retention" as the stratification criterion, this paper attempts to construct an evaluation framework covering the three stages of "input – training – application", and puts forward a corresponding dynamic rights allocation scheme.

**Keywords:** generative artificial intelligence, work control, parametric utilization, whole-process evaluation, dynamic rights allocation

---

## 1. Introduction

In late 2023, *The New York Times* filed a copyright infringement lawsuit against OpenAI, alleging that it used a large number of journalistic works for the training of the GPT model, resulting in verbatim reproduction in the output content<sup>1</sup>. This case touches upon a deeper institutional dilemma. The defendant did not fully reproduce or distribute the original works; instead, it deconstructed the texts into training materials and extracted patterns of expression and logic of information organization from them. When the mode of using works shifts from market dissemination to model training, the existing legal framework shows obvious limitations in applicability.

Domestic judicial practice faces the same dilemma. The Guangzhou Internet Court characterized model training as an infringement of the right of reproduction and the right of adaptation, while the Hangzhou

---

<sup>1</sup> The New York Times Company v. Microsoft Corporation, No. 1:23-cv-1195 (S.D.N.Y. Dec. 27, 2023).

Internet Court advocated classified and stratified handling<sup>2</sup>. The divergence in judicial approaches indicates that the existing rule system has not yet formed a unified evaluation standard for this new type of utilization act.

The deep root of this dilemma lies in the misalignment of the institutional structure. The underlying logic of copyright law is the control of expression, whereby right holders obtain economic returns from the dissemination of works by controlling statutory acts such as reproduction and distribution [1]. Model training follows a completely different logic. Works are transformed into numerical representations to serve parameter iteration and the generation of model capabilities, rather than to disseminate expressive content to the public. Since the nature of the object of evaluation has changed while the evaluation criteria have not been adjusted accordingly, adaptation difficulties are inevitable.

Academic and practical circles have attempted three analytical approaches, each with its own limitations. The reproduction right approach incorporates digital reproduction in training into the scope of rights; however, reproduction is only a preliminary step of parameter optimization, and equating it with infringing use may lead to over-expansion of right boundaries. The fair use approach is constrained by the closed enumeration model of Article 24 of China's Copyright Law, as AI training acts are not explicitly included in the statutory exemptions. The licensing approach also faces operational obstacles, since obtaining authorization on a case-by-case basis is practically infeasible when faced with tens of millions of works [2]. The deep commonality of these dilemmas is that, no matter the angle of approach, existing conceptual tools can hardly provide an appropriate legal evaluation of work utilization in model training. This indicates that the problem may not lie in the deviation in the application of specific rules, but in the need to expand the basic cognitive structure of copyright law.

## 2. The predicament and root causes of the work control paradigm

The "work control" paradigm centers on the exclusive control of original expression, and its operation relies on three basic presuppositions: works are legal objects with clear boundaries; right holders can only control utilization acts explicitly enumerated by law<sup>3</sup>; and the institutional goal targets economic returns in the work dissemination market. The premise for the effective operation of this logic is that a stable correspondence can be formed between the modes of work utilization and the types of statutory acts. Once a mode of utilization neither belongs to existing act types nor is oriented toward the dissemination market, this structure becomes difficult to apply.

AI training precisely constitutes a typical sample of such "paradigm mismatch", which can be examined from multiple dimensions.

In terms of act types, training is a continuous process spanning data acquisition, parameter optimization, and model convergence. Reproduction in the sense of copyright law only occurs at the very front end, while subsequent links such as feature extraction and parameter adjustment are not within the scope of statutory act types [3]. As a result, the core interval of model capability generation becomes a regulatory gap.

In terms of the basis of imputation, the capability of AI models derives from the aggregation effect of massive data, and model parameters carry statistical patterns jointly contributed by hundreds of millions of training samples [4]. Under this technical structure, the marginal contribution of a single work is negligible,

<sup>2</sup> Guangzhou Internet Court, Civil Judgment (2024) Yue 0192 Min Chu No. 113; Hangzhou Internet Court, Civil Judgment (2024) Zhe 0192 Min Chu No. 1587.

<sup>3</sup> Copyright Law of the People's Republic of China (2020 Amendment), art. 10.

which is both difficult to identify and impossible to quantify. Claiming exclusive control over individual works lacks even the most basic factual premise.

In terms of market orientation, the protection target of the traditional paradigm is the dissemination market of works, while data in training does not enter the dissemination link, and the value of works is reflected in their contribution to model capabilities. The rise of the training data licensing market indicates that the path of interest realization for works has extended from "dissemination" to "input" [5]. Unfortunately, this emerging market happens to be beyond the vision of the traditional paradigm.

In terms of the time dimension, the traditional paradigm takes the time of the act as the evaluation benchmark, while the reproduction act in training and the generative output with market substitution effects are often separated by months or even years. In this case, the causal chain between the act and the damage is difficult to trace effectively [6].

In sum, the above dilemmas do not stem from technical deviations at the level of rule application, but from the fact that the work control paradigm itself is an institutional structure designed for content dissemination, while AI training is a technical process aimed at knowledge production. The fundamental differences between the two in terms of control objects, act types, and forms of interests make the application of the traditional paradigm structurally mismatched from the outset.

### **3. Perspective shift: from work control to capability regulation**

#### **3.1. Basic orientation of capability regulation**

Since the problem lies in the structural misalignment between the framework and the object, the response should not be limited to patching existing rules, but requires re-examining the footing of evaluation. This paper summarizes this repositioning as "capability regulation". Its core essence is to shift the focus of legal evaluation from "control of work expression" to "the complete process of works participating in the generation of model capabilities", taking the contribution of works to model capabilities as the basic fulcrum and the degree of expression retention as the stratification criterion [7]. This is not to abandon the traditional paradigm, but to supplement a new evaluation dimension in areas where it has insufficient coverage. Correspondingly, the evaluation unit expands from "act" to "process", the normative baseline adjusts from "exclusive rights" to "contributory interests", and the regulatory mode changes from "all-or-nothing" to gradient allocation.

#### **3.2. Conceptual definition of parametric utilization**

After the perspective shift, a basic factual question needs to be answered: in what way exactly do works participate in the training process? "Parametric utilization" is a concept proposed precisely for this purpose. It refers to the process in which work data, through technical processing such as numerical representation, feature extraction, and parameter iteration, is ultimately embedded in the model parameter structure and participates in the construction of generative capabilities in AI training [8].

Specifically, this concept can be understood from four levels. At the object level, works are transformed from perceptible expressions into computable data. At the behavioral level, data continuously participates in the iterations of forward propagation and backpropagation. At the process level, it covers the complete chain from data acquisition to model convergence. At the outcome level, after being abstracted and transformed layer by layer, expressions are no longer stored in a recognizable manner, but are integrated into model capabilities as parametric traces [9]. Compared with traditional utilization modes, parametric utilization has three distinctive features: it is a continuous process rather than an instantaneous act; its effectiveness derives

from the superposition of massive data rather than the contribution of a single work; and the form of expression shifts from "perceptible" to "unreadable parameters" [10].

### 3.3. Doctrinal distinction from similar concepts

Distinguishing parametric utilization from existing concepts in copyright law helps to establish its independent normative status.

The core of the reproduction act is the reproduction of expression on a tangible medium. After training is completed, the expression of works does not remain in the model in a perceptible form. Reproduction increases the number of copies of works, while parametric utilization updates parameter weights; the two have entirely different normative orientations.

The goal of text and data mining is to extract information from data, and the original data can be discarded after the analysis is completed<sup>4</sup>. The goal of parametric utilization, by contrast, is to form generative capabilities, and data is embedded in the model as statistical traces rather than existing in the form of copies.

Transformative use presupposes the expression of the original work as the object of operation<sup>5</sup>. The object of parametric utilization is numerical data rather than expression itself, and the two operate at different levels.

The temporary reproduction exception requires that the reproduction act has no independent economic significance<sup>6</sup>. Although reproduction in training is part of the technical process, it is precisely the basic link for the formation of model capabilities and the operation of the business model, and thus does not meet the conditions for applying this exception.

## 4. Whole-process stratified evaluation system

Traditional evaluation takes single, clearly-bounded acts as its object, while parametric utilization runs through the entire training process. Legal methodology requires that evaluation methods adapt to the structural characteristics of the object [11]. If the evaluation horizon is limited to the reproduction act at the entry point, the core link of value generation will always be outside the scope of evaluation. Since different stages of training bear different normative tasks, they need to be examined separately.

The input stage examines the acquisition methods of materials. Public domain works and materials without originality are not subject to copyright constraints, and the focus of the problem lies in works within the protection period. It should be noted that the legality of acquisition does not equal the legality of training utilization. Materials obtained by circumventing technological measures, breaching agreements, or hacking into non-public databases are illegal in terms of the acquisition act itself. As for whether public data scraping can be used unconditionally, no consensus has been formed across countries [12]. For works that are legally obtained and within the protection period, prior authorization is the most normative approach, but the cost of individual authorization is prohibitively high. In comparative law, the EU has adopted an "opt-out" model, and Japan has included non-commercial AI training in the scope of statutory exemption<sup>7</sup>. The common judgment behind these practices is that when individual authorization is not feasible, alternative institutional channels should be provided.

<sup>4</sup> Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the digital single market, OJ L 130, 17.5.2019, p. 92, art. 2(2).

<sup>5</sup> *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994).

<sup>6</sup> *Infopaq International A/S v. Danske Dagblades Forening*, Case C-5/08, [2009] ECR I-6569.

<sup>7</sup> Directive (EU) 2019/790, art. 4; Copyright Law of Japan (2018 Amendment), art. 30-4.

The training stage requires identifying the degree of expression retention. The abstract learning type only extracts general patterns such as grammatical structures, does not retain recognizable original expression, and its retention approaches zero. The feature retention type absorbs stylistic and structural features but does not store substantive content, and can independently generate brand-new content with similar styles. The expression memory type directly retains complete expressions due to overfitting, and can achieve accurate reproduction under specific prompts [13]. This classification provides a factual basis for the subsequent judgment of market impact.

The application stage unfolds around three progressive dimensions. Relevance is the prerequisite: only when the output forms a corresponding relationship with a specific work in content can the evaluation of market impact have a factual basis. Reproduction stability is the key: coincidental overlap has limited market impact, while the reproducible generation of substantially similar content will significantly raise risks. The degree of market substitution is the ultimate criterion: direct substitution is manifested in homogeneous functions and interchangeable uses, while indirect substitution is realized by compressing the derivative market space of right holders or weakening their data licensing revenue [14].

The integrated logic of the three-stage examination can be summarized as "three filtrations" and "one calibration". Specifically, it goes through the legitimacy threshold, the baseline of expression retention, and the terminal judgment of market substitution in sequence, and is then calibrated in combination with case-specific factors such as the public attribute of the training purpose and the scope of the substitution effect [15].

## 5. Dynamic rights allocation mechanism

### 5.1. Basic logic of allocation

The dual structure of traditional exclusive rights plus tort liability faces dual difficulties in this scenario. On the one hand, the contribution of a single work is difficult to discern, and right holders can hardly prove how their works are utilized and what damage is caused. On the other hand, the utilization of works in AI training is often in an intermediate state, and the either-or liability model either leads to vacancy of rights due to failure of proof, or over-regulates technical acts due to rigid liability [16]. The reform idea of "separation of three rights" provides a useful reference. Its core lies in diluting ownership and emphasizing utilization rights, shifting from the question of "who owns it" to a functionalist approach of "how to utilize it effectively"<sup>8</sup>. Based on this, three principles can be extracted: stage adaptation, intensity matching, and damage correspondence [17].

### 5.2. Input stage: delimiting legal boundaries

Materials obtained by circumventing technological measures, breaching agreements, or hacking into non-public databases lack legitimacy from the outset. Article 7 of the *Interim Measures for the Management of Generative Artificial Intelligence Services* already requires the use of data from legal sources<sup>9</sup>. For works with legal sources and within the protection period, the exclusive control of right holders should be respected. However, under circumstances of prohibitively high transaction costs, a transition from "absolute exclusivity"

---

<sup>8</sup> Central Committee of the Communist Party of China & State Council, Opinions on building a basic data system to better leverage the role of data elements (Dec. 19, 2022).

<sup>9</sup> Interim Measures for the Management of Generative Artificial Intelligence Services (Order No. 15 of Seven Departments including the Cyberspace Administration of China, July 13, 2023), art. 7.

to "conditional exclusivity" can be made, with alternative arrangements including collective licensing, statutory licensing, and right holder declaration mechanisms [18].

### 5.3. Training stage: gradient allocation

The abstract learning type has the weakest interest connection, so there is no need to set strict exclusive control, and right holders only retain procedural rights of notification and objection. The interest connection of the feature retention type is enhanced, and the focus of coordination lies in transparency and opt-out channels. Developers should disclose the types and source scope of copyrighted works included in the training dataset, and right holders have the right to opt out of subsequent training. Article 53 of the EU Artificial Intelligence Act already requires providers of general-purpose AI models to publish a summary of training data<sup>10</sup>. The expression memory type has the strongest interest connection, and prior permission should apply; in principle, training that may produce expression memory shall not be carried out without permission, and a compensation mechanism should be established for utilization that has already occurred.

### 5.4. Application stage: responding to market impact

The control checkpoint for the risk of content reproduction should be moved forward. Developers shall bear the obligation of output filtering and adopt technical measures at the model deployment phase to prevent outputs that can stably reproduce the expression of original works from entering circulation. Article 14 of the *Interim Measures* already provides for the disposal of illegal content<sup>11</sup>. The correction of market substitution damage varies according to its form. For direct substitution, related functions should be suspended or parameter correction should be carried out; for indirect substitution, the focus is on compensating right holders for losses suffered due to compressed market opportunities, and mandatory suspension of model operation is not a prerequisite.

Pecuniary relief needs to distinguish between two functions: damages and disgorgement of benefits. The former compensates right holders for actual losses suffered due to market substitution, while the latter returns the unjust benefits obtained by infringers through using works for training. When actual losses can be proved, damages shall apply preferentially; when actual losses are difficult to prove but the infringer has indeed obtained unjust benefits, disgorgement of benefits can serve as a supplementary remedy [19].

### 5.5. Supporting guarantee mechanisms

The transparency obligation is the premise for the exercise of rights. Developers shall disclose the basic categories and source scope of training data, indicate whether copyrighted works are included, and provide channels for right inquiry<sup>12</sup>. The opt-out mechanism is a means for right holders to achieve self-control: right holders can indicate whether their works can be used for training in a machine-readable format, and can also request removal from subsequent training after learning that their works have been included in the training dataset<sup>13</sup>. Risk testing is a key link in precautionary prevention: expression memory risk assessment shall be conducted before deployment to evaluate the stability and likelihood of the model reproducing the expression

<sup>10</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, art. 53(1)(d).

<sup>11</sup> Interim Measures for the Management of Generative Artificial Intelligence Services, art. 14.

<sup>12</sup> Regulation (EU) 2024/1689, art. 53(1).

<sup>13</sup> Directive (EU) 2019/790, art. 4(3); Regulation (EU) 2024/1689, art. 53(2).

of training works under specific conditions [20].

## 6. Conclusion

In the final analysis, the copyright disputes triggered by generative AI training lie in the changed functional positioning of works, which have shifted from content for human consumption to basic material driving model generation. When works become elements participating in capability generation, the normative structure centered on expression control can hardly fulfill the evaluative function.

This paper provides a set of preliminary analytical tools. "Parametric utilization" depicts the way works participate in training; "whole-process stratified evaluation" expands the evaluation horizon from single-point acts to whole-process review; "dynamic rights allocation" replaces the single "all-or-nothing" pattern with differentiated schemes. The three follow the same logic: the evaluation structure should match the actual form of the object of evaluation.

It should be noted that the analysis in this paper is limited to work utilization acts in generative AI training, and does not involve discriminative AI or the copyright ownership of generative content. The determination of the degree of expression retention also requires further intervention of interdisciplinary methods.

## References

- [1] Wang, Q. (2021). *Intellectual property law textbook* (7th ed.). Renmin University of China Press.
- [2] Long, W. (2024). Copyright infringement risks and resolution paths of generative artificial intelligence public legal services. *China Public Legal Service*, (4), 45–58.
- [3] Feng, X. (2023). Research on copyright issues of generative artificial intelligence. *China Copyright*, (2), 17–26.
- [4] Wang, S., & Zhang, J. (2024). Research on fair use of copyright in large model data training. *Journal of East China University of Political Science and Law*, (4), 87–102.
- [5] Yi, J. (2024). Judicial protection of data property rights and interests. *Law Application*, (6), 56–71.
- [6] Cai, Y. (2024). The application basis and rule construction of statutory licensing for machine learning copyright. *Intellectual Property*, (11), 63–79.
- [7] Liang, Z. (2025). The characterization of generative artificial intelligence output in copyright law and its liability rules. *Law Science*, (6), 102–119.
- [8] Liu, Y. (2024). *Peking University intellectual property review: 2024 volume*. Intellectual Property Press.
- [9] Wang, Q. (2023). On the copyright law analysis of artificial intelligence generated content. *China Copyright*, (5), 22–31.
- [10] Zhang, X. (2024). Research on personal information protection of generative artificial intelligence training corpora. *China Legal Science*, (6), 56–73.
- [11] Larenz, K. (2003). *Methodenlehre der rechtswissenschaft* (A. Chen, Trans.). The Commercial Press.
- [12] Cui, G. (2023). Legal characterization of internet anti-crawling measures. *China Law Review*, (6), 89–104.
- [13] Zhang, J., & Wang, S. (2024). Fair use of copyright in generative artificial intelligence training data. *Law Application*, (8), 72–88.
- [14] Wu, H. (2024). Copyright law issues of artificial intelligence generated content. *Legal Science*, (3), 135–148.
- [15] Wang, L. (2016). *Research on tort law*. Renmin University of China Press.
- [16] Wu, H. (2022). Institutional innovation and legal improvement of intellectual property law. *Studies in Law and Business*, (4), 89–103.
- [17] Calabresi, G., & Melamed, A. D. (1972). Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review*, 85(6), 1089–1128.

- [18] Xiong, Q. (2024). Statutory licensing system of copyright for artificial intelligence training data. *Law Science*, (8), 90–106.
- [19] Cui, J. (2012). *Research on unjust enrichment*. Renmin University of China Press.
- [20] Zhang, J. (2025). Algorithmic governance and risk regulation of generative artificial intelligence. *China Legal Science*, (2), 108–125.