# Research on the Copyright Fair Use of Text Data Mining in Generative Artificial Intelligence Training

Jiayu Guo[1], Wei Lin[2,*] and Xuan Liu[3]
[1]Law School, Wenzhou University, Wenzhou, Zhejiang, 325035, China
[2]Civil and Commercial Law School, Southwest University of Political Science and Law, Chongqing, 401120, China
[3]Law School of Guangzhou University, Guangzhou University, Guangzhou, Guangdong, 511400, China
*

Keywords: Generative Artificial Intelligence, Text Data Mining, Copyright, Fair Use, Infringement Risks.

Abstract: This paper focuses on the fair use of copyright in the text data mining in generative artificial intelligence training, makes staged analysis the infringement risks of TDM, explores the reasonableness of the fair use system for TDM and proposes a localized construction strategy by drawing on the overseas legislative experience. In China, Article 24 of the Copyright Law of the People's Republic of China (2020 Amendment) is difficult to cover its subject, purpose and data scale requirements. In other countries, EU adopts a "dual-track system" to distinguish between scientific research and general purposes, Japan expands the scope of exemption through the "generalization + enumeration + coverage" model, and the U.S. expands the scope of exemption through the "Transformative use" principle with the help of case law. Based on this, China needs to clarify the boundaries of the fair use of TDM and balance the rights and interests of copyright holders and the development of the AI industry and establish a data security mechanism to promote a dynamic balance between technological innovation and copyright protection.

## 1 INTRODUCTION

As generative artificial intelligence (hereinafter referred to as "GenAI") technology transitions from code-defined to data-trained, a series of problems and challenges emerge gradually. GenAI relies on a large amount of data training and achieves automatic analysis and content generation with the help of text data mining (hereinafter referred to as "TDM") technology. The training data used by GenAI includes content that is not original or has entered the public domain, which is not subject to copyright restrictions, as well as a large number of works protected by copyright. The use of such data can easily lead to conflicts of rights and infringement disputes (Yao, 2024).

In recent years, scholars from all over the world have conducted active research on the relevant issues about TDM copyright and have come to different paths. At the same time, various countries have also successively introduced policies and regulations to express their attitudes on TDM copyright issues. However, there is a structural conflict between the closed fair use clause in Article 24 of the Copyright Law of the People's Republic of China (2020 Amendment) (hereinafter referred to as the Copyright Law) and the technical characteristics of TDM. On the one hand, the existing exceptions cannot completely cover the subject behaviour and the scale of TDM technology, which can lead to a dual dilemma for juridical practice, which is the subject limitation of the “personal use” clause and the rigidity of the “appropriate citation” quantity standard. On the other hand, China’s legislative level has not responded to international rule innovation yet. Neither has it established the case law rules of “Transformative use”, nor does it have a systematic design for the commercial TDM authorization mechanism and balance of interests, which restricts the selection of compliance paths for technology research and development (Chinese Government Website, 2021).

Based on the above conflicts and practical difficulties, this article intends to start from the perspectives of the comparative method, we will

* Corresponding author



Guo, J., Lin, W. and Liu, X.
Research on the Copyright Fair Use of Text Data Mining in Generative Artificial Intelligence Training.
DOI: 10.5220/0014360000004859

In *Proceedings of the 1st International Conference on Politics, Law, and Social Science (ICPLSS 2025)*, pages 232-240
ISBN: 978-989-758-785-6

explore the legal boundaries of TDM behavior under the current legal framework of China and the rationality of its application in the fair use system, in order to propose suggestions for improving China's TDM fair use system.

# 2 INFRINGEMENT RISKS ASSOCIATED WITH TDM

TDM is a collective activity involving multiple processes, divided into three stages: data collection, data processing, and data aggregation and output (Fan, 2024).

## 2.1 Infringement Risks in the Data Collection Stage

There is a high risk of infringement of reproduction rights during the data collection phase of TDM behavior. At this stage, large-scale text data is often automatically captured using web crawlers and other technological means. Although authorized or unprotected content can be legally collected, the actual collected data is often mixed data due to the algorithm's indiscriminate recognition of the data, and it is difficult to obtain usage licenses one by one, which can easily constitute infringement of the right holder's reproduction right (Fan, 2024). In particular, long-term storage of source text data for repeated calls to the behavior of more clearly considered a violation of the right to copy. In addition, the data collection process often needs to circumvent the "Control and utilization" technology protection measures, such as bypassing access restrictions, traffic monitoring, etc., which also constitutes a violation of the right to copy. Even short-term, indirect temporary copy, because it may cause the loss of work data and bring potential economic damage, more and more is included in the protection of the right to copy (Ma & Zhao, 2021). Therefore, in the data collection phase, TDM behavior faces a substantial legal risk of infringement of the right of reproduction.

## 2.2 Infringement Risks in the Data Processing Stage

In the data processing stage of TDM behavior, the original data is transformed into a structured form that can be recognized by the algorithm through data cleaning, data labeling and data collation, and then serves the subsequent analysis. However, the treatment at this stage may involve the adaptation, translation, modification and reproduction of the protected works, which may constitute a potential infringement of copyright. On the one hand, data cleansing often deletes non-target information such as advertisements, comments, and codes to delete, translate, and store the original work, the rights of reproduction, translation, adaptation and the protection of the integrity of the work are easily infringed. On the other hand, data marks may also infringe the right of deduction by changing the original expression form by adding labels or notes (Fan, 2024). In addition, data collation generates structured data through "Transcoding" and other means, which is highly homogeneous with the translation and adaptation of works from the perspective of external performance and internal mechanism, therefore, it may constitute a right to the right of adaptation, translation of the infringement (Ma & Zhao, 2021). In general, the automatic and deep processing characteristics of the TDM data processing stage make it easy to cause the risk of deductive copyright infringement without authorization.

## 2.3 Infringement Risks in the Data Aggregation and Output Stage

In the TDM behavior, the data summary output stage mainly includes the collation and external output of the analysis results, and there are multiple risks of copyright infringement. First of all, data aggregation does not usually constitute infringement if it only involves the Quantitative analysis and independent expression of the relationship between the original data, but if the content of the original work itself is selected and arranged, it may infringe upon the right of compilation of the copyright owner. Secondly, in the stage of data output, if the results containing the content of the original work or its adapted content are disseminated to the public through the network platform or other means, it may constitute an infringement of the right of information network communication or the right of broadcasting (Fan, 2024). In particular, if the expression content protected by copyright is embedded in the analysis results, its network release behavior is easy to touch the "Copyright law" and "Regulations on the protection of the right of communication of Information Network" the relevant provisions of the protection of the dissemination of property rights (Chinese Government Website, 2021 & Chinese Government Website, 2013). In summary, in the stage of TDM data collection, whether it is content

collation or achievement dissemination, it is necessary to be alert to the potential infringement of the right to compile works and the right to disseminate information network.

Although there are multiple infringement risks in TDM behavior, the balance between the practical needs of technological development and the legal value has led to the discussion of the rationality of its application of the fair use system. After clarifying the risk boundary, it is necessary to systematically demonstrate the legitimacy basis of legal exemption, which is the key link to solve the contradiction between technological innovation and copyright protection.

# 3 THE RATIONALITY OF THE TDM FAIR USE SYSTEM

## 3.1 The Realistic Demand for Technological Innovation

### 3.1.1 Institutional Barriers to Data Supply

The training of GenAI relies on massive text and data. However, the current copyright system forms dual restrictions. Firstly, according to the Copyright Law, the protection of citizens' works by law extends back to the author's lifetime and 50 years after their death. As a result, a large number of advanced works cannot be used for model training, and it's obviously difficult to meet the technical requirements of timeliness and technical diversity if we only rely on the texts that have entered the public domain (such as classical literature or early journal) (Fan, 2024). Secondly, it's difficult for the traditional copyright trading model of "prior authorization, payment for use" to meet the demand for massive data, which will establish an institutional barrier to technological innovation (Xie, 2024).

### 3.1.2 The Inevitable Choice of International Rule Competition

The development of GenAI has rebuilt the landscape of international competitive, which requires China to make changes to traditional authorization mechanisms. Nowadays, special TDM rules have been established in major jurisdictions. The United States has passed a theory named "Transformative use" to extend the scope of fair use. The EU sets exemption clauses for research institutions by introducing the Directive on Copyright in the Digital Single Market. Japan amends law to add exception for "computer information analysis" (EUR-Lex, 2019). International practice indicates that the fair use system can reduce the legal cost of technology research and development. If China adheres to traditional authorization mechanisms, it might lose institutional advantage in global AI competition.

## 3.2 The Realization of the Coordination of Legal Values

### 3.2.1 Extended Protection of Constitutional Rights

With the development of AI technology, the public no longer solely relies on individual reading as a way of acquiring knowledge. Instead, they increasingly choose the algorithms that can extract content and analyze knowledge based on their training data to meet their requirements of "reading". In this context, the traditional "Reading right" has shown an extension trend of instrumentalization, collectivization and digitalization, which is manifested in the new derivative right form of "Text mining right", that is, the right of the public to conduct technical analysis of legally obtained works (Chinese Government Website, 2018). By ensuring the acquisition of works and the utilization of information, the fair use system not only maintains the cultural rights as stipulated in Article 47 of the Constitution, but it also promotes the public value of knowledge dissemination, which forms a value loop with the legislative purpose of "encourage the dissemination of works" of the Copyright Law (Chinese Government Website, 2021 & Chinese Government Website, 2018).

### 3.2.2 The Dynamic Balance Between Rights Protection and Technological Innovation

TDM involves a game of three parties' interests: the exclusive right of the copyright owner, the data requirements of the development of GenAI and the citizens' right to acquire knowledge. The strict interpretation of traditional "Author centralism" and "Three-step Test" excessively expands the scope of control of the rights holder, resulting in limited data available for training. The fair use system applies to TDM behavior, giving the TDM subjects varying degrees of exemption and obligation to protect the interests of the copyright owner while meeting the requirements of the miner. This design not only breaks through the limitations of the "prior

authorization" pattern on data usage amount, but also avoids excessive erosion of rights through hierarchical obligations.

### 3.3 Correction Mechanism of Market Failure

#### 3.3.1 Breaking Through the Dilemma of Transaction Costs

GenAI training involves licensing a huge amount of work, and the traditional licensing model has a triple cost: cost of rights identification (confirming the ownership of massive works), negotiation cost (making a contract with dispersed rights holders) and supervision cost (ensuring compliance in use). Microeconomic analysis indicates that transaction costs in the scenario of massive data have become a substantial obstacle to the development of technology (Mas-Colell & Whinston et al, 1995). Fair use systems that allow the use of data under certain conditions without cumbersome authorization procedures simplify the process of data acquisition and authorization and reduce transaction costs, it makes it more convenient for mining people to obtain the required data, improves the operation efficiency of the market, and thus promotes the development of the market.

#### 3.3.2 The Institutional Response of Positive Externalities

TDM generates significant social benefits: the industrial upgrading of promoting the breakthrough of technology, increasing the efficiency of public access to information and so on. However, it's difficult for private research and development institutions to obtain these external benefits completely, which will cause a shortage of market investment. By lowering the threshold of obtaining data, the fair use system makes social benefits and private costs of technological research and development tend towards equilibrium (Chinese Government Website, 2018).

Although the fair use system has legal legitimacy, there is a structural contradiction between the closed legislative model in the current Copyright Law and the development needs of AI technology. The current situation of insufficient supply of the system, urgently needs to be addressed through comparative research and practical dilemma analysis to find a solution (Chinese Government Website, 2021).

## 4 INSTITUTIONAL CHALLENGES IN APPLYING TDM FAIR USE UNDER COPYRIGHT LAW ARTICLE 24

Article 24 of the Copyright Law adopts a closed enumeration model for fair use, listing only 12 specific situations, which lacks a targeted response to the application needs of TDM in the development of generative artificial intelligence (Chinese Government Website, 2021). Article 24, paragraph 1, subparagraph 1 (personal use), subparagraph 2 (appropriate citation), subparagraph 6 (teaching and research), subparagraph 8 (cultural institutions) and other provisions on fair use can not be met by TDM, as follows (Guan, 2024).

### 4.1 Dual Constraints in Article 24(1): Subject-Type Limitations and Purpose Restrictions on Personal Use

Article 24(1) of the Copyright Law provides that an individual's use of a work for the purpose of learning, research, or appreciation does not constitute infringement (Chinese Government Website, 2021). However, TDM is mostly completed by enterprises or scientific research institutions, and its technical operation involves complex system deployment and large-scale data processing, which can not be completed by an individual. Therefore, the subject of its use is clearly beyond the scope of the "Individual" as defined by the law (Fan, 2024). In addition, the main purpose of TDM is often directly related to commercial development, technology optimization, market competition and so on, which is difficult to classify as "Learning, research or appreciation" non-profit category. This makes it difficult to apply the clause to TDM behaviour in practice.

### 4.2 Article 24(2)'S Compliance Burden: Purpose Specification and Quantitative Thresholds for Appropriate Citation

Article 24(2) fair use clause of the Copyright Law allows appropriate citations only for specific purposes such as introduction, commentary, or exposition (Chinese Government Website, 2021). The purpose of using TDM is usually to serve model training or application system building by analyzing big data extraction patterns and trends, it is not about

"Introducing", "Commenting" or "Describing" the work of others (Ma & Zhao, 2021). At the same time, the TDM training process often involves systematic, batch replication of thousands of works, far beyond the number of "Appropriate citations". Therefore, this clause does not provide an effective space for copyright exemption for TDM activities.

### 4.3 Functional Limitations of Article 24(6): Teaching/Research Exceptions in TDM Contexts

Article 24(6) fair use clause of the Copyright Law stipulates that teaching or research personnel may make a small number of copies or adaptations of works for teaching or research purposes (Fan, 2024). However, the application of TDM has already gone beyond the traditional teaching and scientific research, and has penetrated into the digital transformation process of many industries, such as medical, finance, manufacturing, and media. The purpose is not limited to classroom teaching or academic research. At the same time, the main body of TDM operation includes not only scientific researchers, but also enterprise engineers, technical teams and other groups. It is therefore difficult for this provision to cover TDM conduct in practice.

### 4.4 Regulatory Obsolescence: Article 24(3)(4)(5)(8)'s Incompatibility with Evolving TDM Requirements

Paragraphs 3,4,5 and 8 of Article 24 of the Copyright Law establish exemptions for the reasonable reproduction of specific works by the media and for libraries to preserve copies of the collection, respectively, however, in the specific application, it is faced with the limitations of the type of work and the purpose of use (Fan, 2024). In order to protect their commercial interests, media and publishing organizations often set up technical and legal barriers to API services and data interfaces to restrict the use of TDM. Although libraries and other cultural institutions are allowed to copy works for preservation purposes, it is difficult to cover the systematic and functional data mining tasks required by TDM. This too narrow use of purpose setting, in fact, weakened the library to fulfill the social functions of knowledge services and promote learning (Fan, 2024).

To sum up, Article 24 of China's Copyright Law imposes great restrictions on the fair use of TDM in terms of the system of provisions, the object of application, the purpose of use and the way of use, it is difficult to respond to the realistic demand for the legitimacy of big data mining in the context of the current development of artificial intelligence (Chinese Government Website, 2021).

In the face of the dilemma of the lack of localization rules, it is of great reference value to learn from the experience of foreign legislation. The United States, Europe, Japan and other major jurisdictions have constructed TDM rule systems through different paths, and their system design logic and implementation effect provide a multidimensional mirror for China's rule innovation.

## 5 EXTRATERRITORIAL PRACTICE OF THE TDM FAIR USE SYSTEM

### 5.1 European Union

#### 5.1.1 Current Status of Legislation

Article 3 and Article 4 of the EU Digital Single Market Copyright Directive (hereinafter referred to as "Copyright Directive") provide for "text and data mining for scientific research purposes" and "exceptions or limitations to text and data mining" respectively, i.e., a "two-track system" is adopted. The "two-track system", which distinguishes between scientific research purposes and general purposes, includes TDM in the scope of fair use (EUR-Lex, 2019 & Bao & Xiao, 2025). Liu Xiaochun pointed out that although there are relevant exceptions in the Directive, the scope of application is narrow and the conditions are strict, and it fails to completely solve the problem of the legality of data training behavior (Liu, 2024). In addition, the Copyright Directive also sets up an "opt-out" mechanism for copyright owners (EUR-Lex, 2019). However, Quintais points out that this "opt-out" mechanism exacerbates the imbalance of rights due to the lack of technical standards. He argues that the current opt-out mechanism does not solve the problem of creators' remuneration, and that collective bargaining and statutory licenses are needed to restructure the distribution of benefits (Quintais, 2025).

#### 5.1.2 Causes

In order to solve the legislative differences among member states and promote the modernization of copyright and related rights in the digital era, the EU

has formulated a unified TDM rule, i.e., the Copyright Directive (EUR-Lex, 2019). From a practical point of view, TDM technology has a key role in the field of scientific research, which can accelerate scientific discovery and help technological innovation. The EU expects to use this system to open up space for researchers and AI developers to use data legally and promote scientific research and technological innovation. At the same time, in order to safeguard the interests of copyright holders and to avoid overuse of their works to the detriment of their rights and interests, an "opt-out" mechanism for rights holders has been established.

## 5.2 Japan

### 5.2.1 Current Status of Legislation

Japan adopts the legislative model of "generalization + enumeration + underlining", and the Copyright Law of Japan has formed a system of copyright restriction rules on artificial intelligence technology with Article 30-4 as the core. This article, in conjunction with Article 47-5, includes information analysis behavior within the scope of fair use (Japanese Law Translation, 2021).

Article 30-4 establishes general criteria for determining non-appreciative use, with two specific lists of circumstances that qualify as non-appreciative use of a work and a supplementary explanation of the concept by means of an escape clause. The first paragraph of Article 47-5 is a general provision that establishes general criteria for the analysis of computer information for the small amount of use of a work, with the first and second subparagraphs of the first paragraph listing two situations that qualify for the general provision, and the third subparagraph serving as an underpinning provision.

### 5.2.2 Causes

Japan was the first to implement the concept of prioritizing the development of AI technology by expanding the copyright fair use system through legislation in order to give machine learning a break (Xie, 2024). Japan considers that TDM behavior is mainly for the purpose of obtaining information and knowledge in the data, and is not directly used for the enjoyment of the work itself, and does not cause substantial damage to the core interests of the copyright owner, and therefore gives TDM a wider scope of application, expecting to vigorously promote the rapid development of AI and related technologies through the construction of a lenient legal environment.

## 5.3 United States of America

### 5.3.1 Balance of Interests in Technological Innovation Orientation

The United States, as a case law country, does not have a statutory exemption specifically for TDM, but instead relies on judicial precedent to interpret the four-factor rule of fair use in Section 107 of the Copyright Act expansively (U.S. Copyright office, 1976). The four elements refer to purpose and nature (i.e., whether it is for non-profit purposes and whether it is commercial in nature), nature of the work (i.e., whether the work used is a copyrighted work), weight (i. e., the amount of content of the work used as a proportion of the complete work), and value and market (i. e., the extent to which the use of the copyrighted work has an impact on the value of the work or on the potential market for the work) (U.S. Copyright office, 1976 & Xiong, 2018). From the legal text of the principle itself, there is no explicit prohibition of commercial use.

In actual judicial practice, U.S. courts have also demonstrated a relatively tolerant attitude toward the AI training behavior of commercial subjects. If the TDM behavior meets the above four elements, especially if it is characterized by transformative use (i.e., adding new meaning or value to the original work), even if it is a commercial subject's use, it may be found by the court to be fair use. However, the transformative rule has its drawbacks, and Thongmeensuk, taking into account the U.S. jurisprudence (e.g., Andersen v. Stability AI), reveals the limitations of the "transformative" standard of fair use in the scenario of competing AI outputs, and argues that it is difficult to cope with the risk of market substitution solely relying on the principle of fair use, and that a layered design with exceptions is needed. It needs to be supplemented with a layered design of exception rules (Thongmeensuk, 2024).

### 5.3.2 Causes

The U.S. legal system is dominated by case law, and judicial precedent is central to the application of the law, a flexible legal tradition that allows for precise judgments on TDM behavior based on the circumstances of specific cases. The U.S. technology industry is highly developed and the pursuit of innovation is extremely strong. Therefore, the U.S. tends to give TDM users more space for their rights,

and through loose criteria for judging fair use, incentivize enterprises and scientific research institutions to carry out innovative activities by using TDM technology, so as to solidify its leading position in the global scientific and technological field.

# 6 STRATEGIES FOR BUILDING A TDM RATIONAL USE SYSTEM IN CHINA

The third amendment to the Copyright Law introduced a saving clause in Article 24(13)--"other circumstances provided for by laws and regulations", reserving space for China to create exceptions for text and data mining (Chinese Government Website, 2021). Therefore, the most feasible option is to use the touting clause as an interface to introduce a fair use clause for generative AI through the Regulations for the Implementation of the Copyright Law of the People's Republic of China (Revised in 2013), and to refine the relevant rules (Chinese Government Website, 2013).

## 6.1 Purpose of TDM Fair Use: Scientific Research and Knowledge Innovation

When China constructs rules for the fair use of TDM, it is not appropriate to limit the purpose of use to "non-commercial purposes", as the definition of "non-commercial use" is ambiguous in practice, and may restrict behaviors that have public interest objectives but have a certain degree of profitability. Therefore, the more intrinsically oriented "for the purpose of scientific research or knowledge innovation" should be the criterion for defining the legitimate purposes of TDM use. Due to the natural profit-driven nature of enterprises, the restriction of "non-commercial purpose" alone will not prevent them from building training datasets, but rather jeopardize the transparency of the training datasets and even form an industry monopoly. In the future, it is possible to consider "use for the purpose of scientific research or intellectual innovation" as the purpose of fair use of TDM, and to restrict secondary use to the initial market of the work, leaving the function outside the initial market to society (Guan, 2024).

## 6.2 Subject Scope of TDM Fair Use: Legitimate Access Holders

The subject of use should not be limited to "scientific research institutions", but should be extended to any subject that can legally access the work (e.g., public cultural research institutions such as libraries and market entities such as enterprises). At this point, emphasis should be placed on the legality of the means of access, requiring the relevant subjects to have "lawful access" to the work, not to bypass the relevant technical measures to access the work unlawfully, and not to presume that the work "may be reasonably used" just because it "exists openly on the Internet". Legitimate access to works includes, but is not limited to, access based on subscription behavior, access based on license agreements, access based on works being made available online for free (except where the right holder has made a reservation statement), access based on the needs of national development or the needs of the public interest of society, etc. (Guan, 2024).

## 6.3 Behavioral Requirements for Fair Use of TDM: Not Limited to “Replication” but Not Including “Propagation”

When China builds a fair use system for TDM, the behavioral elements should be defined as not limited to "copying", but not including "dissemination". Reproduction is the basic behavior of TDM, and the processing, analysis and storage based on the reproductions are also necessary for the implementation of the TDM process (Bao & Xiao, 2025). Therefore, when constructing a fair use clause for TDM, the elements of conduct should not be limited to "copying", but may include subsequent acts of analysis and research, including electronic transcoding, compiling, extracting, parsing, analyzing, reorganizing, etc. Moreover, the act of "dissemination" should be strictly excluded. The purpose of GenAI data acquisition and training is to analyze and learn, and ultimately to output a generated product. This is similar to the behavior of a natural person who reads, studies, etc., and eventually creates a work. The limit of the Copyright Act's tolerance for natural persons is to allow them to "study, research or enjoy". Similarly, in the case of GenAI, the extension of the behavioral elements to the behavior of information network dissemination would objectively result in "superhuman treatment".

### 6.4 Post-TDM Disposition of Technical Copies: Deletion or Transfer to Designated Institutions

The French Intellectual Property Code requires that technical reproductions made in the course of text and data mining should be placed at the disposal of a specific institution at the end of the research. Germany has similar provisions: "Once research is completed, follow-up and copies of source material should be removed and made inaccessible to the public (Chinese Government Website, 2013)." French and German practices reflect concerns about data security and apply to preventing copyright abuses arising from training data breaches. China can, on this basis and in conjunction with the characteristics of the network environment, establish a mechanism for centralized processing of TDM copies by a national-level trusted third party (such as an authorized agency of the State Copyright Administration) to prevent the leakage and dissemination of works and to establish a mechanism for safeguarding data security.

## 7 CONCLUSION

GenAI's TDM poses a systematic challenge to the current system of fair use of copyright. The research shows that TDM behavior faces the risk of copyright infringement at all stages of data collection, processing and output. However, the closed enumeration mode of article 24 of China's Copyright Law is difficult to adapt to the needs of technological development due to the limitation of subjects, the dislocation of purposes and the rigidity of behavioral elements. The experience of comparative law shows that the EU's "Dual-track system" distinguishes between scientific research and commercial use, Japan expands the boundary of unappreciative use through general clauses, and the United States achieves Dynamic equilibrium through "Transformative use" cases. The core of the system points to the dual goals of "Technology neutrality" and "Balance of interests". Based on the local practice, the construction of the TDM fair use system in China should focus on four aspects: First, the purpose element should anchor the purpose of "Scientific research or knowledge innovation" and break through the narrow limit of "Non-commercial purpose". Secondly, the scope of the subject should be extended to all subjects who legally obtain the work, and the dispute over the subject qualification should be resolved through the "Legal contact" rule, the requirements of behavior must cover the necessary technical behaviors such as Data pre-processing and structured processing, but strictly exclude the dissemination of use; Prevention of data leakage and secondary infringement of rights.

## AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

## REFERENCES


A. Mas-Colell, M. D. Whinston, J. R. Green, 1995. *Microeconomic theory.* Oxford University Press, 307-308

C. Guan, 2024. Research on the Fair Use of Copyright in Generative Artificial Intelligence Training: International Trends, Local Development and Rule Construction. *Publishing Research,* (12), 94-96.

Chinese Government Website, 2013. Regulations for the Implementation of the Copyright Law of the People's Republic of China, https://www.gov.cn/zhengce/zhengceku/2013-02/08/content_5423.htm

Chinese Government Website, 2013. Regulations on the Protection of the Right of Communication through Information Networks, https://flk.npc.gov.cn/detail2.html?ZmY4MDgwODE2ZjNjYmIzYzAxNmY0MTM5OTJiMjFkYjk

Chinese Government Website, 2018. the Constitution of the People's Republic of China, https://www.gov.cn/guoqing/2018-03/22/content_5276318.htm

Chinese Government Website, 2021. Copyright Law of the People's Republic of China, https://www.gov.cn/guoqing/2021-10/29/content_5647633.htm

EUR-Lex, 2019. 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 and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790

H. Fan, 2024. Construction of Copyright Exception Rules for Text and Data Mining in Digital Environment, *Library Work and Study,* (9), 27-30.

J. P. Quintais, 2025. Generative AI, copyright and the AI Act. *COMPUT LAW SECUR REV,* 56 (106107), 4.

Japanese Law Translation, 2021. Copyright Act (Act No. 48 of 1970), https://www.japaneselawtranslation.go.jp/ja/laws/view/4207#je_ch2sc3sb5at1

Q. Xiong, 2018. On the Judicial Standards of Fair Use of Copyright. *Law Science,* (1), 185.
S. Bao, D. Xiao, 2025. Copyright law response to Generative Artificial Intelligence Training Data: EU copyright exception rules and its enlightenment to China. *LIBRARY TRIBUNE,* 1-9.
S. Thongmeensuk, 2024. Rethinking Copyright Exceptions in the Era of Generative AI: Balancing Innovation and Intellectual Property Protection. *J WORLD INTELLECT PR,* 27 (2), 286.
U.S. Copyright office, 1976. Copyright Law of the United States (Title 17), https://www.copyright.gov/title17/92chap1.html#107
X. Liu, 2024. "Non-work use" in Data Training of Generative Artificial Intelligence and its Legal Justification. *Legal Forum,* 39 (3), 68.
Y. Xie, 2024. Copyright disputes and solutions of generative artificial intelligence work training. Chinese Editors Journal, (11), 38-41.
Y. Yao, 2024. On the Construction of Fair Use Rules of "Text and data mining". *STL,* (1), 32-37.
Z. Ma, L. Zhao, 2021. Impact of Text and Data Mining on Copyright Exception System and Countermeasures. *JNNU(SS),* 58 (4), 108-109.