



The Library

Berkeley, CA 94720-6000

October 30, 2023

Suzanne V. Wilson
General Counsel and Associate Register of Copyrights
U.S. Copyright Office
101 Independence Ave, S.E. Washington, D.C. 20559-6000

Re: Notice of inquiry (“NOI”) and request for comments, Artificial Intelligence and Copyright, Docket No. 2023-6

Dear Associate Register Wilson:

The University Library of the University of California, Berkeley (“UCB Library”) is grateful for the opportunity to submit this response (“Response”) to the Copyright Office’s NOI.¹ The copyright law and policy landscape underpinning the use of AI models is complex, and regulatory decision-making in the copyright sphere will bear ramifications for global enterprise, innovation, and trade. The NOI thus raises important and timely legal questions, many of which we are only beginning to understand.

For these reasons, we take a cautious and narrow approach with our Response. We address key precepts known about fair use and licensing, as these issues bear upon the nonprofit education, research, and scholarship undertaken by scholars who rely on AI models. In this regard, we limit our Response to two foundational principles essential for guiding any subsequent Copyright Office rulemaking:

- First, training AI models by using copyright-protected inputs falls squarely within what courts have determined to be fair use, and particularly so when that AI training is for nonprofit educational or research purposes (**NOI Question 8**).
- Second, to the extent that the Copyright Office is inclined to create a regulatory right for content creators to opt out of having their works included in AI training, such an opt-out provision should not be extended to any AI training or processes that constitute fair uses (**NOI Question 9**).

To establish these cornerstones, it will be helpful first for us to contextualize the kind of AI model-reliant research and educational activities undertaken by scholars (**NOI Question 6**) and explain an academic library’s role in supporting these activities (**NOI Question 7**).

I. NOI Questions 6 and 7: AI model-reliant research and education, and libraries’ roles in supporting it

The University of California produces over 8% of all scholarly publishing in the U.S. and performs 10%

¹ The University of California, Merced Library has also endorsed and signed this Response.

of all U.S. academic research and development.² Our campus scholars include faculty, graduate and undergraduate students, postdoctoral researchers, and professional staff who rely on access to a broad array of academic, cultural, and creative content to power their research and scholarship. The resources they utilize include scholarly publications, databases, data, literary works, social media, and more—all of which they draw upon, incorporate, cite, and expand upon in their research, publishing, and teaching. In performing this work, scholars require a flexible and permissive statutory and regulatory environment, as they rely heavily on the exercise of limitations and exceptions to copyright such as fair use, and must leverage institutional license agreements that UCB Library negotiates at great cost to preserve essential copyright exceptions (as discussed further below).

A. Text and data mining; AI model-reliant research and education

With increasing importance and frequency, scholars at institutions like ours have employed computational tools, algorithms, and automated techniques to extract revelatory information from large sets of unstructured or thinly-structured digital content. This process is broadly known as text and data mining (TDM). Text and data mining allows researchers to identify and analyze patterns, trends, and relationships across volumes of data that would otherwise be impossible to sift through.³ TDM enables the exploration of issues like: racial disparity evidenced through police body camera footage;⁴ changes in gender significance in fiction;⁵ and public discussions of social justice issues like violence against women.⁶ The particular TDM methodologies employed continue to expand, enabling advancements across education, literature, society, politics, and beyond.⁷

Not all TDM research methodologies necessitate usage of AI models. For instance, sometimes TDM can be performed by developing algorithms to detect the frequency of certain words within a corpus, or to parse sentiments based on the proximity of various words to each other.⁸ In other cases, though, scholars must employ machine learning techniques to train AI models before the models can make a variety of assessments.

The following example illustrates this distinction: Imagine a scholar wishes to assess the prevalence with which 20th century fiction authors write about notions of happiness. The scholar likely would

² University of California. *Accountability Report 2019: Research*. Retrieved October 3, 2023, from <https://accountability.universityofcalifornia.edu/2019/chapters/chapter-9.html>

³ Hearst, Marti A. (2003, October 17). What is text mining? <http://people.ischool.berkeley.edu/~hearst/text-mining.html>

⁴ Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., and Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25), 6521. <https://doi.org/10.1073/pnas.1702413114>

⁵ Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. *Cultural analytics*. <https://doi.org/10.22148/16.019>

⁶ Xue, J., Macropol, K., Jia, Y., Zhu, T., and Gelles, R. J. (2019). Harnessing big data for social justice: An exploration of violence against women-related conversations on Twitter. *Human Behavior and Emerging Technologies*, 1(3), 269–279. <https://doi.org/10.1002/hbe2.160>

⁷ Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, 4(1). <https://doi.org/10.3390/bdcc4010001>

⁸ See, for example Google Research. *Google Books Ngram Viewer*. Retrieved October 3, 2023, from <https://books.google.com/ngrams/info> and Sentiment analysis. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=1178380470

compile a corpus of thousands or tens of thousands of works of fiction, and then run a search algorithm across the corpus to detect the occurrence or frequency of words like “happiness,” “joy,” “mirth,” “contentment,” and synonyms and variations thereof. But if a scholar instead wanted to establish the presence of fictional characters who embody or display characteristics of *being* happy, the scholar would need to employ “discriminative modeling” (a classification and regression technique) that can train AI to recognize the appearance of happiness by looking for recurring indicia of character psychology, behavior, attitude, conversational tone, demeanor, appearance, and more.⁹

This latter scenario of AI modeling is supported by a growing number of UCB research grants. For instance, Professor David Bamman has obtained a grant to leverage AI modeling to investigate the representation of race, gender, and place in both popular and prestige films and TV shows, and answer questions such as: How are race and gender tied to the depiction of characters on screen, and how has this changed over the past 50 years? How much attention is given to cities vs. rural environments? And how might this kind of representation on screen shape the development of stereotypes in viewers?¹⁰ The UCB Library’s role in this grant project will be to analyze how these techniques demonstrate the value and limitations of a new federal regulatory exemption that allows researchers to break the digital locks on electronic materials, making the kind of research Bamman is undertaking possible.¹¹

B. Role of academic libraries

To support TDM and other computational research, the UCB Library invests approximately \$12 million each year licensing electronic content. We secure campus access to a wide range of scholarly resources including books (print and digitized), scientific journals, databases, multimedia resources, and other materials. In doing so, we endeavor to negotiate “researcher-friendly” licensing terms that preserve fair use and TDM rights, which are essential to ensuring that scholars can lawfully make use of the materials we have procured.

The UCB Library’s role is not just one of content acquirer, however: We are also instructors and advisors. Our Office of Scholarly Communication Services (OSCS) is a team of copyright law and information policy (i.e. licensing, privacy, and ethics) experts who help scholars understand the ecosystem of scholarly resources available to them, and then navigate legal, ethical, and policy considerations in utilizing these resources in their research.¹² Given the research and publishing volume on a campus like UC Berkeley, in FY22-23, OSCS were called upon to provide 2,838

⁹ The authors differentiate discriminative modeling (classification and regression) from generative modeling (systems capable of producing outputs such as text or images). Lee, K., Cooper, A. F., & Grimmelmann, J. (2023). *Talkin’ Bout AI Generation: Copyright and the Generative AI Supply Chain* (SSRN Scholarly Paper 4523551). <https://doi.org/10.2139/ssrn.4523551>, p. 11.

¹⁰ UC Berkeley Library Communications. (2023, January 26). *Poetry, mining Hollywood, and digital books: Mellon Foundation grants will support groundbreaking work at the UC Berkeley Library and beyond*. <https://www.lib.berkeley.edu/about/news/mellon-grants>

¹¹ For more examples of AI model-reliant research, please see the NOI response being submitted by the Authors Alliance organization which incorporates perspectives from researchers.

¹² UC Berkeley Library. *Scholarly Communication Services*. Retrieved October 3, 2023, from <https://www.lib.berkeley.edu/research/scholarly-communication>

consultations regarding copyright and information policy, including TDM and AI.¹³ OSCS have also developed informational guides, workshops, and videos related to TDM, copyright, and information policy issues in research.¹⁴

These services build upon OSCS' national and international reputation as leaders supporting TDM research. In 2020, OSCS led an NEH-sponsored institute to train scholars in "Building Legal Literacies for Text Data Mining (Building LLTDM)," ¹⁵ which led to follow-on NEH-funded work in 2022 to develop guidance for scholars conducting TDM across international borders (Legal Literacies for Text Data Mining—Cross-border (LLTDM-X))¹⁶. The outputs of these projects (an open educational resource¹⁷, case study¹⁸, and analytical white papers¹⁹) are now regularly relied upon by U.S. and global scholars in navigating copyright issues in TDM research. OSCS also collaborated with the UC Berkeley Samuelson Law, Technology & Public Policy Clinic to help secure a new Digital Millennium Copyright Act exemption that allows scholars to conduct text and data mining on literary works and motion pictures—materials that otherwise would have been off-limits for computational analysis because of technological protection measures.²⁰ And OSCS continues to advocate for harmonization in TDM legal protections worldwide.²¹

A key takeaway from the substantial work that OSCS (and the entire UCB Library) does to support TDM and AI model-reliant research and education is: scholars engaging in computational research (i) rely on fair use to create and analyze corpora of copyright-protected texts and materials; and (ii) rely on a licensing and regulatory environment that does not impinge upon their statutory rights. These two issues are discussed further in our responses to Questions 8 and 9, respectively.

II. NOI Question 8: Training AI models with copyrighted works is a fair use

The fair use provision of the U.S. Copyright Act (17 U.S.C. § 107), and its first factor in particular, requires an analysis of the specific "use" of a copyrighted work that is alleged to be an infringement.

¹³ UC Berkeley Library Office of Scholarly Communication. *Annual Report FY22–23*. Retrieved September 23, 2023, from

<https://docs.google.com/document/d/1WCVSU6jNj8Kkt3zheY84LI8l5LS5fyFVXhycaQ7p0U/edit?usp=sharing>

¹⁴ See, e.g., UC Berkeley Library Office of Scholarly Communication. *Text Data Mining*. Retrieved September 23, 2023, from <https://www.lib.berkeley.edu/research/scholarly-communication/copyright?section=text-data-mining> and UC Berkeley Library Office of Scholarly Communication. *YouTube*. Retrieved September 23, 2023, from

<https://www.youtube.com/channel/UCNUMwTyK0raTNNZVjhqB7KA>

¹⁵ *Building LLTDM Institute*. <https://buildinglltdm.org/institute/>

¹⁶ *LLTDM-X*. <https://buildinglltdm.org/lltdmx/>

¹⁷ Althaus, S., Bamman, D., Benson, S., Butler, B., et al. (2021). *Building Legal Literacies for Text Data Mining*. University of California, Berkeley. <https://doi.org/10.48451/S1159P>

¹⁸ Samberg, R., Vollmer, T., & Padilla, T. (2023). *Legal Literacies for Text Data Mining – Cross-Border ("LLTDM-X"): Case Study*. <https://escholarship.org/uc/item/1w03f9r2>

¹⁹ See Samberg, R., & Vollmer, T. (2021). *Building Legal Literacies for Text Data Mining: Institute White Paper*. <https://escholarship.org/uc/item/1db5350t> and Samberg, R., Vollmer, T., & Padilla, T. (2023). *Legal Literacies for Text Data Mining – Cross-Border ("LLTDM-X"): White Paper*. <https://escholarship.org/uc/item/5k91r1s1>

²⁰ Authors Alliance. (2021, October 27). Update: Librarian of Congress Grants 1201 Exemption to Enable Text Data Mining Research. *Authors Alliance*. <https://www.authorsalliance.org/2021/10/27/update-librarian-of-congress-grants-1201-exemption-to-enable-text-data-mining-research/>

²¹ Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., et al. (2022). Legal reform to enhance global text and data mining research. *Science*, 378(6623), 951–953. <https://doi.org/10.1126/science.add6124>

(*Warhol v. Goldsmith*, 143 S. Ct. 1258, 1263-1264 (2023)). In the context of AI models, training the model to predict or classify aspects of copyright-protected inputs is a distinct purpose, and one that is highly transformative from the original “consumptive” purpose of those copyrighted works.

We agree with legal scholars who explain that the determination of transformativeness and the overall fair use of generative AI *outputs* cannot always be predicted in advance: The mechanics of LLM operability (framed perhaps best by scholar Matthew Sag in his recent paper²²) suggest that there are limited instances in which generative AI outputs could indeed be substantially similar to (and potentially infringing of) the underlying works used for training; this substantial similarity is possible typically only when a training corpus is rife with numerous copies of the same work.²³ Given the case-by-case nature of evaluating substantial similarity of outputs, we take no position as to whether *output* content created by generative AI is transformative or infringing.

However, we likewise align with legal scholars who explain that the **training of AI LLMs by using copyright-protected inputs falls squarely within what courts have determined to be a transformative fair use, especially when that training is for nonprofit educational or research purposes**.²⁴ And it is essential to protect the fair use rights of scholars and researchers to make these uses of copyright-protected works when training AI, as we detail below.

Because exclusivity can have negative consequences on the progress of science and knowledge, the law has “limited the scope of copyright protection to ensure that a copyright holder’s monopoly does not harm the public interest.” (*Google LLC v. Oracle Am., Inc.*, 141 S.Ct. 1183, 1187 (2021)). One way Congress and the courts have protected the public interest is through the fair use right, which is a “flexible” right to adapt to developments in technology. (*Id.* at 1197). Over the past decade, fair use has been interpreted by the courts²⁵ and the Copyright Office²⁶ to permit the reproduction of copyrighted works to create and mine a corpus of copyright-protected works—including works that can be used to train generative or non-generative AI. These authorities further hold that making derived data, results, abstractions, metadata, or analysis from the copyright-protected corpus available to the public is also fair use, as long as the research methodologies or data distribution processes do not re-express the

²² Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper 4438593). <https://doi.org/10.2139/ssrn.4438593>

²³ *Id.* at 7.

²⁴ That an AI model could arguably be used for some other, infringing purposes is of no moment to the assessment of the particular use—i.e., the training of AI—at issue. *Warhol*, 143 S. Ct. 1258 (2023) (distinguishing between usage purposes for fair use analysis). In any event, LLMs do not create derivative works of or redistribute the corpus materials, but rather learn *from* them. See Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper 4438593). <https://doi.org/10.2139/ssrn.4438593> p. 6-7. As such, even subsequent licensing of a trained AI model would not infringe the original works with distributed reproductions; only particular *outputs* of the model could (in rare cases) produce anything substantially similar to an original work. *Id.*

²⁵ See *Authors Guild v. Google, Inc.*, 804 F.3d 202, 215 (2d Cir. 2015); *Authors Guild, Inc. v. HathiTrust* 755 F.3d 87, 105 (2d Cir. 2014); *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630 (4th Cir. 2009).

²⁶ In evaluating the proposed DMCA § 1201 exemption to circumvent technological protection measures on DVDs and eBooks for the purpose of conducting TDM, the USCO writes: “Balancing the four fair use factors, with the limitations discussed, the Register concludes that the proposed use is likely to be a fair use.” See U.S. Copyright Office. *Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention – Recommendation of the Register of Copyrights – October 2021*. https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf

underlying works to the public in a way that could supplant the market for the originals.²⁷

Of particular relevance for the AI training fair use inquiry will be Factors One and Four. This is because courts have already opined on the non-determinative nature of Factors Two and Three for computational uses:

- As a preliminary matter, fair use's "second factor [i.e., the nature of the underlying work] has rarely played a significant role in the determination of a fair use dispute" (*Authors Guild v. Google, Inc.* 804 F.3d 202 at 220). This is particularly true in TDM scenarios when the nature of the work is secondary to the transformativeness of the purpose (i.e., to understand relationships, patterns, etc.); in these cases, the content being studied or mined can inherently be either factual or expressive, and the nature of the work is ancillary to the non-protectable information being extracted *about* the work. Indeed, the courts adjudicating corpus-mining scenarios would not have been persuaded by Factor Two even if all of the works at issue had been highly expressive works of fiction. (*Id.* at 220). For this reason, Factor Two has a neutral outcome for fair use balancing.
- Likewise, for Factor Three, the courts have "rejected any categorical rule that a copying of the entirety cannot be a fair use. Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original." (*Id.* at 221). Accordingly, courts have ruled that the use of the entire work is "literally necessary" to achieve the intended transformative search purposes of corpus mining.²⁸ So, like Factor Two, Factor Three also has a neutral outcome for fair use balancing.

With fair use Factors Two and Three having minimal impact in computational research cases, fair use Factors One and Four remain most consequential. To that end, Factor One considers "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes." As the Supreme Court made clear in *Google*, 141 S. Ct. at 1183, "[t]he inquiry into 'the purpose and character' of the use turns in large measure on whether the copying at issue was 'transformative,' i.e., whether it 'adds something new, with a further purpose or different character.'" [citing *Campbell v. Acuff-Rose*, 510 U.S. 569 (1994)]. In the context of expressive texts or images that are originally intended to be "consumed," reproduction for the purposes of identifying or teaching facts or trends *about* those texts or images represents such a further purpose.

Accordingly, all of the following relevant uses have been deemed transformative for purposes of Factor One: (a) creation of a full-text searchable digital library that displays word count results (*Authors Guild*

²⁷ The findings in these matters also reinforce that copying done as part of a process to produce non-expressive or non-infringing content (such as patterns or data) is not an infringement in part because copyright protection does not extend to facts or ideas. *Google*, 141 S. Ct. at 1187.

²⁸ In *Authors Guild v. Google* (804 F.3d at 221), the Court explained: "In *HathiTrust*, our court concluded in its discussion of the third factor that '[b]ecause it was reasonably necessary for the [HathiTrust Digital Library] to make use of the entirety of the works in order to enable the full-text search function, we do not believe the copying was excessive.' 755 F.3d at 98. As with *HathiTrust*, not only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose. If Google copied less than the totality of the originals, its search function could not advise searchers reliably whether their searched term appears in a book (or how many times)."

v. HathiTrust, 755 F.3d 87 (2d Cir. 2014)); (b) creation of a full-text searchable database with “snippet view” and “ngram viewer” [search strings] *Authors Guild v. Google*, 804 F.3d 202 (2d Cir. 2015); and (c) use of machine learning on a corpus so that the AI can detect the occurrence of plagiarism in other inputs (*A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630 (4th Cir. 2009))²⁹. Not only would training AI for any purpose be considered transformative under Factor One for each of these holdings, but also the specific *scholarly* purposes of AI training contemplated by this Response further satisfy the “nonprofit educational purpose” prong of Factor One, again favoring fair use.

Factor Four considers “the effect of the use upon the potential market for or value of the copyrighted work.” The inquiry for this factor is not whether there is a suppression of the market for the copyrighted works, but rather whether the use supplants or serves as a substitute for that market.³⁰ The courts and the Copyright Office have consistently recognized that the non-expressive use of copyrighted works does not substitute for the copyright owners’ ability to sell or communicate their copyrighted works to the public.³¹ The lawful creation of corpora through reproduction of copyright-protected works means that copyright owners have already made lawful sales or licenses of their works (or have made them available at no cost to the public). There is no supplantation of that market, but rather, that market has functioned successfully. Moreover, when the use is highly transformative under Factor One—as courts have found corpus creation to extract patterns and information to be—“the more it serves copyright’s goal of enriching public knowledge and the less likely it is that the appropriation will serve as a substitute for the original or its plausible derivatives” for purposes of Factor Four. (804 F.3d at 214).

To further affirm the application of these TDM holdings to AI training for research and educational purposes,³² we can return to the context of a scholar building an AI tool to detect “happy characters” in contemporary fiction. The scholar may teach the AI model what happiness “looks like” by displaying textual passages embodying the psychology, speech patterns, or decisions made by characters perceived to be “happy,” and then applying regression techniques to affirm or discount the model’s attempts at recognition. Once the AI becomes an “expert” in its happiness detection skills, the scholar

²⁹ Although the term “machine learning” is not mentioned in the *iParadigms* opinion, scholars understand the *iParadigms* algorithm to function only having once been trained to characterize passages.

³⁰ See e.g. *Authors Guild v. Google*, 804 F.3d at 207 (“Google’s making of a digital copy to provide a search [or snippet] function is a transformative use, which augments public knowledge by making available information about Plaintiffs’ books without providing the public with a substantial substitute for matter protected by the Plaintiffs’ copyright interests in the original works or derivatives of them.”) Moreover, the suppression of the market matters only where the unauthorized uses are those protectable by copyright to begin with, rather than unprotectable interests like facts, or non-expressive uses. See Sag, M. (2019). *The New Legal Landscape for Text Mining and Machine Learning* (SSRN Scholarly Paper 3331606). <https://doi.org/10.2139/ssrn.3331606>

³¹ *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *Authors Guild v. Google*, 804 F.3d 202 (2d Cir. 2015); *A.V. ex rel. Vanderhye v. iParadigms*, 562 F.3d 630 (4th Cir. 2009)

³² Prof. Matthew Sag writes, “The suggestion that the broad affordance for text data mining as fair use announced in *HathiTrust* does not apply to machine learning is confounding. There is no principled reason why deriving metadata through technical acts of copying and analyzing that metadata through logistic regression should be fair use, but analyzing that data by training a machine learning classifier to perform a different kind of logistic regression that produces a predictive model wouldn’t be.” Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper 4438593). <https://doi.org/10.2139/ssrn.4438593>, p. 10. See also Butler, B. (2023, May 8). What AI can teach us about copyright and fair use. *Freethink*.

<https://www.freethink.com/robots-ai/what-ai-can-teach-us-about-copyright-and-fair-use> (“there’s “no meaningful difference between these [AI] tools and the other ‘non-consumptive’ / computational uses that courts have already blessed as fair use many times over.”)

may then present it with new works of fiction to perform TDM analysis upon to detect any occurrences of happy characters. Here, the AI is being trained for predictive and analytical purposes, thus transformative under Factor One (and also a nonprofit scholarly or educational use in this context). Provided that the scholar does not reproduce or distribute the training corpus to the public, nothing about the scholar's training of the AI or their subsequent TDM analysis re-expresses the copyrighted works or supplants the market for the sales of those works, securing fairness under Factor Four.³³

Were these fair use rights overridden by limiting AI training access to only "safe" materials (like public domain works or works for which training permission has been granted via license), this would exacerbate bias in the nature of research questions able to be studied and the methodologies available to study them, and amplify the views of an unrepresentative set of creators given the limited types of materials available with which to conduct the studies.³⁴ For instance, if a scholar were able to license for AI training only those books published by a certain publisher (with fair use otherwise circumscribed), the comprehensiveness and objectivity of the AI training would be degraded, and the scholar would be subsequently limited in the scope of inquiries that can be made using the AI tool.

III. NOI Question 9: Any opt-out rights granted should not extend to fair uses

The fair use provision of Section 107 does not afford copyright owners a right to opt out of allowing other people to use their works for good reason: if content creators were able to opt out, the provision for fair use would be undermined, and little content would be available to build upon for the advancement of science and the useful arts. Accordingly, to the extent that the Copyright Office is inclined to create a regulatory right for creators to opt out of having their works included in AI training, it is paramount that such opt-out provision not be extended to any AI training or activities that constitute fair use, particularly in the nonprofit educational and research contexts.³⁵

Research, scholarship, and teaching for nonprofit educational purposes are all favored purposes of fair

³³ Even supposing the scholar instead wanted to utilize the AI tool to generate new passages that express the tool's own version of "happy characters," the inputs or training processes for the AI tool do not re-express the copyrighted works for the public. The *outputs* of the AI tool may or may not be substantially similar on a case-by-case basis—but the reproduction of works to *train* the AI does not reproduce, distribute, or express anything to the public. Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper 4438593). <https://doi.org/10.2139/ssrn.4438593>

³⁴ See, e.g., Levendowski, A. (2018). How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. *Washington Law Review*, 93(2), 579. <https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/>; Authors Alliance. (2023, August 30). Copyright and Generative AI: Our Views Today. *Authors Alliance*. <https://authorsalliance.substack.com/p/copyright-and-generative-ai-our-views>

³⁵ This would also align with European Union law on this point. Article 3 of the Directive on Copyright in the Digital Single Market requires that Member States enact an exception to copyright for "reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access" and Article 7 ensures that "Any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6 shall be unenforceable." Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. Retrieved August 18, 2023, from <https://eur-lex.europa.eu/eli/dir/2019/790/oj>

use, and cited in the statute's preamble.³⁶ Yet, fair use in these contexts is becoming an out-of-reach luxury even for the wealthiest institutions.³⁷ This is because academic libraries are forced to pay significant sums each year to try to preserve fair use rights for campus scholars within the content license agreements that libraries sign. In the U.S., the prospect of "contractual override" means that, although fair use is statutorily prescribed—serving as a fundamental accommodation to the First Amendment³⁸—the prevailing theory is that private parties may contract to derogate the fair use exception.³⁹

And so, rightsholders, academic publishers, and content aggregators regularly "contract around" fair use by requiring libraries to negotiate for otherwise lawful activities (such as conducting TDM or training AI for research), and often to pay additional fees for the right to conduct these lawful activities on top of the cost of licensing the content, itself. When such costs are beyond institutional reach, the publisher or vendor may then offer similar contractual terms directly to research teams, who may feel obliged to agree in order to get access to the content they need.⁴⁰ Vendors may charge tens or even hundreds of thousands of dollars for this type of access.⁴¹

This "pay-to-play" landscape of charging institutions for the opportunity to rely on existing statutory rights is particularly detrimental for TDM research methodologies. As global scholars have recently explained:

Licensing is not an affordable or viable option for many critical TDM projects. TDM research often requires use of massive datasets with works from many publishers, including copyright owners that cannot be identified or are unwilling to grant licenses. Forcing researchers to use only licensed or public domain content (i.e., content in which there is no enforceable copyright) can restrict topics of study, hamper reproducibility and validation, bias results, and dissuade

³⁶ See also, e.g., *Warhol*, 143 S. Ct. at 1274 (the purposes of teaching, scholarship, and research are the 'sorts of copying that courts and Congress most commonly have found to be fair uses')

³⁷ Palmedo, M. (2019). *The Impact of Copyright Exceptions for Researchers on Scholarly Output* (SSRN Scholarly Paper 3090022). <https://doi.org/10.2139/ssrn.3090022>

³⁸ Contract terms that would restrict fair use rights may also inherently restrict the Constitutional right to free speech, given that fair use is "built-in First Amendment accommodation" that "eases the tension" between free speech and copyright law. *De Fontbrune v. Wofsy*, 39 F.4th 1214, 1223 (9th Cir. 2022). See also 5 Nimmer on Copyright § 19E.05 (2023) (addressing concerns for the sufficiency of fair use to protect free speech).

³⁹ In a white paper on copyright and contractual issues faced by libraries, the Association of Research Libraries states, "Arguably, contract terms that seek to limit exceptions under the Copyright Act are preempted under a conflict-preemption theory" but notes there is "surprisingly little case law on this point." Klosek, K. (2022). *Copyright and Contracts: Issues & Strategies*. Association of Research Libraries.

<https://www.arl.org/wp-content/uploads/2022/07/Copyright-and-Contracts-Paper.pdf>. By contrast, this is not the case in the European Union. The EU's Copyright in the Digital Single Market Directive provides that, "Any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6 shall be unenforceable," referencing articles that govern text mining and data mining, digital cross-border teaching, and preservation by cultural heritage organizations, respectively. <https://eur-lex.europa.eu/eli/dir/2019/790/oj>. See also Singapore's Copyright Bill, which similarly prohibits contractual override of copyright exceptions.

<https://sso.agc.gov.sg/Bills-Supp/17-2021/Published/20210706?DocDate=20210706&ProvIds=P15-#pr187->

⁴⁰ Klosek, K. (2022). *Copyright and Contracts: Issues & Strategies*. Association of Research Libraries. <https://www.arl.org/wp-content/uploads/2022/07/Copyright-and-Contracts-Paper.pdf>, p. 10.

⁴¹ *Id.*

researchers from undertaking projects.⁴²

If the Copyright Office were to enable rightsholders to opt-out of training AI for research and teaching fair uses, then academic institutions and scholars would face even greater hurdles in licensing content for research purposes.⁴³ It would be operationally difficult for academic publishers and content aggregators to amass and license the “leftover” body of copyrighted works that remain eligible for AI training. Costs associated with publishers’ efforts in compiling “AI-training-eligible” content would be passed along as additional fees charged to academic libraries, who are already financially constrained to preserve TDM and other fair uses for scholars. In addition, rightsholders might opt out of allowing their work to be used for AI training fair uses, and then turn around and charge AI usage fees to scholars (or libraries)—essentially licensing back fair uses for research. These scenarios would impede scholarship by or for research teams who lack grant or institutional funds to cover these additional expenses; penalize research in or about underfunded disciplines or geographical regions; and result in bias as to the topics and regions studied.

Copyright exceptions like Section 107 matter for research: they result in a higher production of new works of scholarship, and drive TDM mining as a means of extracting information and advancing knowledge.^{44, 45} Scholars need to be able to utilize existing knowledge resources to create new knowledge goods.⁴⁶ Indeed, the availability of openly accessible scholarship (without legal or technical barriers) during the COVID-19 pandemic directly benefited public health policy- and decision-making.⁴⁷ Congress and the Copyright Office clearly understand the importance of facilitating access and usage rights, having implemented Section 107 without any statutory or regulatory exclusions or opt-outs. This status quo should be preserved: to promote the progress of science and useful arts, copyright holders should not be permitted to opt out of having their works used for AI training when such AI training or practices would be a fair use—and particularly in the nonprofit educational or research contexts.

We would be pleased to provide any additional information that would assist the Copyright Office’s inquiry.

Sincerely,

⁴² Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., et al. (2022). Legal reform to enhance global text and data mining research. *Science*, 378(6623), 951–953. <https://doi.org/10.1126/science.add6124> (internal citations omitted)

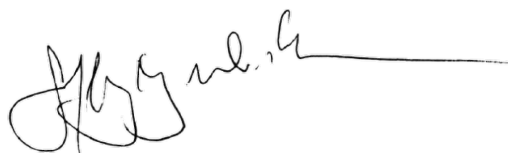
⁴³ We also believe that because an author’s derivative rights under copyright “do not include an exclusive right to supply information...about her works” (*Authors Guild v. Google*, 804 F.3d at 207), were the Copyright Office to permit opt-outs from AI training, the Copyright Office would be de facto creating a right under copyright law to protect non-protectable information rather than expression.

⁴⁴ Palmedo, M. (2019). *The Impact of Copyright Exceptions for Researchers on Scholarly Output* (SSRN Scholarly Paper 3090022). <https://doi.org/10.2139/ssrn.3090022>

⁴⁵ Handke, C., Guibault, L., & Vallbé, J.J. (2021). Copyright’s impact on data mining in academic research. *Managerial and Decision Economics*, 42(8), 1999–2016. <https://doi.org/10.1002/mde.3354>

⁴⁶ Palmedo, M. (2019). *The Impact of Copyright Exceptions for Researchers on Scholarly Output* (SSRN Scholarly Paper 3090022). <https://doi.org/10.2139/ssrn.3090022>

⁴⁷ Yin, Y., Gao, J., Jones, B. F., & Wang, D. (2021). Coevolution of policy and science during the pandemic. *Science*, 371(6525), 128–130. <https://doi.org/10.1126/science.abe3084>

A handwritten signature in black ink, appearing to read 'Jeffrey Mackie-Mason', followed by a long horizontal line.

Jeffrey Mackie-Mason
on behalf of the University Library
University of California, Berkeley

A handwritten signature in black ink, appearing to read 'Haipeng Li'.

Haipeng Li
on behalf of the Library
University of California, Merced