HANKUN

Legal Commentary

October 28, 2026

BEIJING | SHANGHAI | SHENZHEN | HANGZHOU | WUHAN | HAIKOU | HONG KONG | SINGAPORE | NEW YORK | SILICON VALLEY | LONDON

AI Litigation Watch (Part I): Copyright Limits on Using Copyrighted Works to Train AI Models — A China–U.S. Litigation Perspective

Authors: Yan WANG | Lin YU | Haocheng XUE | Mingwei LI | Zifan ZHAO | Yu ZHANG

Over the past year, and particularly in the last two months, copyright disputes over artificial intelligence have reached landmark developments almost in parallel in China and the United States. For developers and service providers of large models—whose training data comes from a wide range of sources and passes through a complex processing chain—these disputes concern not only the use of specific works, but also the legal certainty surrounding model research and development, product iteration, and service operation.

In the United States, in *Bartz v. Anthropic*, Anthropic reached a USD 1.5 billion settlement with the author class in September 2025 over claims relating to the acquisition and retention of books from pirated sources; the court granted final approval of the settlement in July 2026, and the settlement has been described as "the largest copyright class action settlement in history." The settlement, however, did not end the controversy, and a new wave of lawsuits against Anthropic followed in quick succession: in January 2026, Universal Music Publishing Group and other publishers, having first sued in 2023, filed another action covering more than 20,000 works and claiming over USD 3 billion in damages; BMG and Round Hill Music followed suit. On August 28, 2026, Sony Music Publishing and Warner Chappell Music, together with 35 affiliated publishing entities, sued Anthropic and two of its founders in the U.S. District Court for the Northern District of California, alleging that Anthropic had acquired tens of thousands of lyrics and musical scores on a massive scale through BitTorrent downloads, web scraping, and other means to train its Claude model, and describing the conduct in the complaint as "one of the largest, most brazen, and ongoing thefts of intellectual property in history"[1].

In China, the Supreme People's Court issued the Opinions on Lawfully Adjudicating Disputes Involving Artificial Intelligence (《关于依法审理涉人工智能纠纷案件的意见》, the "AI Opinions") on September 7, 2026, systematically setting out rules on liability determination and the burden of proof in AI-related cases, while prudently leaving open the core question of how to characterize the unauthorized use of others' works to train large models. As for pending cases, actions that directly target the legality of training data

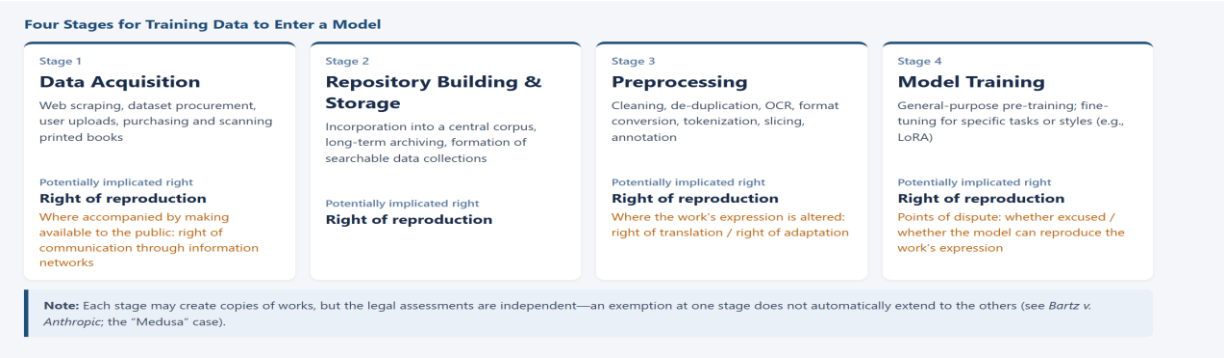
[1] *See Sony Music Publishing et al. v. Anthropic PBC*, No. 5:26-cv-09217 (N.D. Cal. filed Aug. 28, 2026). As of September 2026, this case and the actions against Anthropic referred to herein by Universal Music Publishing Group, BMG, and Round Hill Music have entered litigation, with no publicly available decisions on the merits to date.

for general-purpose models—such as iQIYI's suit against MiniMax and a suit by four illustrators against Trik AI—are in litigation, and no decisions on the merits have yet been made public; the "Medusa" case and the Hangzhou "Ultraman" case, by contrast, have already produced decisions in the context of model fine-tuning.

As the two markets where the global AI industry is most concentrated and rule developments are most closely watched, China and the United States will, to a considerable extent, shape how AI enterprises arrange their training-data use, legal defenses, and cross-border operations. Centered on the question of whether using works to train large models constitutes infringement, this article offers a comparative analysis of the rules and judicial practice of the two countries.

Breaking Down the Training Chain and Mapping the Rights Implicated

Any legal assessment of model training must begin with the training process itself. From a technical perspective, training materials typically pass through four stages—data acquisition, repository building and storage, preprocessing, and model training—before ultimately being distilled into model parameters. From a copyright perspective, each stage may produce copies of works, but the types of rights implicated and the applicable legal analysis differ from stage to stage. To determine whether a training activity infringes, one cannot ask in the abstract whether "training is legal"; the proper inquiry is what specific acts the developer has performed with respect to the works, and which exclusive right's scope of control each act falls within.



(Figure 1: Key Acts at Each Stage of the Training Chain and the Rights Potentially Implicated)

Specifically, the reproduction occurring at the data acquisition stage is the most intuitive: downloading and saving webpages, e-books, and images to servers, or disassembling printed books and scanning them page by page, both create new copies of works; and where the acquisition process is accompanied by making content available to the public—for example, uploading while downloading, or publicly sharing a curated dataset—the activity may further fall within the scope of control of the right of communication through information networks. Second, repository building and storage may appear to be merely static storage, but the source of the stored copies and the purpose of retention themselves require independent assessment—in *Bartz v. Anthropic*, the court held that Anthropic's downloading of works from pirated sources into a central repository intended for permanent retention would not automatically constitute fair use merely because some of the copies were later used for training, and allowed that portion of the claims to proceed to trial. Third, cleaning, de-duplication, format conversion, tokenization, and annotation at the

preprocessing stage principally implicate the right of reproduction; but where the processing alters the expression of the work itself—for example, translating a foreign-language work for training, or making an original adaptation of a work—the right of translation and the right of adaptation may be separately implicated. Finally, at the model training stage, reproduction issues are deeply intertwined with technical facts: during pre-training and fine-tuning, the system continuously reads and processes training materials and may create caches, intermediate copies, and the like, and whether such reproduction can be excused depends on the two countries' respective rules on limitations and exceptions; whether a trained model has "memorized" and can reproduce the protected expression of a work was the key factual issue in the second-instance "Medusa" decision on whether reproduction had occurred.

The United States: Case-by-Case Evolution Under the Four-Factor Fair Use Framework

U.S. copyright law provides no specific exception for text and data mining (TDM) or AI training; the fair use regime established by Section 107 of the Copyright Act, 17 U.S.C. § 107, is the principal defense for training activities. That section prescribes four factors: first, the purpose and character of the use, including whether the use is of a commercial nature, and courts also focus on whether the challenged use has a purpose or character distinct from that of the original work—that is, whether it is "transformative"[2]; second, the nature of the copyrighted work, i.e., whether the work leans factual or creative; third, the amount and substantiality of the portion used, i.e., how much of the work was used and whether its core expression is implicated; and fourth, the effect of the use on the potential market for or value of the original work. The four factors must be weighed holistically in light of the facts of each case and may not be applied in isolation.

Factor & Guiding Questions	Representative Case	Adjudicative Focus	Factor Assessment
01 Purpose and character of the use Commercial or not; whether it has a "transformative" purpose distinct from the original work	**Thomson Reuters v. ROSS**	Competing in the same arena as the rights holder; insufficiently transformative	**Against defendant**
	Bartz v. Anthropic	Training highly transformative, but pirated books assessed separately	**Favors defendant (training) / Against (pirated repository)**
	Kadrey v. Meta	Transformativeness acknowledged, but not the end of the analysis	**Favors defendant**
02 Nature of the copyrighted work More factual or more creative; the more creative the work, the thicker the protection	**Thomson Reuters v. ROSS**	Limited creativity in the headnotes	**Favors defendant**
	Bartz v. Anthropic / Kadrey v. Meta	Books at issue mostly creative works	**Limited weight**
03 Amount and substantiality of the portion used How much was used; whether the work's core expression is implicated	**Thomson Reuters v. ROSS**	Copying complete summaries exceeded what was necessary	**Against defendant (on appeal)**
	Kadrey v. Meta	Complete copying consonant with the transformative training purpose	**Favors defendant**
04 Effect on the potential market or value Whether it substitutes for the original; whether it harms markets the rights holder could reasonably develop	**Thomson Reuters v. ROSS**	Erodes the potential licensing market for training data	**Against defendant**
	Kadrey v. Meta	Distinguishes substitution / licensing-market / market-dilution harms; plaintiffs' proof insufficient	**Outcome-determinative**
	Bartz v. Anthropic	Training reproduction and pirated repository assessed separately	**Favors defendant (training); Against (pirated repository)**

Note: The four factors must be weighed holistically; no single factor is dispositive (17 U.S.C. § 107).
Colors indicate direction only: **Green = favors defendant** **Red = against defendant** **Orange = evidence-dependent**

(Figure 2: The Four Fair Use Factors and Their Points of Application in the Three Representative Cases)

[2] *See* 17 U.S.C. §§ 106, 107, U.S. Copyright Office, https://www.copyright.gov/title17/92chap1.html.

Three representative cases since 2025 approach the issue from different factors and, together, have gradually delineated the fair use boundaries of AI training.

I. *Thomson Reuters v. ROSS*[3]: AI Training Is No "Free Pass"

In that case, where the challenged use directly competed with the rights holder's legal research business, the court rejected ROSS's fair use defense. To develop a legal research tool competing with Westlaw, ROSS obtained approximately 25,000 legal question-and-answer items through a third party for training purposes, some of which were found to have actually copied headnotes written by Westlaw's editors (the court found 2,243 headnotes implicated). In February 2025, the U.S. District Court for the District of Delaware issued what is regarded as the first U.S. decision on the merits regarding fair use in AI training: because the parties served the same legal research market, the challenged use lacked a sufficiently transformative purpose, and the first factor weighed against ROSS; because the use was aimed at developing a competing product and risked eroding the potential licensing market for training data, the fourth factor likewise weighed against it. Following interlocutory appellate proceedings, the U.S. Court of Appeals for the Third Circuit affirmed in September 2026. That said, the system at issue was not generative AI, and the challenged use involved a legal research product in direct competition with the rights holder; the case's rejection of fair use should not be extended mechanically to the training of general-purpose large models.

II. *Bartz v. Anthropic*[4]: Source and Use Must Be Assessed Separately

In that case, the court held that reproduction for training a specific large language model constituted fair use, while separately assessing the acquisition and permanent retention of works from pirated sources. In June 2025, in the first decision directly addressing the pre-training of a general-purpose large language model, the U.S. District Court for the Northern District of California centered its analysis on transformativeness under the first factor[5] and evaluated the training chain stage by stage: the copies actually selected into the training set and created during cleaning, tokenization, and training were made in order to build entirely new model capabilities, were highly transformative, and constituted fair use; the digitization of lawfully purchased print books by "destroying the physical copy while retaining a digital substitute" likewise constituted fair use. Copies acquired from pirated sources and permanently retained in a central repository, however, were not sheltered by their subsequent training use—the transformative nature of a later training use cannot, as a matter of course, render the earlier

---

[3] *See Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.*, No. 1:20-cv-613-SB, Memorandum Opinion (D. Del. Feb. 11, 2025), https://law.justia.com/cases/federal/district-courts/delaware/dedce/1%3A2020cv00613/72109/770/; *see also* Memorandum Opinion (D. Del. May 23, 2025), https://docs.justia.com/cases/federal/district-courts/delaware/dedce/1%3A2020cv00613/72109/804 (certifying interlocutory appeal on originality and fair use and staying the district court proceedings). The U.S. Court of Appeals for the Third Circuit heard oral argument on June 11, 2026 and affirmed on September 29, 2026; as of this writing, the appellate opinion remained temporarily under seal pending the parties' proposed redactions. The docket is available at https://www.courtlistener.com/docket/70622297/thomson-reuters-enterprise-centre-gmbh-v-ross-intelligence-inc/.

[4] *See Bartz v. Anthropic PBC*, No. C 24-05417 WHA, Order on Fair Use (N.D. Cal. June 23, 2025), https://docs.justia.com/cases/federal/district-courts/california/candce/3%3A2024cv05417/434709/231; *see also* Order Granting Final Approval of Class Action Settlement and Judgment, Dkt. No. 680 (N.D. Cal. July 20, 2026), https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.680.0.pdf. The settlement's release does not cover future conduct or claims based on AI outputs.

[5] For the analytical framework of "transformative" use, *see Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023), https://www.supremecourt.gov/opinions/22pdf/21-869_87ad.pdf.

acquisition of the data fair use. The parties subsequently reached a USD 1.5 billion class settlement on the claims relating to pirated books, and the decision's reasoning has been directly invoked in the new wave of lawsuits brought by music publishers such as Sony and Warner. The case thus shows that the unit of fair use analysis is not "training" in the abstract, but each specific act along the chain.

III. *Kadrey v. Meta*[6]: Market Impact and Evidence Can Prove Outcome-Determinative

In that case, because the plaintiffs failed to adequately prove market harm from the training reproduction, the court upheld Meta's fair use defense on the relevant training reproduction. In the action in which 13 authors sued Meta regarding the training of its Llama model on books, the court acknowledged that the training use was strongly transformative and that the first factor generally favored Meta, but placed the center of its analysis on the fourth factor, distinguishing three categories of potential market harm: the model's direct regurgitation of the original works, producing market substitution; unauthorized training harming the licensing market for the works; and the model's generation, at extremely low cost, of large volumes of content that, while not infringing, competes with human works for sales and attention, diluting the market for the original works. The first two theories failed for insufficient evidence and for circularity concerns, respectively; the third was regarded by the court as the most significant, but the plaintiffs failed to present sufficient empirical evidence of the market impact actually caused by Llama. The court expressly confined its conclusion to the existing evidentiary record—had the plaintiffs presented fuller evidence of market dilution, the fourth-factor analysis might have come out differently. A subsequent order in March 2026 allowed the plaintiffs to add a contributory infringement claim relating to BitTorrent uploading, and the case remains pending.

Although the three cases reached different outcomes, they all demonstrate that fair use must be assessed on the specific facts: commerciality does not automatically defeat fair use; data acquisition and training reproduction must be assessed separately; and market harm must be established through evidence. Policy-level responses have also emerged, though the positions are not entirely aligned: the pre-publication version of *Copyright and Artificial Intelligence, Part 3: Generative AI Training*, released by the U.S. Copyright Office in May 2025[7], advocates case-by-case assessment based on the type of work, data sources, model uses, and market impact; in September 2026, the U.S. Department of Justice filed a Statement of Interest of the United States in *In re OpenAI Copyright Litigation*[8], arguing on the one hand that using works to train large language models should generally constitute fair use, while emphasizing on the other hand that the initial acquisition, the reproduction during training, and the generation of outputs to the public are distinct uses that must be analyzed

---

[6] *See Kadrey v. Meta Platforms, Inc.*, No. 23-cv-03417-VC, Order Denying Plaintiffs' Motion for Partial Summary Judgment and Granting Meta's Cross-Motion for Partial Summary Judgment (N.D. Cal. June 25, 2025), https://law.justia.com/cases/federal/district-courts/california/candce/3%3A2023cv03417/415175/598/; *see also* Order Granting Motion for Leave to File Fourth Amended Complaint; Denying Motion for Class Discovery, Dkt. No. 700 (N.D. Cal. Mar. 25, 2026), https://law.justia.com/cases/federal/district-courts/california/candce/3%3A2023cv03417/415175/700/.

[7] *See* U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training*, Pre-Publication Version (May 2025); U.S. Copyright Office AI Initiative, https://www.copyright.gov/ai/.

[8] *See* Statement of Interest of the United States, *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (S.D.N.Y. Sept. 1, 2026), https://storage.courtlistener.com/recap/gov.uscourts.nysd.640396/gov.uscourts.nysd.640396.1682.0.pdf.

separately—a stage-by-stage approach that echoes the analytical method of *Bartz v. Anthropic*.

China: Adjudication Rules Gradually Taking Shape Through Case-by-Case Exploration

Current Chinese law likewise provides no specific exception for text and data mining or model training. Article 24 of the Copyright Law of the People's Republic of China (《中华人民共和国著作权法》, the "Copyright Law") establishes the fair use regime on the basis of statutorily enumerated circumstances[9]; Article 7 of the Interim Measures for the Management of Generative Artificial Intelligence Services (《生成式人工智能服务管理暂行办法》, the "Interim Measures") requires providers of generative AI services offered to the public within China to use data and foundation models from lawful sources in their training-data processing activities and not to infringe intellectual property rights[10]. The AI Opinions, issued by the Supreme People's Court in September 2026, further refine the determination of liability and evidentiary requirements—in disputes over infringement of copyright by AI-generated content, where a developer raises a non-infringement defense, it must provide supporting evidence on matters such as the sources of its training data, the training process, and the manner in which the model operates[11]. On the characterization of unauthorized training using works, the AI Opinions prudently leave the question open for now, reserving room for judicial practice to explore[12]. Against this backdrop, the publicly available decisions have taken the lead in giving refined answers in the context of model fine-tuning.

I. The "Medusa" Case[13]: A Stage-by-Stage Determination Anchored on "Reproduction"

In its second-instance judgment, the Shanghai Intellectual Property Court conducted a stage-by-stage analysis of the training chain, holding that the creation of the training image set infringed the right of reproduction, while LoRA training itself did not necessarily constitute reproduction. A platform user captured more than twenty images of the "Medusa" character from the Battle Through the Heavens (《斗破苍穹》) animation, removed the backgrounds, uploaded the images to the platform, and trained and released two LoRA models. Taking as its doctrinal anchor the principle that "the essence of the right of reproduction is the right to reproduce the work," the second-instance court divided the chain

[9] *See* Copyright Law of the People's Republic of China (《中华人民共和国著作权法》), arts. 10, 24, 25, National Copyright Administration, https://www.ncac.gov.cn/xxfb/flfg/flfg_532/202103/t20210309_50530.html.

[10] *See* Interim Measures for the Management of Generative Artificial Intelligence Services (《生成式人工智能服务管理暂行办法》), arts. 2, 7, Cyberspace Administration of China and six other departments, July 13, 2023, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[11] *See* Supreme People's Court, Opinions on Lawfully Adjudicating Disputes Involving Artificial Intelligence (《关于依法审理涉人工智能纠纷案件的意见》) (Fa Fa [2026] No. 10), art. 12, September 7, 2026, https://www.court.gov.cn/zixun/xiangqing/511101.html.

[12] *See* Zhou Jiahai, Si Yanli, Qin Yuanming, Jia Yuhui & Zhang Yin, Understanding and Applying the SPC Opinions on Lawfully Adjudicating AI-Related Disputes (《〈最高人民法院关于依法审理涉人工智能纠纷案件的意见〉的理解与适用》), Intellectual Property Court of the Supreme People's Court, September 2026, https://ipc.court.gov.cn/zh-cn/news/view-6039.html.

[13] *See* Jinshan District People's Court of Shanghai, (2025) Hu 0116 Min Chu No. 2354, and Shanghai Intellectual Property Court, (2025) Hu 73 Min Zhong No. 1189; *see also* Does Using AI to "Replicate" the Medusa Anime Character Constitute Infringement? Second-Instance Judgment Announced in Shanghai's First Copyright Infringement Case Involving a Large AI Model (《用 AI"复刻"美杜莎动漫角色，算侵权吗？上海首例人工智能大模型著作权侵权案二审宣判！》), Shanghai High People's Court, April 29, 2026 (second instance), reposted on ciplawyer.cn: https://www.ciplawyer.cn/articles/159022.html; *see also* Liu Junhua, Reflections on Adjudicating "Medusa," Shanghai's First Copyright Case Involving a Large AI Model (《上海首例 AI 大模型著作权侵权案"美杜莎"案审理心得》), *Law Application* (《法律适用》), September 18, 2026, https://www.ciplawyer.cn/articles/159706.html.

into three stages for assessment: capturing the images to form an image set constituted infringement of the right of reproduction; feeding the image set into the model for LoRA training did not in itself necessarily constitute reproduction—LoRA learns weight offsets through low-rank matrices, and what the model "remembers" is "a form of low-rank statistical compression, rather than an encoding of the original images; these parameters cannot be pieced together to reconstruct the original images"; reproduction occurs only when a user issues a further prompt and the model generates an image substantially similar to the original work. This reproduction standard places technical facts at the core of adjudication, and also means that fact-finding on model architecture and training methods will play a pivotal role in such litigation.

II. The Hangzhou "Ultraman" Case[14]: A Fine-Grained Division Between Training Conduct and Platform Liability

The case distinguishes between users' directed training and the platform's provision of technical services; the court did not find the platform infringing merely for offering training functionality, but found liability for contributory infringement in light of its duty of care. In that case, the LoRA models at issue were trained by users who uploaded Ultraman images, selected a base model, and adjusted parameters, while the platform provided the base model, computing power, and distribution channels. In determining the scope of injunctive relief, the Hangzhou Internet Court of first instance made an exploratory determination on training at the input stage: absent evidence that the training aimed to exploit the original expression of the works, and where the training did not affect the normal use of the works or unreasonably prejudice the legitimate interests of the rights holders, the training on the data could be considered fair use. Although the appellate court, constrained by the scope of the appeal, did not comment further, this line of reasoning may serve as a reference for future cases. The more significant liability rule in the case lies in the division of liability: directed training is carried out by users, and a platform does not incur liability automatically by virtue of providing training functionality; its liability turns on whether it has fulfilled a duty of care commensurate with its information-management capabilities with respect to users' posting and dissemination of infringing content. The platform in this case was found liable for contributory infringement precisely because it had failed to discharge that duty.

III. The Trik AI Case and the *iQIYI v. MiniMax* Case[15]: Head-On Battles over General-Purpose Model Training Have Begun

In the case brought by four illustrators against Trik AI, the defendant has raised three lines of defense: no substantial similarity in the generated results, that the training constitutes fair use, and that the data

[14] *See* Hangzhou Internet Court, (2024) Zhe 0192 Min Chu No. 1587, and Hangzhou Intermediate People's Court, (2024) Zhe 01 Min Zhong No. 10332; *see also* item 29, Determination of Infringement Liability of Generative AI Service Providers ("生成式人工智能服务提供者的侵权责任认定"), in Summary of the Annual Report on the Application of Law in Intellectual Property Cases in Courts Nationwide (2024) (《全国法院知识产权案件法律适用问题年度报告（2024）摘要》), Supreme People's Court, https://www.court.gov.cn/zixun/xiangqing/462901.html.

[15] *See* Wang Jian, In the AI Era, Intellectual Property Adjudication Faces New Challenges (《AI 时代 知识产权司法迎来新挑战》), *Democracy and Legal System Weekly* (《民主与法制》), Issue 25 (2025), https://ipc.court.gov.cn/zh-cn/news/view-4513.html (describing the disputes and the then-procedural status of the lawsuit brought by four illustrators against Trik AI and the iQIYI v. MiniMax/Hailuo AI case). As of September 2026, we had not located any publicly available judgment on the merits in these cases that could be verified.

came from authorized sources; iQIYI, for its part, alleges that MiniMax and its Hailuo AI product used materials from its audiovisual works for model training without permission. Questions concerning data sources, use of works, and fair use in both cases await clarification in future decisions on the merits.

Institutional proposals are already under discussion in response. Writing from a legislative policy perspective, Judge Qi Lei of the Supreme People's Court has proposed a framework of "lenient entry and strict exit" ("宽进严出")—on the input side, constructing corresponding fair use regimes for data in the copyright, personal-information, and enterprise-data contexts, while permitting rights holders to deploy technical measures to restrict scraping; on the output side, strengthening risk control through mechanisms such as tort liability—thereby providing a reference point for the eventual development of the rules[16]. Academic discussion is likewise unfolding. Professor Wang Qian of the School of Intellectual Property at East China University of Political Science and Law argues that the fundamental theories of copyright law can hardly justify the unauthorized creation of training corpora containing massive volumes of works; that concepts and mechanisms imported from abroad, such as "non-expressive use," "transformative use," and "opt-out," can offer no once-and-for-all solution to this problem; and that using works to train artificial intelligence should not be categorically deemed fair use—instead, mechanisms should be improved and the rights and interests of rights holders safeguarded on the basis of a sound grasp of China's national conditions[17].

Doctrinal Lineage and Reach: The Evolution of Two Threads of Rules and Their Boundaries in the AI Era

The division of platform liability in the Hangzhou "Ultraman" case discussed above still operates within the existing "notice and necessary measures" framework. Yet the evolution of copyright rules in response to the large-scale digitized exploitation of works does not stop there. Only by tracing the origins and development of the relevant threads of rules can one see clearly the reach and boundaries of the existing rules as applied to the training stage.

The first thread concerns the indirect infringement liability of network service providers. In the late 1990s, the crawling, editing, organizing, and ranking of works by search engines and online platforms made "whether platforms should be liable for infringing content disseminated by users" a difficult question for copyright law. Section 512 of the U.S. Digital Millennium Copyright Act of 1998 (the "DMCA") established conditional limitations on liability for different types of online services, including "notice and takedown," and in principle, service providers are not required to monitor user activity generally or proactively[18]. China

[16] *See* Qi Lei, Legal Risks and Institutional Supply for AI Training Data (《人工智能训练数据的法律风险与制度供给》), Intellectual Property Court of the Supreme People's Court, 2025, https://ipc.court.gov.cn/zh-cn/news/view-5030.html. The article expressly proposes the "lenient entry, strict exit" institutional design from a legislative policy perspective.

[17] Wang Qian, A Rule-of-Law Response to Copyright Disputes over Generative AI (《生成式人工智能著作权争议的法治回应》), *China Rule of Law* (《中国法治》), Issue 4 (2026). Citing *American Geophysical Union v. Texaco*, the article notes that large-scale internal reproduction that materially affects the market for the works does not constitute fair use.

[18] Wang Qian, Key Issues in Platform Liability for Copyright Infringement and Their Resolution: Fault Determination and Necessary Measures (《著作权侵权平台责任的关键问题及其化解：过错认定与必要措施》), *Peking University Law Journal* (《中外法学》), Issue 3 (2026). For a systematic review of the legislative history of the "notice and takedown" mechanism and the two cornerstones of platform liability (the distinction between direct and indirect infringement, and

broadly followed this path: the Supreme People's Court's 2000 Interpretation on Several Issues Concerning the Application of Law in the Trial of Copyright Dispute Cases Involving Computer Networks first established a liability rule for those who "knowingly fail to remove" infringing content or fail to remove it "after a substantiated warning"; the 2006 Regulation on the Protection of the Right of Communication through Information Networks, drawing chiefly on the DMCA and the EU E-Commerce Directive, created "safe harbor" provisions, which have been carried forward through the Tort Liability Law to Articles 1195 and 1197 of the Civil Code. The series of cases during this period over whether search-engine "ranking lists" gave rise to knowledge is a typical example of these rules taking root in China[19].

The second thread concerns service providers' own exploitation of works and sits closer to the subject of this article. During the same period, U.S. courts addressed search engines' own crawling, copying, and display of thumbnails chiefly through the fair use doctrine—a platform's "providing a conduit for others" and its "own use of works" have been assessed separately from the outset[20]. The 2015 decision in *Authors Guild v. Google* established the analytical approach of "copying works in their entirety in order to provide information about the works" and has become one of the most frequently cited precedents in later AI training defenses[21]. China has not yet created a specific exception for text and data mining, but the Outline of the 15th Five-Year Plan has already proposed "establishing a fair use regime for AI training data," and a legislative response is on the agenda[22].

Viewing the questions discussed in this article through these two threads, one finds that the existing rules have developed distinct analytical approaches for platforms' own use of works and for their provision of technical services to third parties—a distinction that remains significant in AI scenarios.

First, the objects of regulation differ. The safe harbor rests on two cornerstones—the distinction between direct and indirect infringement, and the absence of a general duty to monitor on platforms—and its typical scenario is "users upload; the platform stores and disseminates": the platform is a conduit for others' infringing content. Model developers, by contrast, themselves collect works, build corpora, and conduct training, which involves their own use of works and must be assessed under the exclusive rights of copyright together with limitation rules such as fair use; as for infringement liability arising from a platform's provision of technical services to users, it must be separately assessed according to the services at issue

---

the absence of a general monitoring duty), *see* Parts I and II of that article.

[19] Professor Wang Qian proposed as early as 2005 the distinction between the "blank search box" and "ranking lists"—the latter, because they embody the service provider's editing, organizing, and ranking, support a presumption that the provider should have known the linked content was infringing; that view was later adopted by the Beijing High People's Court in the *Panasia v. Baidu* judgment and reflected in judgments such as the *Eleven Record Companies v. Yahoo* case and the *Newpass Online v. Tudou* case. *See* Wang Qian, *supra* note 18.

[20] *See Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003); *Field v. Google Inc.*, 412 F. Supp. 2d 1106 (D. Nev. 2006); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146 (9th Cir. 2007). *See* also Wang Qian, *Research on Copyright Protection in the Network Environment* (《网络环境中的著作权保护研究》), Law Press China, 2011.

[21] *See Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

[22] *See* the Outline of the 15th Five-Year Plan for National Economic and Social Development of the People's Republic of China (《中华人民共和国国民经济和社会发展第十五个五年规划纲要》) (March 2026): "improve the data standards system and quality management system, accelerate the building of AI corpora … establish a fair use regime for AI training data." On foreign text-and-data-mining exceptions, *see* Japan Copyright Act, art. 30-4 (successor to the "non-appreciative use" exception in art. 47-7 of the Copyright Act as amended in 2009, consolidated in the 2018 amendment); Directive (EU) 2019/790, arts. 3-4; Singapore Copyright Act 2021, s. 244. The UK Intellectual Property Office proposed in June 2022 a broad text-and-data-mining exception without a rights-holder reservation, but announced in February 2023 that it would not proceed, following strong opposition from the creative industries.

and the corresponding liability rules.

Second, the triggers of liability and the timing of remedies differ. "Notice and takedown" is a quintessential *ex post* remedy: the infringing content is in a state of public dissemination, and the rights holder can discover it, send a notice, and demand its removal, with necessary measures taken to prevent further harm. Training activities typically occur within the developer's internal environment; compared with public dissemination scenarios, the discovery, evidentiary challenges, and remediation of such acts differ, and must still be assessed in light of the training methods and technical facts involved.

It should be noted that existing platform liability-limitation mechanisms such as the safe harbor do not automatically lose their value simply because of advances in AI technology: for platform liability at the model-service level, judicial practice continues to operate within the "notice and necessary measures" framework (as in the division of platform liability in the Hangzhou "Ultraman" case), and scholars are likewise discussing its adaptation and application within the existing framework[23]. Improved technical capabilities do not, in themselves, mean that platforms must assume a general *ex ante* duty to monitor user conduct; developers' own training activities should still be assessed independently under the exclusive-rights and fair-use rules.

Three Observations and an Outlook in Comparative Perspective

First, the two countries' courts enter the analysis at different levels, and the focus of contention differs accordingly. U.S. courts proceed on the premise that reproduction occurs during training and concentrate on whether that reproduction can be excused by fair use, with the disputed issues revolving around the purpose of the use, data sources, and market evidence; existing Chinese decisions, by contrast, begin with a fine-grained technical characterization of the training conduct itself—the second-instance "Medusa" decision, for example, applied the "reproduction" standard to determine whether reproduction occurred, with the dispute focusing on technical facts such as model architecture and training methods (the first-instance courts in that case and in the Hangzhou "Ultraman" case took a different view). This difference has direct implications for cross-border litigation strategy: in Chinese forums, technical fact-finding and expert assistance may prove outcome-determinative; in U.S. forums, evidentiary contests over transformativeness and market harm may be more critical. Response strategies in the two jurisdictions should be built around their respective centers of gravity.

Second, the independent assessment of training-data sources is converging by different routes in the two countries. In China, Article 7 of the Interim Measures establishes the "lawful source" requirement at the regulatory level; in the United States, *Bartz v. Anthropic* made clear at the judicial level that "lawful training cannot automatically excuse pirated acquisition." The direction is aligned, but the points of impact differ: in China, the risk manifests both as administrative compliance exposure and potentially as direct civil infringement liability (in the "Medusa" case, capturing images to form the image package was itself found to infringe the right of reproduction); in the United States, it manifests directly as exposure to large damages awards and settlement pressure. For enterprises, data acquisition and subsequent training should be assessed separately: a training purpose cannot automatically legitimize the earlier acquisition,

[23] *See* the discussion of the Hangzhou "Ultraman" case in Part III of this article; Wang Qian, *supra* note 17.

nor do disputed data sources automatically foreclose a fair use defense that may apply to the subsequent training.

Third, the two countries' existing precedents correspond to different training methods, and their conclusions should not be mechanically transposed onto one another. *Bartz v. Anthropic* and *Kadrey v. Meta* concern general-purpose pre-training: massive data from a wide range of sources, a tenuous connection between any single work and the final outputs, and disputes centering on whether the training conduct itself can be excused; the "Medusa" case and the Hangzhou "Ultraman" case concern directed fine-tuning: a direct correspondence between specific materials and the generated results, with the core questions being whether the model reproduced the expression in the works and how liability should be allocated between users who train and post the models and the platform. Different factual structures frame different legal questions.

Dimension	United States	China
Observation 1 **Different levels of analysis; different centers of gravity**	**Point of dispute: whether reproduction can be excused** Reproduction during training is acknowledged; the discussion centers on whether fair use is established, with evidentiary contests over transformativeness and market evidence (*Thomson Reuters v. ROSS*; *Bartz v. Anthropic*; *Kadrey v. Meta*)	**Point of dispute: whether reproduction occurred** In the existing LoRA cases, the training conduct is technically characterized, and the "reproduction" standard determines whether reproduction occurred; technical fact-finding and expert assistance may prove outcome-determinative (the "Medusa" case)
Observation 2 **Independent assessment of data sources—convergence by different routes**	**Judicial route** *Bartz v. Anthropic* makes clear that "lawful training cannot automatically excuse pirated acquisition"; the risk manifests directly as high damages and settlement pressure	**Regulatory + civil dual route** Article 7 of the Interim Measures for the Management of Generative AI Services establishes the "lawful source" requirement; the relevant conduct involves administrative compliance exposure, and whether civil infringement is constituted requires independent assessment
Observation 3 **Precedents correspond to different training methods; conclusions not directly interchangeable**	**General-purpose pre-training** Massive data from a wide range of sources; tenuous link between any single work and the outputs; the dispute centers on whether the training conduct itself can be excused (*Bartz v. Anthropic*; *Kadrey v. Meta*)	**Directed fine-tuning** Direct correspondence between specific materials and generated results; the core issues are whether the model reproduced the expression in the works and the division of liability between users and the platform (the "Medusa" case; the "Ultraman" case)

Takeaway: Different factual structures frame different legal questions—the two countries' adjudicative conclusions should not be directly interchanged, and response strategies should be built around each jurisdiction's own center of gravity.

(Figure 3: Three Observations on China–U.S. Copyright Disputes over Training Data)

Summary and outlook: neither China nor the United States has created a specific exception for the training stage under current law, and the governing rules still depend on fair use, lawful-source requirements, and case-by-case adjudication. On the U.S. side, the reasoning of *Bartz v. Anthropic* and the Department of Justice's Statement of Interest have pushed the "separate assessment of source and use" to center stage; on the Chinese side, the Interim Measures and the AI Opinions reserve room on the two fronts of administrative compliance and evidentiary burdens, respectively, and the merits decisions in cases such as *iQIYI v. MiniMax* and the Trik AI case will remain key windows for observing the formation of Chinese rules.

Conclusion: Four Things Enterprises Can Do Now

For enterprises that continuously conduct large-scale model training and cross-border operations, identifying copyright risks and preparing the bases for lawful use are equally important. Drawing on the existing rules and cases in China and the United States, we suggest that compliance work can rest on at least four pillars. First, move the review upstream into the training-data supply chain: verify data sources and the scope of authorization, and set bright-line rules barring the admission of data clearly obtained from pirated repositories or by circumventing technical measures—the USD 1.5 billion settlement in *Bartz v.*

Anthropic underscores the real-world stakes. Second, keep stage-by-stage records: maintain complete documentation of data acquisition, cleaning, training versions, and model evaluations, which can both support defenses such as fair use and delineate the scope of liability, and also help meet evidentiary requirements in related litigation. Third, before launch, conduct regurgitation testing and set output controls for models—particularly fine-tuned models—that are capable of consistently reproducing specific works; this responds directly to the reasoning of the "Medusa" case. Fourth, for enterprises operating in both China and the United States, litigation and compliance strategies in the two jurisdictions should be designed in a coordinated manner, and the organization of technical explanations and market evidence should take account of the adjudicatory needs of both forums.

At present, the relevant judicial rules are still being gradually clarified. The new wave of U.S. lawsuits brought against Anthropic by Sony Music Publishing, Universal Music Publishing Group, and others is ongoing, and no publicly available decisions on the merits have yet been issued in China's *iQIYI v. MiniMax* case or the Trik AI case. These cases' determinations on the use of training data and the fair use defense merit the attention of enterprises engaged in model R&D and cross-border operations. We will continue to follow the relevant judicial developments.

This issue has addressed the legal boundaries of works entering models; when model outputs overlap with pre-existing works, how should liability be allocated among users, platforms, and developers? We will take up that question in the next issue.

Important Announcement

This Legal Commentary has been prepared for clients and professional associates of Han Kun Law Offices. Whilst every effort has been made to ensure accuracy, no responsibility can be accepted for errors and omissions, however caused. The information contained in this publication should not be relied on as legal advice and should not be regarded as a substitute for detailed advice in individual cases.

If you have any questions regarding this publication, please contact:

Yan WANG

Tel: +86 21 6080 0200
Email: yan.wang@hankunlaw.com

Lin YU

Tel: +86 21 6016 9780
Email: lin.yu@hankunlaw.com

© 2026 Han Kun Law Offices. All Rights Reserved.