Beyond Fair Use and Opt-Out: Forging a Hybrid Copyright Path for Generative AI in China

Chang Ge[1*]

[1] Chizhou University, Chizhou, 247000, China

* gechang9988@foxmail.com

https://doi.org/10.70695/10.70695/IAAI202503A9

Abstract

This paper addresses the global copyright infringement challenges posed by Text and Data Mining (TDM) during the training of Generative AI models. It seeks to develop a legal framework for China that conforms to international legal standards while supporting the growth of its domestic AI industry. Using a normative comparative legal methodology, the study examines the legal approaches of the United States, the European Union, and China—representing common law, civil law, and mixed legal traditions, respectively. Through analysis of legal texts, key cases, and scholarly discourse, the paper identifies strengths and weaknesses in each system's treatment of TDM. Findings indicate that U.S.-style "Fair Use" offers flexibility but creates legal uncertainty; the EU's DSM Directive provides clarity but may hinder innovation via its opt-out mechanism; and China's closed-list copyright exceptions are ill-suited to TDM, resulting in a legal vacuum. The study's original contribution is a "Hybrid Model" that integrates rule-based certainty with factor-based flexibility, proposing a TDM exception under Chinese copyright law conditioned on lawful access and incorporating "Transformative Use" as a key judicial criterion. This approach offers a legislative pathway for China and a globally relevant third way for balancing technological innovation with author rights.

Keywords Generative AI; Text and Data Mining; Copyright Law; Fair Use; Transformative Use

1 Introduction

The rapid advancement of Generative Artificial Intelligence is profoundly altering global economic and social frameworks. Central to this technology is large-scale TDM, which involves copying and analyzing extensive datasets to train AI models [1]. This process inherently conflicts with copyright's reproduction right, making TDM a central issue in international intellectual property debates. Leading legal systems have developed divergent responses: the United States employs a judicially applied "Fair Use" standard offering adaptability at the cost of predictability; the European Union has adopted a detailed legislative framework under the DSM Directive that ensures legal certainty but may inhibit commercial innovation [2];and China's rigid, closed-list system of copyright exceptions fails to accommodate TDM, creating significant legal risks for its AI industry [3]. This paper argues that China must avoid simply adopting foreign models and should instead craft a tailored legal framework that reconciles copyright protection with AI development. Subsequent sections will explore theoretical justifications for TDM exceptions, compare the three legal systems, propose a hybrid model for China, and conclude with implications for global AI governance.

2 Theoretical Framework: Justification for a TDM Copyright Exception

Granting a specific copyright exception for TDM should not be seen as an act that undermines the foundations of copyright law. Instead, it represents a necessary recalibration of the delicate balance between protecting authorial rights and promoting public welfare in light of profound technological change [4]. Compelling justifications for such an exception arise from three complementary theoretical perspectives: law and economics, legal and utilitarian philosophy, and established principles of international law.

2.1 Law and Economics Perspective: Overcoming Market Failure and High Transaction Costs

From a law and economics viewpoint, mandating individual licensing for every piece of data used in TDM presents a classic and insurmountable case of market failure [5]. The sheer volume and diversity of data required to effectively train a competitive generative AI model—often involving millions or even billions of individual works from disparate sources — make it logistically impossible and financially prohibitive to identify and negotiate licenses with every single copyright holder. The transaction costs associated with such an endeavor—including search costs to identify rightsholders, negotiation costs to agree on terms, and enforcement costs to ensure compliance — would be astronomically high. These costs create an insurmountable barrier to entry for all but the most heavily capitalized corporations, effectively stifling the very innovation that could otherwise generate immense social and economic value. A thoughtfully designed TDM exception acts as a crucial corrective to this market failure. By removing the need for atomized licensing, it dramatically lowers transaction costs, democratizes access to data for training purposes, and thereby fosters the kind of socially beneficial AI innovation that the market, left to its own devices, would otherwise prevent.

2.2 Utilitarian Philosophy: Non-Expressive Use and the Ultimate Goal of Copyright

From a utilitarian philosophical standpoint, it is essential to return to the ultimate purpose of the copyright system, which, as articulated in the U.S. Constitution, is to "promote the Progress of Science and useful Arts [6]." Copyright achieves this not merely by rewarding creators as an end in itself, but by incentivizing the creation and dissemination of new works for the public good. TDM's engagement with copyrighted works is fundamentally non-expressive; it does not seek to reproduce or publicly communicate the creative expression of the original works to a human audience. An AI model does not "read" a book for its plot or "view" a photograph for its aesthetic beauty. Instead, its purpose is to computationally analyze these works to extract unprotectable elements such as facts, data, patterns, linguistic information, and stylistic tendencies.

This type of use does not supplant the primary or secondary markets for the original creative works; a person seeking to train an AI on legal documents is not a lost customer for a John Grisham novel. Because TDM operates on a different functional and economic plane, it aligns perfectly with copyright's ultimate utilitarian goal of fostering the creation of new knowledge and derivative creativity. To prohibit such a use would be to protect the "letter" of copyright law while frustrating its fundamental "spirit".

2.3 Compliance with International Law: The Berne Convention's Three-Step Test

From the perspective of international law, a carefully tailored TDM exception can be designed to be in full compliance with the globally accepted framework of the Berne Convention's "Three-Step Test". This test, articulated in Article 9(2) of the Convention, serves as the international standard for permitting exceptions to the exclusive right of reproduction [7]. An exception for TDM satisfies all three prongs of this test:(a)It is limited to certain specific cases. The exception would not be a blanket license to copy but would be narrowly confined to the specific act of computational analysis of a work for the purpose of information extraction and model training.(b)It does not conflict with a normal exploitation of the work. As previously discussed, the market for TDM is distinct from the market for human consumption of creative works. The analysis of a work's statistical properties does not substitute for the experience of reading, viewing, or listening to it, and therefore does not harm the primary commercial value.(c)It does not unreasonably prejudice the legitimate interests of the authors. Because the use is non-expressive and non-substitutional, the prejudice to the author's legitimate economic and moral interests is minimal, if any. The societal benefit derived from enabling AI innovation, meanwhile, is substantial.

Therefore, crafting an exception for TDM is not a legally unsupported action but has a solid and defensible basis for legitimacy within the established framework of international copyright treaties.

3 Comparative Analysis: Inherent Logic and Structural Dilemmas of the Three Paradigms

3.1 The United States: The Flexible Boundary and Uncertainty Cost of "Transformative Use"

The U.S. legal system addresses the challenge of TDM through the doctrinal lens of fair use, a famously flexible and context-sensitive standard codified in Section 107 of the Copyright Act. Courts evaluate potential infringements using a four-factor balancing test, with the first factor—the purpose and character of the use—often being paramount. Within this analysis, the concept of "transformative use" has become the central pivot [8]. A use is considered transformative if it adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message. Landmark precedents like Authors Guild v. Google, which affirmed Google's project to digitize millions of books to create a searchable database, have strongly supported TDM as a quintessential transformative use, converting expressive content into a new, functional, and informational tool [9].

However, the very flexibility that makes fair use adaptable also generates its greatest weakness: significant and persistent legal uncertainty. The boundaries of what is "transformative" are not fixed but are constantly being re-litigated and re-interpreted by the courts. This case-by-case adjudication process means that AI developers, especially startups and smaller entities, operate in a state of ambiguity, facing the constant threat of costly and time-consuming litigation. The ongoing wave of lawsuits against major generative AI companies is a testament to this instability. This uncertainty imposes a heavy "risk premium" on innovation, potentially chilling investment and discouraging experimentation in the American AI ecosystem.

3.2 The European Union: The Certainty of Rules and the Practical Shackles of "Opt-Out"

In stark contrast to the American model, the European Union opted for legislative precision with its DSM Directive. Articles 3 and 4 of the Directive create explicit, detailed exceptions for TDM [10]. This approach provides a high degree of legal certainty, clearly delineating the rules of engagement. Article 3 carves out a broad exception for TDM conducted by research organizations and cultural heritage institutions for scientific purposes. Article 4 creates a more constrained exception for general TDM, including commercial uses, but subjects it to a critical condition: the exception does not apply if rightsholders have expressly reserved their rights in a machine-readable format (an "opt-out").

While this bifurcated system benefits public research, the commercial opt-out mechanism creates significant practical shackles. It effectively empowers large rightsholders to unilaterally withdraw their content from the data pool available for commercial TDM, leading to fragmented and incomplete datasets. This creates high compliance costs for developers, who must constantly monitor and respect these opt-outs, and it risks reinforcing the market dominance of incumbent data-rich corporations and large publishing houses, who can either hoard their data or license it at exorbitant prices. Consequently, the EU's quest for certainty may inadvertently stifle competition and innovation by making it harder for new entrants to access the comprehensive data they need.

3.3 China: The Structural Failure of a Closed-List Exception System

China's copyright framework cannot accommodate TDM, as it does not fall within any existing exception [11]. This legislative gap is not a minor oversight but a profound structural failure. It places China's entire AI industry in a precarious legal gray zone, forcing companies to operate with significant legal ambiguity and the constant risk of infringement claims. This uncertainty obstructs the development of a structured, reliable data economy, as the legal status of data inputs remains unresolved. For China to achieve its ambitions in AI, urgent legal reform is imperative to move beyond this outdated framework and create a clear legal basis for TDM that can support both domestic industry growth and compliance with international norms.

4 Institutional Construction: A Hybrid Model Path for China

To resolve its current legislative impasse, China should avoid a simple transplant of either the U.S. or EU model and instead innovate by developing a hybrid system that strategically integrates elements from both. This path would combine the statutory clarity of the European approach with the judicial flexibility inherent in the American system, creating a framework uniquely suited to China's legal culture and industrial policy goals. In the immediate short term, before legislative reform is complete, Chinese courts could provide temporary relief by creatively leveraging existing clauses in the Copyright Law, such as the principles of good faith or the prohibition of abuse of rights, to protect legitimate TDM activities. However, the definitive long-term solution requires the introduction of a new, dedicated TDM exception into the law.

The hybrid model proposed here is built upon two foundational pillars.

4.1 Pillar One: Establishing Legal Certainty Through Clear Statutory Rules

The first pillar of this hybrid model is designed to provide a strong baseline of legal certainty, giving the AI industry a predictable and stable environment in which to operate and invest. This would be achieved through the enactment of clear, ex-ante statutory rules. A key feature of this pillar would be the creation of a unified TDM exception that deliberately avoids the problematic bifurcation between commercial and non-commercial uses seen in the EU. This distinction is often artificial in the context of AI, where research frequently leads to commercial applications.

Crucially, this exception would be firmly conditioned on the prerequisite of lawful access. This means AI developers could only mine works that they have legal permission to access, such as content from the open internet, licensed databases, or lawfully purchased materials. This essential safeguard ensures that the TDM exception cannot be used as a shield for piracy or data theft. Furthermore, to foster a robust and competitive data environment, the rule should explicitly reject an EU-style opt-out mechanism for data that is already publicly available. Such a rejection is vital to prevent the "Balkanization" or fragmentation of the digital data pool, which would disproportionately harm startups and ensure a more level playing field for all innovators.

4.2 Pillar Two: Introducing Judicial Flexibility via the "Transformative Use" Principle

The second pillar complements the statutory certainty of the first by introducing guided judicial discretion, empowering courts to handle complex, novel, or borderline cases that may not fit neatly within the established rule. This would be achieved by formally incorporating the principle of "transformative use" as a key analytical tool for courts. This would not be a full, wholesale transplantation of the four-factor U.S. fair use test, which would be incompatible with China's civil law system. Instead, it would serve as a statutory directive for judges to consider the degree to which a TDM activity transforms the original work — from expressive content into functional data — as a central, though not necessarily sole, criterion in their analysis.

To guide this judicial inquiry and prevent arbitrary decision-making, the legislation could provide a non-exhaustive list of other relevant factors for consideration. These might include: (a) the fundamental purpose and character of the TDM activity (e.g., is it for building a new tool or for directly reproducing content?); (b) the nature of the copyrighted data being used; (c) the potential effect of the use upon the potential market for or value of the original copyrighted work; and (d) the technical necessity of using copyrighted materials to achieve the desired technological outcome. This two-pillared structure achieves the best of both worlds: it provides clear, predictable safe harbors for conventional TDM applications while simultaneously empowering the judiciary to thoughtfully adapt the law to novel AI technologies and unforeseen uses on a principled, case-by-case basis. This ensures the legal framework is both stable and future-proof.

5 Conclusion

The global ascent of generative AI presents not just a technological revolution, but also a fundamental jurisprudential challenge that tests the adaptability of traditional copyright frameworks. The existing international models for addressing this tension each reveal significant trade-offs. The U.S. model prioritizes flexibility through its fair use doctrine but at the steep price of legal uncertainty and high litigation costs. The EU model ensures predictability with its specific directives but risks stifling innovation and competition through its restrictive opt-out mechanism. Meanwhile, China's current system, a relic of a past technological era, is functionally obsolete and inadequate for the needs of its world-class AI industry.The hybrid model proposed in this paper offers China a balanced and strategic path forward. By integrating the principle of transformative use—a concept refined in common law—into a structured civil law framework, this approach harmonizes statutory predictability with judicial adaptability. It would establish clear rules to protect authors' rights and provide legal certainty for developers, while empowering courts to intelligently adjudicate the complex issues that will inevitably arise at the frontier of AI. This carefully constructed synthesis would protect authors' core markets without impeding the technological development essential for national competitiveness. This approach not only addresses China's pressing domestic needs but also offers a valuable contribution to the global discourse on copyright and AI governance. It suggests a viable "third way"—a middle ground between the American and European poles — that could serve as a compelling template for other civil law countries navigating these same turbulent technological and legal waters.

Acknowledgement

This work was supported by the 2024 Chizhou University Teaching Quality Project (Grant No. 2024XJXTD03) and the Intellectual Property New Liberal Arts Research and Reform Practice Project (Grant No. 2022sx101).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Sag, M. (2018). The new legal landscape for text mining and machine learning. Journal of the Copyright Society of the USA, 66, 291.
2. Lu, W., & Wang, Y. (2024, April). Comparative study of TDM technology systems from an international perspective. In 2024 9th International Conference on Computer and Communication Systems (ICCCS) (pp. 1172-1177). IEEE. https://doi.org/10.1109/ICCCS61882.2024.10603045
3. Hua, J. (2022). Copyright exceptions for text and data mining in China: Inspiration from transformative use. Journal of the Copyright Society of the USA, 69, 123.
4. Carroll, M. W. (2019). Copyright and the progress of science: Why text and data mining is lawful. UC Davis Law Review, 53, 893.
5. Landes, W. M., & Posner, R. A. (1989). An economic analysis of copyright law. The Journal of Legal Studies, 18(2), 325-363. https://doi.org/10.1086/468150
6. Arcangeli, A. (2024). Metamorphoses of authorship in the age of the mechanical reproduction of texts. Journal of Early Modern Studies, 1-13.
7. Geiger, C., Gervais, D., & Senftleben, M. (2013). The three-step test revisited: How to use the test's flexibility in national copyright law. American University International Law Review, 29, 581.
8. Leval, P. N. (1990). Toward a fair use standard. Harvard Law Review, 103(5), 1105-1136. https://doi.org/10.2307/1341457
9. Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).
10. European Parliament and Council. (2019, April 17). Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Official Journal of the European Union, L 130/92. http://data.europa.eu/eli/dir/2019/790/oj
11. Copyright Law of the People's Republic of China (2020 Amendment). (2020). World Intellectual Property Organization. https://wipolex.wipo.int/en/text/565736

Biographies

1. **Chang Ge** teaching assistant at Chizhou University. His research direction is intellectual property law, with a particular focus on Generative AI and copyright law.

超越「合理使用」與「選擇退出」：中國生成式AI版權的混合路徑探索

葛暢[1]

[1]池州學院，池州，中國，247000

摘要：本文旨在探討生成式人工智能模型在訓練過程中，因文本和數據挖掘（TDM）所引發的全球性著作權侵權挑戰。本文旨在為中國建構一套既符合國際法律標準，又能支持其國內人工智慧產業發展的法律框架。本研究採用規範性比較法學方法，檢視了分別代表普通法、大陸法及混合型法律傳統的美國、歐盟與中國的法律途徑。透過分析法律文本、關鍵案例及學術論述，本文指出了各體系在處理TDM議題上的優劣之處。研究發現，美國式的「合理使用」原則雖具彈性，卻也帶來法律上的不確定性；歐盟的《數位單一市場著作權指令》雖提供明確性，卻可能透過其「選擇退出」機制阻礙創新；而中國的封閉式著作權例外清單則難以適用於TDM，從而造成法律真空。本研究的獨創貢獻在於提出一個「混合模式」，該模式整合了規則導向的確定性與因素導向的靈活性，建議在中國著作權法下增設一項TDM例外，並以合法存取為前提，同時納入「轉換性使用」作為關鍵的司法審查標準。此一進路不僅為中國提供了一條立法路徑，也為全球在平衡科技創新與作者權利之間，提供了一條具參考價值的第三條路。

關鍵詞：生成式人工智能；文本與數據挖掘；版權法；合理使用；轉換性使用

1. 葛暢，池州學院教學助理。研究方向是知識產權法，尤其關注生成式人工智能與版權法。