Working Group on Scholarly Content & Generative AI - Whitepaper

Primary authors: Claire Schneider (OGC), Katie Zimmerman (Libraries) Alex Curylo (OGC 2025 summer intern)

Introduction

The goal of this white paper is twofold. First, we will describe the various perspectives of the MIT stakeholders that have an interest in the use of scholarly content to train AI systems. Those include researchers who seek to use scholarly content to train AI systems; stewards of MIT content, who seek to ensure sufficient access to scholarly content for MIT researchers to use to train; "in-house" publishing entities that hope to capitalize on opportunities with third parties who want access to MIT scholarly content for training; and those MIT members who create scholarly content.

Second, we have set forth the legal landscape that governs the use of scholarly content for AI training, both in terms of the way the law currently views using scholarly content with and without permission to train AI systems and the legal arguments that a number of varied US stakeholders are making to the courts and legislative bodies relating to what the law *ought* to be.[1]

We hope this white paper will be educational for those MIT community members not steeped in this topic or the pressing and varied needs of the MIT stakeholders we discuss below. We also hope to educate all MIT community members on relevant aspects of the law, and the areas in which the law may be evolving, which may impact these issues in the future.

MIT Perspectives case studies

A range of perspectives on AI use can be found across MIT. In this section, we will go through several examples.

Researchers

Researchers at MIT use machine learning techniques in their research methodology, and may use scholarly content to train machine learning models. Materials science research is a particularly strong example of the ways in which these research routes intersect with published academic content. In one study, MIT researchers designed a generative tool to predict the parameters needed to synthesize specific chemical materials.[2] The researchers identified approximately 12,000 published research papers that identified synthesis conditions for metal oxides. A neural net was trained on a subset of human-annotated data and then used to extract relevant synthesis characteristics from the full dataset of scholarly articles.[3] The database of extracted synthesis parameters is then subsequently used to train a machine learning

1 This paper does not constitute legal advice. You should seek legal counsel before taking any action.

[2] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti, *Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning*, 29(21) Chemistry Materials 9436–9444 (October 2017), https://pubs.acs.org/doi/10.1021/acs.chemmater.7b03500.

[3] Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum and Elsa Olivetti, *Machine-Learned and Codified Synthesis Parameters of Oxide Materials*, 4 Sci. Data 170127 (Sept. 2017), https://doi.org/10.1038/sdata.2017.127.

model that can generate synthesis outcomes on material systems beyond the original training data. The end result is a tool that can take desired material properties and generate a recipe for creating those properties, whether or not such a recipe has been previously identified. Generated recipes can be validated through more traditional methodologies to identify viable syntheses, and information can also be gained through analysis of the conditions and parameters identified by the model.

Another example is research that uses trends in the academic literature to predict when a new study will have a significant scientific impact.[4] This research provides an alternative to citation-based impact metrics by using artificial intelligence to better model patterns of influence in academic discourse. This research trained a machine learning model on a corpus of articles published in a particular academic field, looking at the patterns around seminal technical breakthroughs. The model can then analyze a new paper and generate a prediction, based on this historical data, of how influential the new paper will be.

An additional example is a study that trained a model to more accurately detect fake news.[5] The training data for this model consisted of news-article content and information about the news content, such as the Wikipedia and Twitter pages of the publisher and web traffic data, validated against a human-annotated fact-checking database. The model that was trained on these sources could then be used to evaluate new news items for factual accuracy and bias.

These examples illustrate several research use cases. First is the use of machine learning techniques to extract information from scholarly content. AI tools can be trained to extract information from written content so that it can be used in subsequent research.[6] An algorithm that has been trained to recognize specific content can assist in identifying and extracting that content when it is embedded in an academic paper. The training material for such an algorithm may not itself be scholarly content – it could be a dataset of whatever the researcher is seeking to extract from the literature – but the resulting algorithm would subsequently use scholarly content as input from which information, associations, or inferences could be extracted. This type of machine learning is not necessarily "generative" – it will not create a new example of things like the training material – but it uses machine learning to better identify content that is similar-but-not-identical to the target. This use case uses scholarly content as input, but not necessarily as training data.

Second, and perhaps more common, is the use of scholarly content directly as training data for machine learning algorithms. Academic researchers publish their findings in the scholarly literature in order to share the results of their research as broadly as possible with the rest of the academic community and the public. Much in the same way that a human could read large amounts of academic writing and make inferences and draw connections between them, a machine learning algorithm can do the same at a

---

[4] James W. Weis and Joseph M. Jacobson, *Learning on knowledge graph dynamics provides an early warning of impactful research*, 39 Nature Biotechnology 1300–1307 (2021), https://doi.org/10.1038/s41587-021-00907-6.

[5] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, Preslav Nakov, *Predicting Factuality of Reporting and Bias of News Media Sources*, Proc.2018 Conf. on Empirical Methods Nat. Language Processing 3528–3539 (2018), https://aclanthology.org/D18-1389/.

[6] *See, e.g.,* Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, and Elsa Olivetti, *Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks*, 60(3) J. of Chemical Info. and Modeling 1194-1201 (2020), https://pubs.acs.org/doi/10.1021/acs.jcim.9b00995; Zhengkai Tu, Sourabh J. Choure, Mun Hong Fong, Jihye Roh, Itai Levin, Kevin Yu, Joonyoung F. Joung, Nathan Morgan, Shih-Cheng Li, Xiaoqi Sun, Huiqian Lin, Mark Murnin, Jordan P. Liles, Thomas J. Struble, Michael E. Fortunato, Mengjie Liu, William H. Green, Klavs F. Jensen, and Connor W. Coley, *ASKCOS: Open-Source, Data-Driven Synthesis Planning*, 58(11) Accounts Chemical Res. 1764-1775 (2025), https://pubs.acs.org/doi/10.1021/acs.accounts.5c00155.

massive scale. As in the chemistry use case described above, this can allow for more efficient discovery of patterns and inferences that can lead to new discoveries. A generative model trained on the chemistry literature can produce new chemical syntheses, a model trained to identify patterns in biochemistry publishing can predict high-impact research, and a model trained to recognize factuality and bias can help assess the veracity of news content. This use case generally requires access to large quantities of relevant material to train from. It also may or may not result in a tool that creates outputs that look like the training data. An AI algorithm trained on academic journal articles *could* be designed to generate more journal articles, but it could also generate plausible chemical syntheses or metrics about the target input.

In all cases, academic research may or may not lead to subsequent creation of a commercial product. Regardless of subsequent commercialization, most research conducted in the academic environment starts as a noncommercial project, and the academic publications that describe the research are generally not commercialized by their authors. Commercialization, or the potential for commercialization, can often color thinking about academic applications, but arguably it should not when considering academic uses. Important to note is the distinction between the publication itself, the written work that describes the scientific advance and the scientific advance *itself*, which may be protected as a patent or make reference to copyrightable software, both of which have potential commercialization potential.

Researchers are also the authors of scholarly content, and also find themselves on the other side of the equation, where their content may be used for training in other products, either via licensing or fair use. MIT authors published approximately 7,000 academic articles in calendar year 2024,[7] and the vast majority of academic articles do not result in financial compensation to the author. Several academic publishers that own, or have rights to license books and/or journal content, have announced lucrative licensing deals with AI companies,[8] with little or no financial compensation flowing back to the authors.[9] Where this is sometimes felt most acutely is around use of academic theses as training data. MIT theses are required to be made publicly available.[10] The MIT Libraries has received questions from thesis

[7] Digital Science. (2018-) Dimensions [Software] available from https://app.dimensions.ai. Accessed on August 8, 2025, under license agreement.

[8] *See, e.g., Research Growth, AI Licensing, and Cost Reduction Drive Wiley's Fiscal 2025 Results*, Wiley (June 17, 2025), https://newsroom.wiley.com/press-releases/press-release-details/2025/Research-Growth-AI-Licensing-and-Cost-Reduction-Drive-Wileys-Fiscal-2025-Results/default.aspx; *AI Subsidiary Aggregation - Author FAQs*, Cambridge Univ. Press, https://infogram.com/1p9g1kvndzqkrkt7523yd02wk3b3grmm9mw (last visited Nov. 26, 2025).

[9] Cambridge University Press (CUP) has taken the step of asking authors to opt in or out of inclusion in their AI training corpus. "Royalty-bearing rightsholders" (which generally includes book authors, but not authors of journal articles) are, as of July 2025, offered a 20% royalty for inclusion of their content in a licensed AI training corpus. In an interview, Ben Denne, Director of Publishing, Academic Books, at CUP, explained the math: a $5 million dollar licensing revenue would be divided across the rightsholders of content included in the corpus, either evenly or based on actual usage. Dave Hansen, *What happens when your publisher licenses your work for AI training?*, Authors Alliance (July 30, 2024), https://www.authorsalliance.org/2024/07/30/what-happens-when-your-publisher-licenses-your-work-for-ai-training/. A $5 Million licensing deal consisting of all CUP books published between 2014-2024 (approximately 141,000 titles [Dimensions data]), divided without regard for specific usage, would therefore result in a royalty payment of approximately $7/book to authors.

[10] *See MIT Policies and Procedures 13.1.3 Intellectual Property Not Owned by MIT*, Mass. Inst. Tech. (last updated Jan. 30, 2025), https://policies.mit.edu/policies-procedures/130-information-policies/131-intellectual-property; *MIT Rules and Regulations 2.73*, Mass. Inst. Tech. (last updated Feb. 2023), https://facultygovernance.mit.edu/faculty-rules-page#2-73.

authors concerned that their thesis will be used as training data without their consent or compensation. MIT researchers may have similar concerns about other scholarly publications.

The MIT Press and other MIT publications

MIT is also a leading publisher of scholarly and scientific content. The MIT Press, a department of MIT, publishes hundreds of trade, text, and scholarly books every year for which MIT owns or administers copyright, as well as dozens of journals. Many of its publications are provided in open access digital form, using a variety of open licenses.

Other examples of publishing within MIT include Sloan Review of Management, MIT Open Learning, and the MIT Museum, which houses and licenses a large collection of photography related to the history of science and computing at MIT, and many departments, labs, and centers that publish white papers, tech reports, and preprints on websites and in repositories hosted by MIT.

Publishing entities within MIT may be interested in licensing or otherwise monetizing their content for AI uses. This is a potentially helpful revenue stream, particularly in increasingly difficult budgetary circumstances. AI is radically reshaping how knowledge is created, accessed, and trusted. As a result, academic publishers face declining revenues and diminished control over scholarly content. Publishers may also be in a position to mediate between authors and AI companies in such licensing deals; indeed, in many instances they are contractually obligated to manage training rights and other digital rights on behalf of authors. MIT publishing entities certainly want to ensure that their authors' interests and preferences are being respected and followed. The differences among and between authors in terms of what licensing actions the MIT publishing entities should/shouldn't take are many, and may cause paralysis (or at the very least significant complication) on the part of publishing entities in terms of decisions around licensing.

Publishing entities also need to consider carefully how any licensing revenue would be shared with authors, as well as how transparent publishing entities will be with their authors (and third parties) regarding licensing deals.

MIT Libraries

The MIT Libraries finds itself between many of the positions identified above. The Libraries licenses scholarly content for use by MIT researchers, and is therefore frequently involved in negotiations with publishers around researchers' use of that content. The MIT Libraries, and libraries in general, have long advocated for electronic content licensing that mirrors standard rights under copyright law, because it is advantageous for instruction, research use, and compliance with contractual obligations for the same rules to apply to use of content, regardless of medium.[11] Restrictive license terms can counterintuitively mean that greater rights are available for less convenient media – if, for example, the ebook version restricts computational uses of the text that would be permitted under fair use of a print edition. The Libraries therefore has an interest in consistency around when AI licensing is necessary and consistency across usage terms when licensing is required.

The Libraries also has a longstanding interest in promoting free and open access to knowledge.[12] Libraries exist to provide broad access to information, a mission that has become more difficult in the

[11] *See, e.g.,* Katie Zimmerman, *Fair Use Savings Clauses*, E-Resource Licensing Explained (2024), https://berkeley.pressbooks.pub/eresourcelicensingexplained/chapter/fair-use-savings-clauses/.

[12] *See, e.g.,* MIT Ad Hoc Task Force on the Future of Libraries, *Institute-wide Task Force on the Future of Libraries*, MIT Libraries (last updated Feb. 23, 2023), https://mitl.pubpub.org/pub/future-of-libraries/release/1.

increasingly licensing-driven digital landscape. Print library materials benefit from broad, statutory copyright limitations such as the first sale doctrine,[13] library exceptions such as those allowing interlibrary loan,[14] and fair use.[15] While such exceptions also apply to digital materials, digital library content is more likely to be constrained by restrictive contract terms in content licenses, which can effectively eliminate those statutory rights.[16] The Libraries therefore generally seeks to support fair uses to the fullest extent of the law, and to support researchers in making full use of library content. Supporting fair use supports academic freedom and innovation, since it allows use of materials regardless of proprietary interests which can skew results. Traditional impact metrics, for example, are based on proprietary bibliographic databases and have received much criticism for reflecting biases and oversimplification of research impact.[17] If rightsholders could inhibit development of alternative metrics by withholding a license for their material to be included in training data, it would create biases in the training data and would inhibit both the academic freedom of those researchers to accurately explore this research space, and inhibit innovation in bibliometrics that could ultimately benefit the whole scholarly ecosystem.

In addition to its values-based positions, the MIT Libraries is also the distributor and caretaker of a large quantity of scholarly content that is either in the public domain or for which MIT is the copyright holder. This content is potentially useful as training data for AI models.[18] The Libraries generally makes content as open as possible and seeks to facilitate computational uses of library content.[19] However, hosting and facilitating that access has costs. Licensing for AI uses could cover those costs and potentially increase access to content that would not otherwise be available. Additionally, some content, such as archival content, may present privacy concerns even if all copyright concerns are accounted for. Sharing of archival content for AI training only under contractual agreements could present a way to allow for greater control over dissemination and use of this more sensitive content.

Potential conflicts between these positions

Several potential tensions present themselves between these MIT stakeholders. Most notably, there is tension between positions that would benefit from licensing or ensuring contractual controls

---

[13] 17 U.S.C. § 109.

[14] 17 U.S.C. § 108.

[15] 17 U.S.C. § 107.

[16] *See, e.g.,* Dave Hansen, Yuanxiao Xu, and Rachael G. Samberg, *Contractual Override: How private contracts undermine the goals of the Copyright Act for libraries and researchers, and what we can do about it*, 72 J. Copyright Soc'y 675 (2025), https://copyrightsociety.org/journal-entries/contract-override/.

[17] Journal Impact Factor, probably the most commonly used impact metric, is produced by Clarivate based on their proprietary Web of Science database. Journal Impact Factor and Web of Science have both faced significant criticism for reflecting biases, being under-inclusive of content produced in the Global South, and being easily subject to manipulation in the editorial process. Its largest competitor, CiteScore, is produced by Elsevier, based on the proprietary Scopus database. *See, e.g.,* Charles J. Gomez, Andrew C. Herman and Paolo Parigi, *Leading countries in global science increasingly receive more citations than other countries doing similar research*, 6 Nat Hum Behav 919–929 (2022), https://doi.org/10.1038/s41562-022-01351-5. *See generally Promotion & Tenure and Open Scholarship: How is "impact" measured & valued?*, MIT Libraries (last updated May 13 2022), https://libguides.mit.edu/c.php?g=1162307&p=8489698.

[18] *See* Chris Bourg, Sue Kriegsman, Nick Lindsay, Heather Sardis, Erin Stalberg, and Micah Altman, *Generative AI for Trustworthy, Open, and Equitable Scholarship*, An MIT Exploration of Generative AI: From Novel Chemicals to Opera (Mar. 27, 2024), https://doi.org/10.21428/e4baedd9.567bfd15.

[19] *See MIT Framework for Publisher Contracts*, MIT Libraries (last updated May 19, 2020), https://libraries.mit.edu/scholarly/publishing/framework/; *Resources and Tools for Computational Research: Home*, MIT Libraries, https://libguides.mit.edu/comptools (last visited Nov. 26, 2025).

around use of content for AI training, and positions that benefit from unrestricted fair use of content for AI training. MIT publishing entities, and possibly the MIT Libraries and researchers creating training datasets, have a financial interest in being able to license content for training. All stakeholders may also have privacy and ethics concerns with the unrestricted training use of all materials. Researchers and the MIT Libraries, and to a lesser extent publishing entities, in contrast, benefit from being able to use content for training without the frequently prohibitive financial and logistical burden of negotiating a licensing deal for each use. All MIT stakeholders see the important need to foster and support incentives for human generated content, which includes supporting mission-driven publishing entities that support high quality academic content.

These positions may not be as diametrically opposed as they would first seem, however. Historically, many things can be a fair use in some circumstances and require licensing in other circumstances. Using content to train LLMs for noncommercial academic research may present a different situation than use of content to train for commercial applications, and uses that can substitute for direct access to the training data are different from uses that enable original discoveries. Where these lines should be drawn is something that the courts are continuing to wrestle with. Given the current status of court decisions, however, clean lines of demarcation may not be forthcoming. MIT must also carefully consider how to weigh the various interests of its numerous stakeholders to the extent there are differences in positions regarding a reliance on fair use versus desire to enter into revenue generating licensing arrangements.

Historical-legal background of copyright and innovation

Copyright is a form of intellectual property protection for various media, text, and other creative works. Copyright is grounded in the U.S. Constitution, art. I, § 8, cl. 8, with a purpose to "promote the progress of science and the useful arts" through the government's protection. Notably, though, law protects only original (including derivative) works according to their *authorship*, not any related claims of ideas. In other words, copyright law protects the tangible expression of an idea by an author (or group of authors) but not the underlying idea or concept.[20] Copyright does not protect facts, systems or methods of operation.[21] The focus of this white paper is on creative or scholarly content.

The legislative landscape of copyright is broad, and copyright is governed by a succession of acts (including most recently the Copyright Act of 1976, though various minor portions of copyright law continue to be updated piecemeal), codified in Title 17 of the United States Code.[22] The Copyright Office (a part of the Library of Congress) also issues specific regulations implementing the law, and various international treaties contribute to some copyright provisions.

As technology has advanced, courts have frequently had to address how copyright would (or would not) extend into new forms. One case in which the Supreme Court has wrestled with assumptions about new technologies is *New York Times Co. v. Tasini* (2001), in which the Supreme Court sided with the author plaintiffs, finding that certain reproduction rights granted to publishers did not, by default, permit publishers to reproduce content in any and all new media formats. This case involved a dispute between authors and the New York Times, who had received a license from the authors to publish and distribute articles in the print newspaper. When electronic news databases were developed, the Times licensed these articles to database providers without additional permission from the authors. The Times

[20] 17 U.S.C. § 102(b).
[21] Id.
[22] 17 U.S.C. Chs. 1-15.

argued that its licensing of an anthology of copyrighted works and the subsequent individual publication of those works in an electronic database was within their authority as a licensee and publisher of a collective work under copyright law. Many similar arguments are found in the pending copyright lawsuits concerning training Artificial Intelligence (AI) systems with copyrighted content without permission from the rightsholders. These cases frequently reflect a conflict between the underlying assumptions of authors and creators about extension of their rights into new technological spaces. Rightsholders are granted certain rights to exclude others from use of their work.[23] The Constitution counterbalances those rights against exceptions and limitations designed to "promote the progress of science and the useful arts."[24] How the courts interpret the interplay between the rights of creators and the rights of subsequent users will shape the development of AI technologies.

Copyright infringement overview

Before focusing on the specific legal issues at play in the lawsuits disputing the legality of non-permissive use of copyrighted content to train Large Language Models (LLMs), this paper steps back to understand the typical legal framework for evaluating a claim of copyright infringement. To make out a claim of copyright infringement, a United States Federal court will analyze whether (A) the underlying work meets the definition of copyrightability, (B) the work has been infringed, and (C) any defenses, such as fair use, are applicable.[25]

Copyrightability

Whether a tangible expression has copyright protection depends on whether the work was created by a human author,[26] rather than copied from other works and possesses at least some minimal degree of creativity.[27] The 'creativity' threshold is low: "[t]he vast majority of works make the grade quite easily, as they possess some creative spark, 'no matter how crude, humble or obvious' it might be."[28] Books, magazines, and other academic publications easily surpass this bar. Facts, ideas, and information[29] as well as information structure such as recipes (including, e.g., culinary recipes and semiconductor process recipes) and methods, are not copyrightable.[30] The threshold of copyrightability is generally met with regard to the training data in most AI training lawsuits of the past few years, to the extent that the training data consists of books, periodicals, and scholarly articles.[31] As will be discussed further, however, AI uses

---

[23] Copyright holders may exclude others from reproducing, adapting (i.e., creating derivatives), distributing, publicly performing, and publicly displaying your work without your permission. 17 U.S.C. § 106.

[24] U.S. Constitution, art. I, § 8, cl. 8.

[25] For purposes of this paper, only the defense of fair use will be discussed.

[26] At least for now copyright is restricted to human creators. Recent examples of individuals attempting to copyright AI-generated works have been denied by the Copyright Office. *See, e.g.,* U.S. Copyright Review Board, Refusal to Register Théâtre D'opéra Spatial (Sept. 2023), https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf; U.S. Copyright Office, *Copyright and Artificial Intelligence – Part 2: Copyrightability* (Jan. 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf.

[27] *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 345 (1991).

[28] *Feist*, 499 U.S. at 345.

[29] 17 U.S.C. § 102(b).

[30] Nimmer on Copyright § 2A.13 (2025); Kurt M. Saunders & Valerie Flugge, *Food for Thought: Intellectual Property Protection for Recipes and Food Designs*, 19 Duke L. & Tech. Rev. 159, 165 (2021) ("As applied to recipes, copyright law affords little protection.").

[31] *See, e.g., The New York Times Company v. Microsoft Corporation, et al.,* No. 1:23-CV-11195 (S.D.N.Y), [hereinafter *NYT v. OpenAI*]; *Bartz v. Anthropic PBC*, No. C 24-05417 WHA, 2025 WL 1741691 (N.D. Cal. June 23, 2025) (Order on Fair Use) [hereinafter *Bartz v. Anthropic* or *Bartz*].

do not necessarily reproduce the copyrightable aspects of the training data in the output of the model. In the material science examples discussed previously, for example, the training data consists of scholarly articles about recipes for creating molecules. While the article texts describing the experimental process and conditions almost certainly meet the low threshold for copyrightability, the chemical ingredients and methods are facts, which do not become proprietary by virtue of being described through copyrightable media. To the extent that an AI model strips the copyrightable content away from unprotectable facts and ideas, the model may interact with copyrightable content in its training phase, but does not necessarily reproduce copyrighted content in the eventual output.

An author becomes the copyright owner immediately upon creating a work that satisfies the creativity threshold for being a copyrightable work. It is not necessary to register copyright in the US Copyright Office in order to own the copyright to a work, but registering copyright is a prerequisite to filing an infringement lawsuit.[32]

Infringement

To make out a claim of direct infringement, one must prove: (1) the copying of original expression from the work, (2) "substantial similarity" between the copied expression from original work and the alleged infringing work and (3) that the alleged infringer had access to the original work.[33]

Copyright liability is typically 'strict' i.e., the analysis does not depend upon whether the alleged infringer knew of the infringement or not; indeed, infringement can be found where evidence demonstrates that the infringer was aware of, or had sufficient access to, the original work, even if it is proven that the alleged infringer was not aware of, let alone intentionally, copied the work.[34]

The copied expression must also be "substantially similar" to the original work. Where the output of an LLM is a verbatim copy of copyrightable input used for training, this is easily established. It is a harder question when an LLM summarizes, translates, or reframes content, or combines content from multiple sources to produce an output. "Substantial similarity" is formulated differently in different jurisdictions but generally involves both an objective and subjective comparison of the original and allegedly infringing copy, and may additionally involve filtering out shared unprotected commonalities (such as public domain content, or content that does not meet the copyrightability standard).[35] Whether a specific AI tool produces outputs that are substantially similar to the training inputs is likely to be highly fact specific and contested in the lawsuits going forward.

[32] 17 U.S.C. § 411.

[33] *Cf. Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, 765 F. Supp. 3d 382 (D. Del. 2025) [hereinafter *Reuters v. Ross* or *Ross*].

[34] *E.g., Bright Tunes Music Corp. v. Harrisongs Music, Ltd.*, 420 F. Supp. 177, 180 (S.D.N.Y. 1976) (George Harrison was found to have infringed the song "He's So Fine," albeit unintentionally, in his song "My Sweet Lord" based on evidence that Harrison unconsciously misappropriated the tune from the song "He's So Fine" after hearing it numerous times given its popularity).

[35] *See, e.g., Rentmeester v. Nike, Inc.*, 883 F.3d 1111, 1117-1124 (9th Cir. 2018); *Boisson v. Banian, Ltd.*, 273 F.3d 262, 272 (2d Cir.2001).

Translations and summaries are also classic examples of a "derivative work,"[36] which is another of the exclusive rights given to copyright holders.[37] In practice, the analysis to determine if something is a "derivative work" also relies on whether there is substantial similarity between the original work and the alleged derivative.[38] This legal definition of a derivative work notably differs from colloquial usage; many things may be colloquially said to be derived from prior works without necessarily infringing on those prior copyrights. Determining whether LLMs are or produce derivative works of training content is another topic that can be expected to be highly contested an AI litigation continues.

Fair Use

Fair use is an affirmative defense that must be invoked by defendants and has been invoked by essentially every AI technology company sued under copyright law. The fair use statute provides that:

> [T]he fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —
>
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[39]

Analyzing a complex, four factor test such as this one can be a subjective exercise, and courts can vary in their analysis of the factors and ultimate conclusions.[40] All factors are not weighted equally and the relative importance of each can be adjusted depending on the facts of the particular case. Fair use doctrine has evolved as technology has evolved.

Several cases have been influential in the modern interpretation of fair use. One distinction that is potentially relevant to fair use is between commercial and non-commercial uses. The Supreme Court differentiated explicitly between commercial and noncommercial use in *Harper & Row* (1985).[41] In that

---

[36] 17 U.S.C. § 101 ("A "derivative work" is a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted. A work consisting of editorial revisions, annotations, elaborations, or other modifications which, as a whole, represent an original work of authorship, is a "derivative work".).

[37] 17 U.S.C. § 106(2).

[38] *See, e.g.*, *Caffey v. Cook*, 409 F. Supp. 2d 484, 496 (S.D.N.Y. 2006); 1 Nimmer on Copyright § 3.01 (2025).

[39] 17 U.S.C. § 107.

[40] For example, the Supreme Court de-emphasized commercial nature in 1990s in *Campbell v. Acuff-Rose* (1994), but more recently considered this commercial use in a greater way in *Andy Warhol Foundation v. Goldsmith* (2023), as discussed below; *see also* Barton Beebe *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 U. PA. L. REV. 549 (2008).

[41] *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539 (1985) [hereinafter *Harper & Row*].

case, the publisher Harper & Row sued magazine company Nation Enterprises over unauthorized use of quotes from a forthcoming book, and the case centered on whether the magazine's quotations were fair use.[42] Commercial use tends to weigh against fair use because the rights granted by copyright are primarily economic rights, and commercial uses are more likely impedes the rightsholder's ability to profit from their work. Copyright grants a monopoly to creators as an incentive to produce creative works; a use that diminishes this economic outcome weakens this incentive. Proving fair use may be a lower bar for non-commercial uses,[43] but there is still potential for infringement. Copyright suits against academic institutions are not unheard of: see, for example, *UAB Planner 5D v. Facebook and Princeton*[44] and *Cambridge University Press v. Patton.*[45] With that said, as to LLM/AI training, there are currently no pending cases involving academic institutions' use of copyrighted material. This is despite a large participation by academic institutions in development and training of LLM systems.

Another important set of cases in the context of LLM/AI training is *Authors Guild v. Google and Authors Guild v. HathiTrust.* In *Authors Guild v. Google*, a 2015 case concerning whether Google's digitization of books for its Google Books service constituted fair use, highly transformative uses that did not publicly display significant portions of copyrighted works, and thus did not displace the original market were found to be fair use, even with commercial use.[46] This was a departure from the commercial use evaluation common a quarter-century earlier in cases like *Harper & Row*,[47] and tracks with the trend of commerciality becoming less important to fair use decisions over time, most notably seen in *Campbell v. Acuff-Rose* (1994).[48] *HathiTrust* dealt with nearly identical facts to *Google*, but notably dealt with academic libraries' copying of books to facilitate full-text search, without providing the actual text to end users. In *HathiTrust* the court concluded that "we can discern little or no resemblance between the original text and the results of the [...] full-text search,"[49] making the use "a quintessentially transformative use" when the result does not "merely repackage[] or republish[] the original[s]."[50]

The most recent Supreme Court fair use opinion comes from *Andy Warhol Foundation v. Goldsmith* (2023), which addressed only the first factor of the fair use analysis. In this case, the Court assessed the rights of a photographer, Lynn Goldsmith, in use of a painting that had used her photograph as a reference image, and had been licensed in similar uses before the alleged infringing use[51] Many commentators were surprised by the case's move away from the transformative nature of a works,[52] and the seeming return to prominence of commerciality in the fair use analysis. The court found that "the first fair use factor [is focused on] whether an allegedly infringing use has a further purpose or different character, which is a matter of degree, and the degree of difference must be weighed against other

---

[42] *Harper & Row*, at 539-40.

[43] *See Harper & Row.*

[44] Compl., *UAB "Planner5D" v. Facebook Inc.*, 3:20-cv-08261 (N.D. Cal Nov 23, 2020) ECF No. 1, available at https://www.courtlistener.com/docket/18689287/1/uab-planner5d-v-facebook-inc/.

[45] *See* Laura Burtle, *GSU Library Copyright Lawsuit*, Georgia State University College of Law Library, https://libguides.law.gsu.edu/gsucopyrightcase (last visited Nov. 26, 2025).

[46] *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), *cert. denied*, 578 U.S. 941 (2016).

[47] *Harper & Row*, at 539-40.

[48] *See Campbell v. Acuff-Rose Music, Inc.*, 510 U. S. 569, 580-85 (1994); *see also* Pierce N. Leval, *Toward a Fair Use Standard*, 103 Harv. L. Rev. 1105 (1990).

[49] *Authors Guild v. HathiTrust*, 755 F.3d 87, 97 (2d Cir. 2014).

[50] *Id.*, (quoting Leval, 103 Harv. L. Rev. at 1111).

[51] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 508-9 (2023) [hereinafter *Warhol*].

[52] *E.g.,* Steve Schindler & Katie Wilson-Milne, *SCOTUS Says Warhol Not So Fast*, The Art Law Podcast (Jun. 5, 2023), https://artlawpodcast.com/2023/06/05/scotus-says-warhol-not-so-fast-the-limitations-of-transformative-use.

considerations, like commercialism."[53] In other words, if the new allegedly infringing use has a similar purpose to the original, the 1st factor "is likely to weigh against fair use" absent extenuating justification.[54]

Training LLMs with copyrighted content

Limited glimpses into generative AI's inner workings have been revealed through recent litigation. Technology companies training LLMs pull their data from a variety of internet and print-derived sources, though many maintain their training data makeup as a closely guarded secret.[55] Some include data obtained through "shadow library" sources such as Library Genesis 1 and 2 (LibGen 1 and 2) and Books3.[56] These online repositories contain hundreds of thousands of books archived from various institutions, not always with permission from those institutions or publishers.

It is important to remember that simply because content is publicly accessible for free on the internet does not mean that permission is *not* required from the copyright owner prior to use that involves copying, downloading or reproducing. Despite colloquial language, if content is found on the internet and is not subject to an open access license (such as a Creative Commons or open source license)[57] it does not necessarily mean that such content is part of the "public domain."[58]

While there have been some large-scale licensing deals, such as OpenAI signing agreements with the Associated Press[59] and Hearst,[60] among others, Microsoft signing with HarperCollins[61] publishers, and Harvard releasing public domain works in its libraries in collaboration with OpenAI and Microsoft,[62] most training has not occurred as a result of licensing. Technology companies have explained that

---

[53] *Warhol*, at 525 (citing *Campbell v. Acuff-Rose Music, Inc.*, 510 U. S. 569, 579 (1994)).

[54] *Andy Warhol*, 598 U.S at 532-33.

[55] *See* Shayne Longpre, et al., *Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?*, An MIT Exploration of Generative AI (Mar. 27, 2024), https://doi.org/10.21428/e4baedd9.a650f77d ("Popular AI systems fail to disclose even basic information about their training data.").

[56] *See, e.g.*, *Kadrey*, at *6.

[57] These licenses typically permit permissive use of content, but have important restrictions and obligations. Notably, Creative Commons licenses almost always require attribution of the author(s), which may be impossible for certain training uses. *See generally About CC Licenses*, Creative Commons (2019), https://creativecommons.org/share-your-work/cclicenses/; *OSI Approved Licenses*, Open Source Initiative, https://opensource.org/licenses.

[58] A work is in the public domain, generally, when the copyright has expired; the content is not eligible for copyright protection; or it has been specifically designated as such by the copyright holder.

[59] *AP, Open AI agree to share select news content and technology in new collaboration*, Associated Press (July 13, 2023), https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/.

[60] *Hearst and OpenAI Announce Strategic Content Partnership*, Hearst (Oct. 8, 2024), https://www.hearst.com/-/hearst-and-openai-announce-strategic-content-partnership.

[61] The three-year deal allows for "limited use of select nonfiction backlist titles" for AI model training and performance improvements, according to statements from HarperCollins. Notably, author opt-in is required and authors would receive a per-book payment as compensation (on the order of a few-thousand USD per book). Andrew Albanese & Jim Milliot, *Agents, Authors Question HarperCollins AI Deal*, *Publishers Weekly* (Nov. 19, 2024), https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/96533-agents-authors-question-harpercollins-ai-deal.html.

[62] Kate Knibbs, *Harvard Is Releasing a Massive Free AI Training Dataset Funded by OpenAI and Microsoft*, Wired (Dec. 12, 2024), https://www.wired.com/story/harvard-ai-training-dataset-openai-microsoft/.

negotiating and entering into licensing arrangements with numerous, disparate copyright owners is inefficient and expensive.[63]

Data "scraped" from the internet and used for training must therefore be assessed as to whether it infringes the copyright of the source material. Assuming the scraped data is copyrightable – as the text in an article, book or other research publication almost always will be – the next question in the AI training context is whether "copying" has occurred. This is also generally easy to establish, as content must be downloaded and saved in its entirety as part of the training process.[64] Less clearly, the content may also be saved in the model's memory, via the statistical application of weights and parameters to the data.[65] Therefore, in most instances, the prima facie case for infringement is undisputed; the real question is over whether it is excused by fair use.[66]

Recent and Pending Litigation

This summary represents a snapshot in time, and judicial analysis of fair use and generative AI will continue to evolve over time. Legislative action is also possible and could greatly affect this space. As demonstrated by the comments the Copyright Office received in advance of their report on generative AI training,[67] technology companies, investors, researchers, publishers, and public interest groups all have strong opinions that may persuade courts or legislators as the legal landscape around generative AI use evolves.

In June 2025 the first two opinions analyzing fair use as a defense to the use of copyrighted content to train LLMs were issued: *Bartz v. Anthropic*, (N.D. Cal. June 23, 2025) and *Kadrey v. Meta*, (N.D. Cal. June 25, 2025). Both opinions are orders on motions by the defendants for summary judgment, which means that the defendants were asking the court to determine that, as a matter of law, defendants' use of the copyrighted content to train was a fair use. Both courts found in favor of the defendants on the question as to whether using copyrighted content without permission to train an LLM is fair use, albeit with somewhat different analysis. These cases highlight both the importance of the particular facts/arguments being presented and the differences in interpretation of fair use analysis even among judges sitting in the same district. Both cases are likely to be appealed to the Ninth Circuit. While these cases wend their way through the Ninth Circuit, other pending cases, such as *NYT v. OpenAI* in the Second Circuit, may issue opinions on fair use.

---

[63] U.S. Copyright Office, *Copyright and Artificial Intelligence – Part 3: Generative AI Training* (pre-publication version) 86 (May 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf [hereinafter *USCO Report Part 3*].

[64] *Id*. at 27.

[65] *USCO Report Part 3* at p. 28-30.

[66] Most plaintiffs are asserting both direct infringement, based on the ingestion (i.e., copying) of the data to train the LLM and also, separately, contributory infringement. Contributory infringement, in this circumstance, would mean that the creator of the LLM is liable for direct infringements committed by end users of the technology. This would require proving that the LLM will respond to user prompts with outputs that copy part or all of a copyrighted work, either because the output is a verbatim copy of the copyrighted content or because the output is substantially similar to the copyrighted content that has been ingested. Defendants, for their part, argue that the likelihood of this occurring is quite small for most large models, and is a "bug" in the system, rather than a core feature of the technology. Smaller models that train on less data or give substantial weight to certain data sources may be more likely to yield outputs that are copies of ingested data.

[67] *USCO Report Part 3*.

Bartz and *Kadrey*

As an illustration in how the fair use analysis is beginning to play out, this paper describes in detail the reasoning set forth in the *Bartz* and *Kadrey* opinions.

In the *Bartz* case, authors filed suit against Anthropic, the developer of Claude and other AI models, alleging that Anthropic pirated copies of authors' books and, separately, that Anthropic used a combination of pirated and legally purchased copies of authors' books for training LLMs.[68] In *Kadrey*, other authors sued Meta, the parent company of Facebook, over allegations that Meta similarly downloaded copies of authors' books from online "shadow libraries" and used said copies for training its Llama LLM product.[69]

Fair Use Analysis in *Bartz* and *Kadrey*

Purpose and Character of the Use

Analysis under the first fair use factor focuses on whether the use by the alleged infringer is essentially "transformative," or, in other words, does it "ad[d] something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[70] Use which is non-commercial, educational, news reporting, commentary, scholarship, or research is also helpful, but not necessary to a finding that this factor favors fair use.[71]

Both the *Bartz* and *Kadrey* courts recognized the clear transformative nature of the use of copyrighted content to train LLMs. The *Bartz* court was most emphatic, noting that copies used to train LLMs are "exceedingly"[72] and "spectacularly"[73] transformative. The court stated:

> "[I]f someone were to read all the modern-day classics because of their exceptional expression, memorize them, and then emulate a blend of their best writing, would that violate the Copyright Act? Of course not." … "Like any reader aspiring to be a writer, Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them — but to turn a hard corner and create something different. If this training process reasonably required making copies within the LLM or otherwise, those copies were engaged in a transformative use."[74]

The *Bartz* court made this emphatic pronouncement based on the ingestion of the data for purposes of training the LLM – in other words the separation of the data into statistical weights and parameters to train an algorithmic model. The court was careful to note that the plaintiffs were *not* asserting that users of the defendant's LLM could retrieve infringing output. "…Authors do not allege that any infringing copy of their works was or would ever be provided to users by the Claude service. Yes, Claude could help less capable writers create works as well-written as Authors' and competing in the same categories. But Claude created no exact copy, nor any substantial knock-off. Nothing traceable to

[68] *Bartz v. Anthropic PBC*, 3:24-cv-05417, (N.D. Cal. Jun 23, 2025) ECF No. 231, at *2 [hereinafter *Bartz v. Anthropic* or *Bartz*].
[69] *Kadrey v. Meta Platforms, Inc.*, No. 23-CV-03417-VC, 2025 WL 1752484, at *2 (N.D. Cal. June 25, 2025) [hereinafter *Kadrey v. Meta* or *Kadrey*].
[70] *Campbell*, 510 U.S. at 579.
[71] *See Campbell*, 510 U.S. at 584.
[72] *Bartz*, at *5.
[73] *Bartz*, at *7.
[74] *Bartz*, at *8.

Authors' works. Such allegations are simply not part of plaintiffs' amended complaint, nor in our record."[75]

While the *Kadre*y court was clear that "[t]here is no serious question that Meta's use of the plaintiffs' books had a 'further purpose' and 'different character' than the books" and thus that use of the works was "transformative,"[76] the court also took the view that the fourth factor (addressing market impact) should be weighted more heavily than the first factor in this case[77] Nevertheless, the *Kadrey* court homed in on the technical differences between the way a human consumes copyrighted content and an LLM: "an LLM's consumption of a book is different than a person's" because it learns "statistical patterns," and because it enables access to that learning without having read the book.[78] The commercial nature of the use does not tip the balance away from it being transformative: "commercialism isn't dispositive of the first factor and tends to be less important when the secondary use is highly transformative."[79]

Nature of the Work

The second fair use factor looks at the nature of the allegedly infringed work and is more likely to favor fair use when the work more factual in nature. Both courts agreed without much ado that this factor weighed against fair use. The copyrighted content in both lawsuits are books, which are highly expressive works "of the type that the copyright laws value and seek to protect."[80]

Amount Used

The third fair use factor looks at the amount of the content used in the alleged infringement and generally favors uses that use less of the original work. Both courts recognized that while the content was copied and used in its entirety, doing so was necessary to fulfill the transformative purpose (in this case, training the LLM).[81]

Market Effect of the Use

The fourth fair use factor looks at the economic effects of the alleged infringement, and whether the use substitutes for economic benefits that should rightly go to the rightsholder. This was the factor that received the most disparate treatment by the two courts. In *Bartz*, the court explained that the copying does not displace the demand for copies of the authors' works "in the way that counts under the Copyright Act"[82] and a licensing market for LLM training "is not one the Copyright Act entitles Authors to exploit."[83] By analogy, the *Bartz* court explained, the "[a]uthors' complaint is no different than it would be if they complained that training schoolchildren to write well would result in an explosion of competing

[75] *Bartz*, at *4.
[76] *Kadrey*, at *9.
[77] *Kadrey*, at *2.
[78] *Kadrey*, at *10.
[79] *Kadrey*, at *11 (citing, *inter alia*, *Google LLC v. Oracle America, Inc.*, 593 U.S. 1, 32 (2021)).
[80] *Kadrey*, at *13 (quoting *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 187 (2d Cir. 2024)); *Bartz*, at *14-15.
[81] *Kadrey*, at *14-*15; *Bartz*, at *15-*16.
[82] *Bartz*, at *16.
[83] *Bartz*, at *17.

works."[84] The *Bartz* court also noted,[85] as did the *Kadrey* court,[86] that there was no evidence of infringing output from use of the LLM, so there was no direct market substitution for the authors' works.

The *Kadrey* court expressed a narrow market view, but used a different rationale than the *Bartz* court, finding that "the plaintiffs are not entitled to the market for licensing their works as AI training data. … [The market to license works for training purposes] is not one that the plaintiffs are legally entitled to monopolize. … [T]o prevent the fourth factor analysis from becoming circular and favoring the rightsholder in every case, harm from the loss of fees paid to license a work for a transformative purpose is not cognizable."[87] The *Kadrey* court did not end its analysis there, however. It expressly noted that, although these plaintiffs had not sufficiently met their evidentiary burden, there was a "potentially winning argument—that Meta has copied their works to create a product that will likely flood the market with similar works, causing market dilution."[88] Unfortunately for the plaintiffs, "they present[ed] no evidence about how the current or expected outputs from Meta's models would dilute the market for their own works."[89]

While the *Kadrey* court found the plaintiffs' evidence lacking, the court also signaled heavily that using copyrighted content for LLM training works against the overall goal of copyright, saying, "by training generative AI models with copyrighted works, companies are creating something that often will dramatically undermine the market for those works, and thus dramatically undermine the incentive for human beings to create things the old-fashioned way.".[90] The opinion concludes by stating:

> "Given the state of the record, the Court has no choice but to grant summary judgment to Meta on the plaintiffs' claim that the company violated copyright law by training its models with their books. … And, as should now be clear, this ruling does not stand for the proposition that Meta's use of copyrighted materials to train its language models is lawful. It stands only for the proposition that these plaintiffs made the wrong arguments and failed to develop a record in support of the right one."[91]

Judge Chhabria also directly criticized the analogy used by the *Bartz* court, "when it comes to market effects, using books to teach children to write is not remotely like using books to create a product that a single individual could employ to generate countless competing works with a miniscule fraction of the time and creativity it would otherwise take. This inapt analogy is not a basis for blowing off the most important factor in the fair use analysis."[92]

Much disagreement still remains over how to define the market relevant to a fair use analysis. In the U.S. Copyright Office's 2025 Report, various viewpoints emerged. The Copyright Alliance, for example, argued that "with generative AI, the harm is often to a creator's overall body of work or even the market more broadly. These harms all impact the creator's incentives…" [93] Meanwhile, tech providers

---

[84] *Bartz*, at *17.
[85] *Bartz*, at *17.
[86] *Kadrey*, at *15.
[87] *Kadrey*, at *2, *16.
[88] *Kadrey*, at *2.
[89] *Kadrey*, at *2.
[90] Kadrey, at *1.
[91] *Kadrey*, at *3.
[92] *Kadrey*, at *2.
[93] *USCO Report Part 3*, at 64.

argue that "the fourth factor analysis considers only harm to markets for the specific copyrighted work."[94] In short, those with less to gain from fair use (e.g., the original copyright holders) generally see a broad market under the fourth fair use factor; those seeking to fairly use text for AI training (e.g., AI providers, among others) generally see a narrow market. With the contradictory handling of the market use factor in *Bartz* and *Kadrey*, and this still-active regulatory debate, whether there is a market for licensing copyrighted content for generative AI purposes—and a market that is cognizable under fair use analysis—is still yet to be seen.

Perhaps noteworthy for future litigation under this factor, the *Kadrey* court includes a potential roadmap for plaintiffs to win on the fourth factor if a court were to adopt the *Kadrey* market dilution approach. The court suggests that answers to these questions would help determine market dilution: "is Llama capable of generating such books?" "[W]hat are these AI-generated books?" And to what extent could the book market be (or has been) diluted by AI-generated titles?[95]

The *Kadrey* court differentiated this market dilution approach from direct substitution,[96] and the court credited the evidence in the case demonstrating that "Llama does not allow users to generate any meaningful portion of the plaintiffs' books… Llama's ability to regurgitate miniscule portions of the plaintiffs' books if manipulated into doing so does not threaten to have a 'meaningful or significant effect 'upon the potential market for or value of' the plaintiffs' books."[97]

Source of Copyrighted Content

Another topic addressed by these opinions is the use of pirated copies. The *Bartz* court found fair use as a valid defense to Anthropic's *training* the LLM on pirated data, but specifically did not find that the defense protected Anthropic from copyright infringement liability for downloading the copyrighted works from pirated sites and using those works to create a central library (from which Anthropic could pull and use various data training sets for training).[98] In *Kadrey*, the court explained that the "bad faith" resulting from the use of pirated content is only "potentially relevant" to the "character" of Meta's use under the first factor, but that such bad faith did not sway the scales against a finding for Meta on the first factor, given the overwhelming transformative nature of the use.[99]

As of the writing of this white paper, the *Bartz* court certified a class of authors and publishers harmed by Anthropic's use of pirated copies. Anthropic and the plaintiffs entered into a preliminary settlement on this discrete issue, which was approved by the court on July 20, 2026. MIT Press is a class member because it is a copyright owner of content Anthropic accessed on pirated websites, and therefore, unless an appeal should reverse the court's final approval of the settlement, the MIT Press will earn a monetary amount as part of the settlement agreement. The amount allocated for the MIT Press works will

[94] *USCO Report Part 3*, at 64.

[95] *Kadrey*, at *19 ("[W]hat impact does this competition actually have on sales of the books it competes with?"). The Court also asked: "how does the threat to the market for the plaintiffs' books in a world where LLM developers can copy those books compare to the threat to the market for the plaintiffs' books in a world where the developers can't copy them?"

[96] *Kadrey*, at *15.

[97] *Kadrey*, at *15 (citing *Authors Guild*, at 224; quoting 17 U.S.C. § 107(4)).

[98] *Bartz*, at *18.

[99] *Kadrey*, at *11.

be shared with the authors of such works, per the court's class certification and the settlement agreement.[100]

Reuters v. Ross

The earlier 2025 opinion in *Reuters v. Ross*[101] provides the only AI-related opinion yet issued which found that the challenged use was not a fair use. The case is briefly described here to underscore that the fair use analysis is complex and different circumstances may sway cases in either direction, and that this legal landscape is still far from settled. In this case, a developer, Ross, allegedly scraped text of "headnotes" (summaries of key case elements) from Reuters' legal information database for use in an AI tool. The court found that the first and fourth fair use factors did not support fair use due to the *non*-transformative use meant to create a similar output that competes with Reuters' product.[102]

Case Law Summary

Given the number of lawsuits still pending on these issues and the early stage of litigation that most cases are in, it is hard to predict where the law will land. However, at least some commercial use has been ruled "fair," and these cases may be persuasive. Further, the current decisions may call into question some major licensing deals, although those deals may still have merit given the unknowns surrounding training on pirated works and other considerations.

Current Copyright Office Guidance

In addition to litigation as a definite authority, the US Copyright Office has published the findings of its studies concerning multiple aspects of copyright with regard to AI.[103] While the legal authority of agency proclamations is complex,[104] the publications at least have some persuasive authority.

The current Copyright Office guidance[105] (a pre-release draft that was made public before verdicts in *Bartz* and *Kadrey* came out) cautions that fair use will be heavily fact dependent, stating "various uses of copyrighted works in AI training are likely to be transformative. The extent to which they are fair, however, will depend on what works were used, from what source, for what purpose, and with what controls on the outputs—all of which can affect the market."[106] Much controversy arose from

[100] *Bartz v. Anthropic*, N.D. Cal. Case No. 3:24-cv-05417-WHA, [Amended Proposed] Order Granting Preliminary Approval of Class Action Settlement (Sept. 25, 2025).
[101] *Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.*, 765 F. Supp. 3d 382 (D. Del. 2025), *motion to certify appeal granted*, No. 1:20-CV-613-SB, 2025 WL 1488015 (D. Del. May 23, 2025) [hereinafter *Reuters v. Ross* or *Ross*].
[102] *Ross*, at 12.
[103] U.S. Copyright Office, *Copyright and Artificial Intelligence – Part 1: Digital Replicas* (July 2024), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf; U.S. Copyright Office, *Copyright and Artificial Intelligence – Part 2: Copyrightability* (Jan. 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf; *USCO Report Part 3.*
[104] *See generally Loper Bright Enters. v. Raimondo*, 143 S. Ct. 2429 (2023); *Skidmore v. Swift & Co.*, 323 U.S. 134, 65 S. Ct. 161 (1944).
[105] *See generally USCO Report Part 3.*
[106] *USCO Report Part 3*, at 107-8.

the report,[107] and it is unclear if, as of this writing, a redaction or revision may come. The report itself, however, highlighted the sharp divide in perspectives on what industry commentators viewed as fair use. A summary of various perspectives for all four fair use factors is listed in Table 1 below. Brackets list the entities which submitted arguments in favor the stated position.

Factor	Arguments For Fair Use	Arguments Against Fair Use
Factor 1: Purpose and Character of the Use[108]	- Training is *highly transformative* as it "predict[s] or classify[ies] aspects of copyright-protected inputs" which is a "distinct purpose" [Computer and Communications Industry Association, , Meta, Electronic Frontier Foundation, University Library of the University of California, et al.] - Use is *non-expressive* and computational and therefore for a different purpose than the original expressive content [Anthropic, Google, et al.] - Substantial *public benefits* support a finding of transformativeness (education, accessibility, economic growth) [IBM]	- Use is *not truly transformative*, just ingestion, similar to compression [National Music Publishers Association (NMPA), Evangelical Christian Publishers Association, NYTimes] - No particular, individual work is needed to train a model, so model developers do not *need* to use copyrighted works for training [Association of American Publishers (AAP)] - the ultimate use of the model may not be distinct from original purpose of creative work [Copyright Alliance] - use is not "non-expressive" because systems lack human creativity and thus need to take the expressive content (e.g. style, themes) from copyrighted works [Authors Guild, AAP]
Factor 2: Nature of the Copyrighted Work[109]	- Models that draw from *factual or functional* works are more likely to support fair use argument [Authors Guild] - Models typically include *published* content, which supports a fair use argument [New Media Rights]	- *Highly creative works* are typically used for training (fiction, music, art) [Author's Guild, NMPA, Digital Media Licensing Association (DMLA)] - The use of unpublished or not fully published content weighs against fair use [Copyright Alliance, Data Provenance Initiative]

[107] *See, e.g.,* Andrew Limbong, The U.S. Copyright Office used to be fairly low-drama. Not anymore, Morning Edition, National Public Radio (June 6, 2025) https://www.npr.org/2025/06/06/nx-s1-5399781/copyright-office-explainer-perlmutter-trump.

[108] *USCO Report Part 3*, at 41-45.

[109] *USCO Report Part 3*, at 53-54.

Factor	Arguments For Fair Use	Arguments Against Fair Use
Factor 3: Amount and Substantiality of the Portion Used[110]	- *Full work copying* is functionally necessary for model accuracy and greatest public benefit [OpenAI] - Even substantial use is fair because it is not being used as a substitute for the work because the output is new [Prof. Samuelson, Sag, Sprigman]	- Models ingest *entire works*, often indiscriminately, differentiating from search engine/Google Books type uses [NMPA] - Scale does not matter; a small, but high quality portion of expressive content may be valuable and either could be excluded from the model's training set or, if necessary, demonstrates its value [A2IM-RIAA, Copyright Alliance, Getty Images] - Unlike Google Books due to substitution potential and lack of safeguards [Copyright Alliance]
Factor 4: Effect on the Market[111]	- *No substitution*—training does not replace the original, output is new type of content with new competitive market [Meta, Author's Alliance, Hugging Face] - *No well-established licensing market* exists yet and quality/quantity needed cannot be obtained through licensing because of administrative and transactional cost, and impracticalities (such as locating the author) [Hugging Face, Anthropic, R Street, EFF, et al.] - Public benefit of training quality models offsets copyright concerns [OpenAI, TechNet, Meta, et al.]	- Produces nearly verbatim copyrighted content and is therefore substitution for the original work [Natl. Assoc. of Broadcasters, Center for Art Law, A2IM Recording Academy-RIAA] - use without permission *undermines licensing opportunities* (e.g., for training rights) [A2IM-RIAA] - *Outputs compete* with originals and dilute market, which can also disincentivize future human creation [ASCAP, Music Workers Alliance, UMG Records, Copyright Alliance] - Threatens authorship entirely and reduces value of copyright; requires human authors to compete with the AI tools trained to mimic them [CISAC, Prof. David Newhoff, AAP] - A licensing market does exist and is growing, as demonstrated by existing deals; therefore, there is a way for model owners to access content and

[110] *USCO Report Part 3*, at 54-60.
[111] *USCO Report Part 3*, at 61-73.

Factor	Arguments For Fair Use	Arguments Against Fair Use
		compensate authors [AAP, Scientific Technical Medical Publishers]

Conclusion

In addition to continuing to monitor pending litigation, it will be important to monitor any new or updated guidance from the US Copyright Office on these issues. Notably, certain interest groups supportive of unregulated AI technology companies have petitioned the Department of Justice to intervene in copyright infringement lawsuits against AI developers in order to prevent AI companies from being saddled with consequential liability.[112] It would not be surprising if other lobbying efforts were underway to encourage legislative action, as well, in favor of technology companies that seek to develop generative AI without hindrance from content creators.

On a smaller scale, we will also be keeping track of content licensing deals as such may occur in the industry more generally and as those opportunities are discussed on campus. The MIT Press continues to evaluate the potential benefits of licensing its content, and the Libraries continues to take part in conversations with third party publishers and its peers in regard to ensuring access to content for all MIT researcher needs.

Meanwhile, these legal and licensing matters will evolve as generative AI continues to advance itself. Factors that have exacerbated the "unfairness" of use of content without permission, such as the inability to attribute a work used to train or the inability to prevent outputs that are substantially similar to original content, may find technical fixes. Simultaneously, as AI tools demand more and more high quality content, there may be greater urgency to preserve and protect human authored content, especially content published by responsible and sophisticated publishers.

[112] *Tech-Backed Group Seeks DOJ's Assistance in AI Copyright Disputes*, International Intellectual Property Law Association, https://iipla.org/tech-backed-group-seeks-dojs-assistance-in-ai-copyright-disputes/ (last accessed, November 10, 2025).