EXHIBIT A

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: All Cases	Case No. 25-md-3143 (SHS) (OTW)

[PROPOSED] AMICUS CURIAE BRIEF OF WIKIMEDIA FOUNDATION IN SUPPORT OF NEITHER PARTY

Rhett O. Millsaps II
LEX LUMINA LLP
745 Fifth Avenue, Suite 500
New York, NY 10151
(646) 898-2055
rhett@lex-lumina.com

Stephen LaPorte
General Counsel
Jacob Rogers
Associate General Counsel
Aly Marino
Legal Counsel
Wikimedia Foundation, Inc.
1 Sansome Street
Suite 1895
San Francisco, CA 94104
(415) 839-6885

Counsel for Amicus Curiae
Wikimedia Foundation, Inc.

TABLE OF CONTENTS

Page

TABLE OF AUTHORITIES……i

CORPORATE DISCLOSURE STATEMENT……..iii

STATEMENT OF INTEREST……1

INTRODUCTION AND SUMMARY OF ARGUMENT……..3

ARGUMENT……..6

I. *Warhol* Requires that the Fair-Use Analysis Be Applied on a Distinct, Use-by-Use Basis Throughout Each Stage of LLM Development…….. 6

II. Scraping—Making a Technical Request to a Website for a Copy of a Work through an Automated Process—Is Not a "Use" and Should Have No Bearing on the Fair Use Analyses Applied to Other Uses of Works in the Creation of a Product…….7

III. Using Copied Works at the Pre-Training Stage is Likely Fair Under the Four-Factor Test……..8

IV. Using Copied Works in a Chatbot's RAG System is Likely Not Fair Under the Four-Factor Test……11

CONCLUSION……..18

TABLE OF AUTHORITIES

Page(s)

Cases

Associated Press v. Meltwater U.S. Holdings, Inc.,
931 F. Supp. 2d 537 (S.D.N.Y. 2013)…….. 12, 16

Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith,
598 U.S. 508 (2023)…….. ……3, 6, 12

Authors Guild v. Google, Inc.,
804 F.3d 202 (2d Cir. 2015) ……..……9, 10, 12, 16

Authors Guild, Inc. v. HathiTrust,
755 F.3d 87 (2d Cir. 2014) ……..9

A.V. ex rel. Vanderhye v. iParadigms, LLC,
562 F.3d 630 (4th Cir. 2009) ……15

Barcroft Media, Ltd. v. Coed Media Group, LLC,
297 F. Supp. 3d 339 (S.D.N.Y. 2017) ……...16

Bartz v. Anthropic PBC,
787 F. Supp. 3d 1007 (N.D. Cal. 2025) …….……6, 7, 10

Cambridge Univ. Press v. Patton,
769 F.3d 1232 (11th Cir. 2014) ……14

Campbell v. Acuff-Rose Music, Inc.,
510 U.S. 569 (1994) …….……10, 13, 15

Fox News Network, LLC v. TVEyes, Inc.,
883 F.3d 169 (2d Cir. 2018) ……9

Hachette Book Grp., Inc. v. Internet Archive,
115 F.4th 163 (2d Cir. 2024) ……16

Harper & Row Publishers, Inc. v. Nation Enterp.,
471 U.S. 539 (1985) ……..13, 14, 15

Infinity Broad. Corp. v. Kirkwood,
150 F.3d 104 (2d Cir. 1998) ……9

Kadrey v. Meta Platforms, Inc.,
788 F. Supp. 3d 1026 (N.D. Cal. 2025) ……9, 10

McGucken v. Pub. Ocean Ltd.,
42 F.4th 1149 (9th Cir. 2022) ……15

Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.,
166 F.3d 65 (2d Cir. 1999) ……14

Penguin Random House LLC v. Colting,
270 F. Supp. 3d 736 (S.D.N.Y. 2017) ……...12

Perfect 10, Inc. v. Amazon.com, Inc.,
508 F.3d 1146 (9th Cir. 2007) ……...15

Philpot v. Indep. J. Rev.,
92 F.4th 252 (4th Cir. 2024) ……..14

Sega Enters. Ltd. v. Accolade, Inc.,
977 F.2d 1510 (9th Cir. 1992) ……7

Sony Corp. of Am. v. Universal City Studios, Inc.,
464 U.S. 417 (1984) ……10, 11

Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc.,
765 F. Supp. 3d 382 (D. Del. 2025) ……..12

Wall Data Inc. v. Los Angeles Cnty. Sheriff's Dep't,
447 F.3d 769 (9th Cir. 2006) ……13

Worldwide Church of God v. Philadelphia Church of God, Inc.,
227 F.3d 1110 (9th Cir. 2000) ……...13

Other Authorities

Wikimedia Foundation Mission, Wikimedia Foundation, https://wikimediafoundation.org/who-we-are/mission/ (last visited Oct. 28, 2026) ……..……..1

Wikimedia Foundation, https://wikimediafoundation.org/ (last visited Oct. 28, 2026) ……..……1

Wikimedia Projects, Wikimedia Foundation, https://wikimediafoundation.org/what-we-do/wikimedia-projects/ (last visited Oct. 3, 2026) ……..1

Wikipedia:About, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:About (last visited Oct. 28, 2026) ……..1

Monthly Overview, Wikimedia Statistics, https://stats.wikimedia.org/#/en.wikipedia.org/reading/total-page-views/normal|bar|2-year|~total|monthly (last visited Oct. 28, 2026) ……..……1

Wikipedia:Notability, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Notability (last visited Oct. 28, 2026) ……..1

Wikipedia:Editing policy, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Editing_policy (last visited Oct. 28, 2026) ……..……1

MediaWiki, https://www.mediawiki.org/wiki/MediaWiki (last visited Oct. 28, 2026)……..……1

Attribution-ShareAlike 4.0 International, Creative Commons, https://creativecommons.org/licenses/by-sa/4.0/deed.en (last visited Oct. 28, 2026)……2

Wikimedia Foundation Terms of Use, Wikimedia Foundation, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use#7._Licensing_of_Content (last visited Oct. 28, 2026) ……2

Copyleft, Wikipedia, https://en.wikipedia.org/wiki/Copyleft (last visited Oct. 28, 2026)……2

Birgit Mueller, Chris Danis & Giuseppe Lavagetto, *How crawlers impact the operations of the Wikimedia projects*, Diff (Apr. 1, 2025), https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/ ……..2

Wikimedia, *The cost of 'free': How Wikimedia Enterprise protects Wikipedia in the AI era*, Medium (July 16, 2026), https://medium.com/freely-sharing-the-sum-of-all-knowledge/the-cost-of-free-how-wikimedia-enterprise-protects-wikipedia-in-the-ai-era-da89552e8927.....2

Jacob Rogers, *How AI Threatens the Social Contract of Free Knowledge*, Diff (July 16, 2026), https://diff.wikimedia.org/2026/07/17/how-ai-threatens-the-social-contract-of-free-knowledge……3

Manual of Style, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last visited Oct. 28, 2026)……8

Wikipedia:Presumptive removal of AI-generated content, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Presumptive_removal_of_AI-generated_content (last visited Oct. 28, 2026) ……8

Machine Learning, MediaWiki, https://www.mediawiki.org/wiki/Machine_Learning (last visited Oct. 28, 2026) ……..8

What is Retrieval-Augmented Generation (RAG)?, Google Cloud, https://cloud.google.com/use-cases/retrieval-augmented-generation (last visited Oct. 28, 2026)……..……11

Wiki Loves Monuments, https://www.wikilovesmonuments.org/ (last visited Oct. 28, 2026).....17

CORPORATE DISCLOSURE STATEMENT

Wikimedia Foundation, Inc. is a non-profit corporation. It does not offer stock and has no parent company.

INTEREST OF AMICUS CURIAE

Wikimedia Foundation, Inc. ("Wikimedia") is a non-profit charitable foundation based in San Francisco, California.[1] Its mission is to empower and engage people around the world to build free, reliable knowledge and to disseminate it globally under a free license.[2] The Foundation accomplishes its mission primarily by hosting thirteen free-knowledge platforms known as the Wikimedia Projects.[3] These Wikimedia Projects host factual and educational content that is created, edited, and moderated by over 260,000 volunteer contributors per month worldwide.[4] Wikipedia is Wikimedia's most well-known platform, serving as a free online encyclopedia that allows human users from all around the world to collaboratively write and edit content.[5] Wikipedia is viewed billions of times every month.[6] Volunteer editors determine whether a topic is notable[7] enough to deserve its own page, ensure that content remains accurate and neutral, and confirm that they are cited by reliable sources.[8] Wikimedia also develops and improves free software, called MediaWiki, that is used to host its projects and by many third parties for hosting their own wikis around the world.[9]

---

[1] No party or party's counsel authored this proposed brief in whole or in part, and no person other than Wikimedia contributed any money to fund its preparation or submission.
[2] *Wikimedia Foundation Mission*, Wikimedia Foundation, https://wikimediafoundation.org/who-we-are/mission/ (last visited Sept. 23, 2026) and Wikimedia Foundation, https://wikimediafoundation.org/ (last visited Oct. 28, 2026).
[3] The Wikimedia Projects include Wikipedia, Wikibooks, Wiktionary, Wikiquote, Wikimedia Commons, Wikisource, Wikiversity, Wikispecies, Wikidata, Wikifunctions, MediaWiki, Meta-Wiki, Wikivoyage. *See Wikimedia Projects,* Wikimedia Foundation, https://wikimediafoundation.org/what-we-do/wikimedia-projects/ (last visited Oct. 28. 2026).
[4] As of January 2025.
[5] *Wikipedia: About*, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:About (last visited Oct. 28, 2026).
[6] *Monthly Overview*, Wikimedia Statistics, https://stats.wikimedia.org/#/en.wikipedia.org/reading/total-page-views/normal|bar|2-year|~total|monthly (last visited Oct. 28, 2026).
[7] *Wikipedia:Notability*, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Notability (last visited Oct. 28, 2026).
[8] *See Wikipedia:Editing policy*, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Editing_policy (last visited Oct. 28, 2026).
[9] MediaWiki, https://www.mediawiki.org/wiki/MediaWiki (last visited Oct. 28, 2026).

Wikimedia has a sustained interest in copyright law because copyright licensing fundamentally defines both the knowledge available in Wikimedia-hosted projects and the terms for its dissemination. Most written contributions to the Wikimedia Projects are licensed under a Creative Commons CC BY-SA 4.0 license,[10] which requires (i) that credit be given to the original creators of the work and (ii) that remixes, adaptations, or any other derivatives of the original works are made under the same free license terms.[11] These requirements form the basis of a "copyleft"[12] model of free content distribution, meaning that people can freely build on Wikimedia contributions following certain terms.

The Wikimedia Projects also rely on copyright law interpretations that incentivize web publishers to keep content openly accessible for both human readers and technological innovation. At the same time that human volunteers contribute to and learn from knowledge on Wikimedia projects, Wikimedia websites have become one of the most critical sources of data for all artificial intelligence and machine learning training. Sustained increases in non-human traffic from scraping and crawling impact Wikimedia's ability to keep Wikimedia sites online for human readers.[13] As partial solutions, Wikimedia has put in place a number of technical safeguards and directed high-volume users of Wikimedia project data to a service called Wikimedia Enterprise to keep the sites in service to human readers.[14] Without the backing of enforceable copyright, web publishers will be pushed to put up more walls against bad actors,

[10] *Attribution-ShareAlike 4.0 International*, Creative Commons, https://creativecommons.org/licenses/by-sa/4.0/deed.en (last visited Oct. 28, 2026).
[11] *Wikimedia Foundation Terms of Use*, Wikimedia Foundation, https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use#7._Licensing_of_Content (last visited Oct. 28, 2026).
[12] *Copyleft,* Wikipedia, https://en.wikipedia.org/wiki/Copyleft (last visited Oct. 28, 2026).
[13] Birgit Mueller, Chris Danis & Giuseppe Lavagetto, *How crawlers impact the operations of the Wikimedia projects*, Diff (Apr. 1, 2025), https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/ .
[14] Wikimedia, *The cost of 'free': How Wikimedia Enterprise protects Wikipedia in the AI era,* Medium, (July 16, 2026), https://medium.com/freely-sharing-the-sum-of-all-knowledge/the-cost-of-free-how-wikimedia-enterprise-protects-wikipedia-in-the-ai-era-da89552e8927.

leading to a more closed internet for people. Thus, how and when AI developers are required to honor CC BY-SA licensing terms impact Wikimedia's dual mission in encouraging people to create free knowledge and protecting the free dissemination of that knowledge.

Wikimedia is concerned that the parties' positions in this case push for overbroad and simplified applications of the fair-use doctrine to uses of copyrighted works throughout AI chatbots' development. Fair use should not be interpreted in a manner that would gatekeep all uses of copied works under proprietary terms to the detriment of efficient innovation. Neither should fair-use doctrine serve as a green light to snub licensing terms and fragment the flow of knowledge between human authors and audiences. In the interest of promoting open access to knowledge, Wikimedia urges the Court to adopt a sharper and more nuanced application of the fair-use doctrine than the parties in this case are seeking.

INTRODUCTION AND SUMMARY OF ARGUMENT

The best way to support Wikimedia's free-knowledge mission is to foster an ecosystem of human-driven knowledge production. Such an ecosystem requires a legal structure that meets the desires of the people contributing to free knowledge by ensuring their work receives credit and remains freely usable for additional knowledge production.[15] To maintain such a legal structure, fair-use analysis of AI tools must not be monolithic. Instead, the Court should treat the multiple allegedly infringing uses identified in this case with their own separate fair-use analyses, as mandated by the Supreme Court's decision in *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508 (2023) ("*Warhol*").

[15] Jacob Rogers, *How AI Threatens the Social Contract of Free Knowledge,* Diff (July 16, 2026), https://diff.wikimedia.org/2026/07/17/how-ai-threatens-the-social-contract-of-free-knowledge/

As a preliminary matter, the act of scraping should not be treated as a "use" of a copyrighted work. Scraping is simply a means of acquiring copies of works, but acquisition in and of itself is not a use of those works. Instead, what one actually *does with the acquired copies* is the "use" to which a fair-use analysis should be applied. In the context of large language models ("LLMs"), copies acquired via scraping are subsequently used by developers at several stages of building and deploying LLM-based services. As most relevant here, there are key distinctions between (1) a copy's use in training a general purpose LLM and (2) using copies to supplement an LLM-based service through retrieval augmented generation ("RAG").[16] RAG is useful to discuss because it is emblematic of a number of similar grounding techniques that make use of copies of sources at time of inference (*i.e.*, in response to individual user queries). Each distinctive use should be assessed with independent fair-use analyses.

First, the use of copyrighted works to train a general purpose LLM is likely to be fair under the four-step analysis. While fair use is always context-specific to each case, the experience that Wikimedia and its volunteer communities have with building and using AI/ML tools suggests that using copied works to build a general purpose LLM is likely transformative and does not inherently create market harm by substituting for the underlying works.

Second, after developers incorporate a general purpose LLM into a consumer-facing service (like a chatbot), the way copies are used by that service differs from the way copies are used in the process of training the underlying LLM. Per the use-by-use approach, while uses of works in some aspects of a deployed LLM-based service may be fair, others may not be.

[16] Other laws apply to how works may be accessed, and so even where data is freely available to the public (as with Wikipedia), scrapers should follow the rules of the road and avoid overburdening websites so that they remain online and accessible to their human users. As a simple comparison: a public park can remain free and open while preventing someone from driving a container truck through it.

It is likely not fair use when an LLM-based service uses copies of original works for their intended purposes—to provide relevant information, perspectives, expressions, and context—while supplanting readers' engagement with those original works. This is particularly the case with a consumer-facing chatbot, like ChatGPT, that relies on RAG techniques to respond to queries. Such a service copies text from third-party websites (like Wikipedia or *The New York Times*), selects and pastes the most relevant passages or "chunks" of that text as supplemental inputs into the LLM's context window alongside the user's query, and often substantially reproduces the original content in its generated output. This is problematic for Wikimedia and the free-knowledge movement when such copies are made and used without compliance with the required attribution and ShareAlike terms of the Creative Commons license.[17]

In summary, Wikimedia respectfully urges the Court to take a more distinct and nuanced approach toward the fair-use question than the parties advocate. This includes recognizing that the act of scraping itself is not a copyright "use," and instead subjecting only what is actually done with those scraped copies to fair-use analyses. The use of works in training an LLM that can be used for many different purposes is typically transformative without immediately causing market harm, making it likely to be a fair use. At the same time, a chatbot that implements RAG to copy and use works in ways that supplant the original works is not likely to be a fair use of those works. Both analyses apply here to ChatGPT: Wikimedia respectfully urges the Court to rule that while the pre-training of ChatGPT's general LLM was fair use, OpenAI's use of copies in ChatGPT's RAG feature is not fair and must abide by the terms of the relevant licenses, such as the Creative Commons license for Wikimedia Project content, to avoid infringement.

[17] *See* Strauss I., Yang J., O'Reilly T., Rosenblat S., Moure I., *The attribution crisis in LLM search results: Estimating ecosystem exploitation*, Data & Policy, Volume 8, e15-12 (2026),https://doi.org/10.1017/dap.2026.10064 ("We find a substantial "attribution gap" between relevant content consumed from websites and attribution practices that threatens the sustainability of the digital content ecosystem.").

ARGUMENT

I. *Warhol* Requires that the Fair-Use Analysis Be Applied on a Distinct, Use-by-Use Basis Throughout Each Stage of LLM Development.

At a high level, Wikimedia agrees with Plaintiff *The New York Times* that the deployment of an LLM can be described in four general stages: data acquisition (*i.e.*, via scraping), pre-training, fine-tuning or post-training, and a final product's output generation. When conducting a fair-use analysis across these stages, courts must identify every allegedly infringing use an LLM developer and/or deployer makes of copyrighted works and apply the four-step analysis separately to each such use. *Warhol*, 598 U.S. at 533. This is because "[t]he same copying may be fair when used for one purpose but not another." *Id.*

Within the four-step fair-use analysis, "the first factor in particular" requires a clear view of how an infringing copy is used at each stage. *Id*. *Bartz v. Anthropic PBC* is instructive in the proper application of *Warhol*. 787 F. Supp. 3d 1007 (N.D. Cal. 2025). First, that court held that the purpose of keeping unauthorized copies of books in an internal library "should they prove useful" is akin to the original works' purpose in "any university or corporate library" as being available for the "various contingent uses" of their expressive content. *Id.* at 1026. Second, the further, subsequent use of a subset of those same copies to train an LLM model had a transformative purpose. *Id.* at 1021-22. That court did not impute the fairness of one use, or the assessment of any of the fair-use factors, to another use of the same copies at a later stage in the process of creating an LLM. Instead, the court made separate determinations that "retain[ing pirated] copies [of books] should they prove useful for one thing or another[] was…not a transformative [use]" *and* that "using copyrighted works to train LLMs to generate new text was quintessentially transformative." *Id.* at 1022, 1026. Ultimately, under *Warhol*, courts must not "swallow" any "distinguishable infringing uses" by blending them together under a singular fair-

use analysis. *Id.* at 1020-28. Nor should courts import or export the 'purpose' of a copy's use between upstream or downstream stages of generative-AI product development.

Wikimedia respectfully asks this Court to align its fair-use approach with the distinctive uses of copies across the multiple separate stages of this process, as in *Bartz*, rather than treating different uses as part of a continuum.

II. Scraping—Making a Technical Request to a Website for a Copy of a Work through an Automated Process—Is Not a "Use" and Should Have No Bearing on the Fair Use Analyses Applied to Other Uses of Works in the Creation of a Product.

Scraping does not involve any copyright-law "use" to which the four-step fair-use analysis should apply. Scraping means making a copy of a work from a website, a necessary precondition to using a work on a computer no matter what the purpose, and thus comes before any type of creation or innovation.[18] What matters for fair-use analysis is what happens with those copies once made. Courts must "attend to different uses of different copies, and even to different uses of the *same copies*." *Id.* at 1028 (emphasis in original). Unauthorized acquisition of copies does not mean later uses of those infringing copies are not fair—indeed, such a rule would undermine the purpose of fair use. Likewise, fair uses of copies at one point do not transform other, unfair uses of those copies at another point into fair ones. Failing to recognize the "direct use of the copyrighted material" at each stage risks over- or under-applying fair use in a way that stymies both innovation and the distribution of information between human authors and audiences. *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992).

[18] This does not mean that scraping activity overall is benign. Some forms of scraping violate site terms of use and thus state contract or tort law. They may also violate statutes like the Digital Millennium Copyright Act (DMCA) and Computer Fraud and Abuse Act (CFAA).

III. Using Copied Works at the Pre-Training Stage Is Likely Fair Under the Four-Factor Test.

While this issue has been well-briefed by the parties, Wikimedia emphasizes that developing new tools and models can be beneficial to a wide range of applications very different from the original source material, including in manners beneficial to the free-knowledge movement.

Although the knowledge in the Wikimedia Projects is free, its creation can take a great deal of work, care, and creativity. Volunteers do careful research and debate the quality of sources, distilling their various viewpoints into the language and style[19] of Wikipedia. While Wikimedia has seen LLMs and other AI tools cause increased difficulty for volunteer contributors[20] when used poorly, Wikimedia also has witnessed the wide range of benefits that can flow from the use of these tools.[21] Tools from the field of machine learning that produced LLMs help volunteer editors fight vandalism, detect copyright violations and inappropriate personal information, determine what articles most need their help for improvement, and find new topic areas to work on. These uses made possible by LLM training help people work more effectively, and there likely are many more use cases that can help volunteers build better encyclopedia articles, find and check sources, and improve their writing. In this context, the ability to train and experiment with AI tools is likely to be a fair use based on the applicable factors.

FACTOR 1 likely weighs toward fair use for pre-training because the use of copies to create a general LLM is highly transformative. The purpose of works on sites like

[19] *Manual of Style*, Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style (last visited Oct. 28, 2026).
[20] *Wikipedia:Presumptive removal of AI-generated content,* Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:Presumptive_removal_of_AI-generated_content (last visited Oct. 28, 2026).
[21] *Machine Learning*, MediaWiki, https://www.mediawiki.org/wiki/Machine_Learning (last visited Oct. 28, 2026).

Wikipedia is to distill and communicate reliable information about the world from a neutral point of view. The purpose of copies of those works for LLM pre-training, however, is to provide linguistic patterns that undergird development of the model. The LLM's use of the works in pre-training is thus transformative because it is "different in purpose, character, expression, meaning, and message from [the works] from which it is drawn." *Authors Guild, Inc. v. HathiTrust,* 755 F.3d 87, 97 (2d Cir. 2014); *see also Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1044 (N.D. Cal. 2025) (purpose of copying books to develop a tool "that can be used to generate diverse text and perform a wide range of functions…is a use with a different purpose and character than the books themselves.").

FACTOR 2 is likely neutral in and is not dispositive to the balanced outcome. As an encyclopedic website, the bases of Wikipedia articles "undoubtedly [are] factual works," *Authors Guild v. Google, Inc.*, 804 F.3d 202, 220 (2d Cir. 2015) ("*Google Books*"). Although works of fiction tend to be "closer to the core of intended copyright protection" *Infinity Broad. Corp. v. Kirkwood*, 150 F.3d 104, 109 (2d Cir. 1998) (cleaned up), factual substance does not automatically exclude works from copyright protections and the rights attached to their licenses. *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 178 (2d Cir. 2018) (holding that while news is factual, that category itself does not exclude news reporting from copyright protections). In particular, Wikipedia editors invest substantial efforts to distill information from other sources, explain it in alignment with Wikipedia's stylistic and attribution standards, and often create new and freely licensed media to illustrate it. This process tends to culminate in collective works made by multiple volunteer authors subject to copyright protection and licensed under a Creative Commons license.

FACTOR 3 likely weighs toward fair use because the amount and substantiality of copied works used is not more than necessary for the transformative purposes of pre-training. Developers need "billions of words to train any given LLM" in order to establish strong linguistic pattern recognition. *Bartz*, 787 F. Supp. 3d at 1030. Using "the entirety of the works" is reasonably necessary because a higher quantity of processed words "does more to train [an LLM] than would feeding it only half" of those same works. *Kadrey*, 788 F. Supp. 3d at 1050.

Finally, FACTOR 4 likely weighs toward fair use because pre-training itself does not produce the substitutive market harm of concern in this case. To "affect the market for the original in a way cognizable under this factor," the use being analyzed must itself "act[] as a substitute for" the copyrighted work." *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 591 (1994). What matters is "significant substitutive competition," not competition generally. *Google Books*, 804 F.3d at 219. Pre-training inputs neither "serve as a substitute for the original," nor "shrink[] the protected market opportunities of the copyrighted work" in a way that would diminish the value of the transformative purpose of the use. *Id*. at 214. This is because copies of works in pre-training are used to analyze various works' syntax and structure to adjust the model's understanding of relationships between words. Such a use does not reproduce expressive content or meaning in a manner substantially similar to and thus substitutive of original works, which is the market harm alleged in this case.[22]

[22] Pre-training could be viewed as contributorily causing market harm because its creation of a general purpose LLM enables other people to make tools capable of substitutive outputs more easily. But the possibility of furthering copyright infringement does not mean the technology itself is a violation of copyright law. *See generally Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417 (1984). Copyright law should not prevent the "legitimate, unobjectionable purposes" of LLM pre-training. *Id*. at 442.

IV. Using Copied Works in a Chatbot's RAG System is Likely Not Fair Under the Four-Factor Test.

LLM-based chatbots like ChatGPT depend on RAG to provide timely and accurate responses to user queries that cannot be answered by the base knowledge in the underlying LLM. The RAG system can search the web (or another database that may be built from copies of protected works) to identify, copy, rank, and then paste relevant content into the LLM's context window alongside the user's query. This process augments a user's prompt with information and context that the LLM lacks.[23] When relevant content is copied from the internet, a chatbot with RAG relies on and reuses the most up-to-date and salient expressions from copies of websites like Wikipedia to inform its responses to a user's queries. Using copies of original web content as RAG inputs to augment system knowledge on a query topic therefore substitutes those original works' purpose in augmenting a human users' knowledge on a query topic. Further, RAG features sometimes respond to user queries by reproducing protected expressions without citing to the sources referenced.[24] This unattributed reproduction likewise substitutes the original works' intended purpose in providing expressive information sought by humans. Under the four-factor test, such uses are likely not transformative and also tend to create substantial harm to the market for the originals.

FACTOR 1 likely weighs against fair use because using copies for their original purpose—to inform readers—is not transformative. First, copies of works used in RAG "do

[23] *What is Retrieval-Augmented Generation (RAG)?*, Google Cloud, https://cloud.google.com/use-cases/retrieval-augmented-generation (last visited Oct. 28, 2026).

[24] There are also services using RAG, including ChatGPT, that do make attempts to consistently cite and link to original sources used. When they do, and thus abide by the CC BY-SA licenses of Wikipedia works, such uses are less likely to infringe public copyright licenses at all. Moreover, some chatbots with RAG draw exclusively from privately hosted databases of works the operator owns or has properly licensed. This is why a fair-use analysis must be applied on a use-by-use basis: depending on the facts, the use of copies for RAG outputs may or may not be fair. *C.f. Sony Corp.*, 464 U.S. at 450 ("The sale of copying equipment…does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes, or, indeed, is merely capable of substantial noninfringing uses.").

not have a further purpose or different character" from that of the originals. *Warhol*, 598 U.S. at 529; *see also Thomson Reuters Enterprise Centre GMBH v. Ross Intelligence Inc*., 765 F. Supp. 3d 382, 397-99 (D. Del. 2025) (copying an annotation system of one legal research tool to develop another legal research tool is not transformative). Chatbots with RAG pull information specifically targeted to a consumer's query on the requested topic from relevant sites. Content copied from sites like Wikipedia serve their original purpose in providing reliable context and information on a searched topic—but within a chatbot, that expression is intercepted by a RAG system rather than directly reaching the human searcher. Because this secondary use of a copy remains a "vehicle for conveying…the [protected works'] original stories and insights," the use is more likely to be non-transformative. *Penguin Random House LLC v. Colting,* 270 F. Supp. 3d 736, 750 (S.D.N.Y. 2017). Content from original sources used as inputs to a RAG system is often reproduced in the system's output as well, whether word for word or substantially similar. Thus, both RAG inputs containing copies of protected works and outputs containing an expression or expressions substantially similar to original work(s) without complying with required provisions "act[] as a substitute for [the original] sites" scraped without "driving [consumers back] to [those] third-party websites[.]" *Associated Press v. Meltwater U.S. Holdings, Inc.,* 931 F. Supp. 2d 537, 554 (S.D.N.Y. 2013). Such uses are not transformative.

RAG outputs, however, are not necessarily the same as summaries of copyrighted works. Many summaries use sufficiently original language and style that they do not violate the expression in the original work at all, creating no copyright infringement claim. Other summaries (to the extent they infringe copyright) can serve a different, transformative purpose when making available "significant information *about those* [*works*]" sufficient to permit a searcher to "identify those…of interest." *Google Books*, 804 F.3d at 217 (emphasis in original). Where

original works serve expressive, entertainment purposes, summarizations typically serve to "shed[] light on [those] work[s.]" *Campbell*, 510 U.S. at 580. Like the snippet functions in *Google Books*, summaries are unlikely to act as "significant substitute[s]" because their "incomplete nature" make it "rare" that the "searcher's interest in the protected aspect of the author's work would be satisfied by what is available from" a summary. 804 F.3d at 225. As with every fair-use analysis, however, generalizations on summarization should not be made. In some cases, uses purporting to be summarizations could be unfair because they contain more of the original expression than is necessary and therefore divert attention away from the originals. Our point is that summarizations should not be treated automatically as unfair either; courts should instead examine whether summarizations serve a different purpose than the underlying works or act as a substitute for those works.

Finally, while the commercial nature of a use does not control this factor, it may tip the scale when for-profit developers "profit from exploitation of the copyrighted material without paying the customary price." *Harper & Row Publishers, Inc. v. Nation Enterp.*, 471 U.S. 539, 562 (1985). This "customary price" can be financial compensation or logistical compliance with limitations of use in licensing terms. *See Worldwide Church of God v. Philadelphia Church of God, Inc*., 227 F.3d 1110, 1118 (9th Cir. 2000) (bypassing licensing terms creates a "benefit" for the user "without having to account to the copyright holder.") (cleaned up); *Wall Data Inc. v. Los Angeles Cnty. Sheriff's Dep't*, 447 F.3d 769, 779 (9th Cir. 2006) (making copies of a software limited to one per computer was commercial because doing so reaped the benefits of the software's use while avoiding the need to purchase authorized copies). Thus, chatbots using RAG that "st[and] to profit" in some way while also refusing to "pay the customary price of

direct attribution" and ShareAlike licensing terms required by Creative Commons licenses are also "exploitative." *Philpot v. Indep. J. Review*, 92 F.4th 252, 260 (4th Cir. 2024) (cleaned up).

FACTOR 2 is likely to be neutral. Again, while "predominantly factual works" are "less close to the core [of intended copyright protections] than more fictional pieces," such works are not categorically excluded from protection. *Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc.*, 166 F.3d 65, 72-73 (2d Cir. 1999) (second factor was neutral because factual news articles are original works but expressive elements were not a dominant feature). When, as with Wikipedia content, the nature of original works contain "evaluative" or "analytical… material that surpasses the bare facts necessary to communicate information[,]" this factor is "neutral, or even weigh[s] against fair use" when secondary uses are "dominated by such material." *Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1270 (11th Cir. 2014). Rather than acting as "bare factual compilations," Wikipedia pages in particular are the product of numerous volunteer authors making collaborative decisions about how to distill and arrange knowledge, communicate information in a Wikipedia-specific style, and create free illustrative content under a Creative Commons license. *Campbell*, 510 U.S. at 586.

FACTOR 3 likely weighs against fair use because the amount and substantiality of works used in a RAG system is not justified by a transformative purpose. This factor weighs against fair use when RAG copies "the heart" of the originals. *Harper & Row*, 471 U.S. at 564–65. This is what RAG does when it searches for content relevant to a user's query, breaks that content into "chunks," and then selects the most relevant chunks to include in the LLM's context window alongside the user's query. Developers decide how much of the retrieved content is copied, but the purpose remains the same: providing the LLM with the most salient expressions of information that are not part of the model so that it can generate a more accurate response. As

in *Harper & Row*, here RAG performs a "heart" transplant by repackaging the crux of original sources. *Id.* Such use does not provide a "transformative purpose that would justify reproducing" any of the works at all. *McGucken v. Pub. Ocean Ltd*., 42 F.4th 1149, 1162 (9th Cir. 2022).

When the "heart" of an original work is used as a RAG system's input, it may thereafter deliver outputs that are substantially similar to the original works that the service extracted from. Here again, RAG runs afoul of fair use because a "finding of fair use is [less] likely…when the copying…encompasses the most important parts of the original. " *Id.* at 221. Thus, this factor is likely to weigh against fair use whether RAG reproduces an original work solely as an input or in both the input and output phases.

FACTOR 4 likely weighs against fair use because market harm is caused when original content is exploited and reproduced in RAG without license compliance. Even when the original works are free, as they are in the Wikimedia projects, attribution and ShareAlike terms of the Creative Commons license encourage people to develop knowledge. This, in turn, allows AI companies to make available new contributions to that knowledge that would otherwise be locked up under proprietary terms. Chatbots with RAG are particularly problematic when structured to search and substantially reproduce from third-party sites (*e.g.*, Wikipedia or *The New York Times*) without acknowledging where that content was acquired. Unlike a variety of search-engine cases finding a fair use, a RAG system's use of protected works as inputs and outputs does not "transform [the use] into a *pointer directing a user to the source of information*[,]" *Perfect 10, Inc. v. Amazon.com, Inc*., 508 F.3d 1146, 1154 (9th Cir. 2007) (emphasis added); help detect characteristics about the work in a manner "completely unrelated to expressive content," *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 640 (4th Cir. 2009) (service archived student works to detect plagiarism); or make available

information about a work to help someone "evaluate whether [the work] falls within the scope of [their] interest." *Meltwater*, 931 F. Supp. 2d at 561. Instead, these chatbots make copies for and sometimes display those copies in RAG outputs "for the very purpose for which they were originally intended[:]" to provide information on a searched topic. *Barcroft Media, Ltd. V. Coed Media Group, LLC,* 297 F. Supp. 3d 339, 355 (S.D.N.Y. 2017). In this way, RAG systems' uses of copied works "'usurp[]' the function of the original works" and thereby harm the originals' market for human web-traffic. *Id.*

The "value of [Wikipedia's] work is further harmed by the fact that" chatbots with RAG "directly compete[] with [Wikipedia] for [viewership]." *Meltwater*, 931 F. Supp. 2d at 561. When AI companies fail to track and follow licensing requirements, they lock human viewership within the chatbot space; the prevalence of this activity has created "significant substitutive competition" for works like Wikipedia. *Google Books*, 804 F.3d at 219. Users diverted from visiting and thus benefiting from the information available on a Wikipedia entry are also diverted from the opportunity to further that knowledge via, for example, volunteer editing or donating. The parallel decrease in human traffic to the original web sources therefore reduces the resources and motivation necessary for content authors to continue their knowledge production efforts. Any "short-term public benefits" of chatbots's substitutive functions "are outweighed not only by harm to…authors but also by the long-term detriments society may suffer" from disincentivizing knowledge sharing. *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 196 (2d Cir. 2024).

Compliance with the provisions of Creative Commons or other copyright licenses would make RAG uses of works noninfringing. A ruling by the Court that requires such compliance would encourage developers to build license compliance into RAG features. This could, for

example, manifest as developers structuring their product design so that outputs using RAG techniques also embed the underlying content's source attribution. In such a world, chatbots using RAG to search and surface information from Wikipedia would more consistently follow attribution terms and/or provide links back to the original source material. This outcome benefits the free-knowledge movement because such LLM-based services better connect human authors to human audiences, thereby motivating continued information production and sharing, which in turn allows AI companies to make available new contributions to that knowledge that would otherwise be locked up under proprietary terms.

CONCLUSION

Wikimedia's experience has been that freedom to work on and with new technologies benefits the free-knowledge community. Trained LLMs, and indeed the broader sweep of trained machine-learning models, have demonstrated substantial benefits to the Wikimedia projects outside of text generation. These tools help address attempts at vandalism (even including detecting potential copyright infringement), help volunteers review new edits to determine the most valuable use of their time, and might be good at detecting undesirable posting of personal information and doxxing attempts, based on recent experimentation within the Wikimedia movement. It would be contrary to the purposes of copyright law to block these capabilities by labeling LLM development as categorically not fair use.

At the same time, fair use principles should not be used to greenlight the displacement of human-produced knowledge in the name of innovation alone. Volunteers create a range of new freely licensed media to illustrate articles, and contests like Wiki Loves Monuments[25] have spurred the creation of some of the most amazing pictures of the world's natural and

[25] Wiki Loves Monuments, https://www.wikilovesmonuments.org/ (last visited Oct. 28, 2026).

architectural wonders. These works are not merely given away, but freely licensed, using the structure of copyright protection to help maintain the freedom of the works over time as others reuse them. Without people driving the process of knowledge production, evolving technologies like LLM-enabled services are unable to make further contributions because they rely on humans to create new information first. But copyright holders and content producers should have ways to prevent and mitigate misuses of their works that might replace their originals, thus ensuring the longevity of the knowledge production ecosystem. Free-knowledge contributors require copyright law structures that encourage developers to play a role in referring readers back to original content sources, thereby facilitating the readers to contribute to, donate to, or otherwise support human-created content. Otherwise, the fragmentation between human authors and audiences will decrease the readily available information that undergirds LLM-based tools' usefulness.

To conclude, Wikimedia respectfully urges the Court not to adopt either parties' position in this case. Rather than blanket determinations that 'all uses by OpenAI are fair' or 'no uses by OpenAI are fair,' Wikimedia respectfully urges the Court to align with the use-by-use basis of fair-use doctrine and apply separate analyses to the multiple distinct copies made during the creation and use of LLMs at issue in this case. Specifically, Wikimedia respectfully encourages the Court to adopt the following positions: 1) what is done with scraped copies – and not the act of scraping – should be treated as the "use" that is analyzed; 2) the pre-training of ChatGPT's LLM models is fair use; 3) the copying and reproduction of expressive sources in responses to user queries by ChatGPT's RAG system is not fair use. Wikimedia submits that this interpretation of copyright law is in line with the guidance from the Supreme Court, and that it is

conducive to maintaining open technological development while deterring substitutive practices that reduce human knowledge creation and sharing.

Dated: October 29, 2026

Respectfully submitted,

/s/ *Rhett O. Millsaps II*
Rhett O. Millsaps II
LEX LUMINA LLP
745 Fifth Avenue, Suite 500
New York, NY 10151
(646) 898-2055
rhett@lex-lumina.com

Stephen LaPorte
General Counsel
Jacob Rogers
Associate General Counsel
Aly Marino
Legal Counsel
Wikimedia Foundation, Inc.
1 Sansome Street
Suite 1895
San Francisco, CA 94104
(415) 839-6885

Counsel for Amicus Curiae
Wikimedia Foundation, Inc.

CERTIFICATE OF COMPLIANCE

In accordance with Local Civil Rule 7.1(c) and the Court's August 7, 2026, Order, (ECF No. 1663), I hereby certify that the foregoing [Proposed Brief] of Amicus Curiae Wikimedia Foundation complies with the word count limit because it contains 5,831 words, excluding the parts exempted by Local Civil Rule 7.1(c). The basis of my knowledge is the word-count feature of the word-processing system used to prepare this proposed brief.

Dated: October 29, 2026

/s/ *Rhett O. Millsaps II*
Rhett O. Millsaps II