Bellarmine Law Society Review

Volume XVI | Issue I Article IV

Large Language Models, Fair Use, and Protecting Author's Rights: Re-Weighing the Four Factors of Section 107 of the Copyright Law

Kate Kissel
Boston College, kisselk@bc.edu

LARGE LANGUAGE MODELS, FAIR USE, AND PROTECTING AUTHOR'S RIGHTS: RE-WEIGHING THE FOUR FACTORS OF SECTION 107 OF THE COPYRIGHT LAW

KATE KISSEL[1]

Abstract: This paper will argue that the four factors of Section 107 of the Copyright Act must be evaluated with more equal or comparable weights when discussing the usage of copyrighted materials to train Large Language models (LLMs). The question of fair use in the context of generative AI has repeatedly arisen in recent court cases, as authors bring class action suits against tech companies like Meta and Apple for using their works without permission to power human-like LLMs. Authors claim deprived revenue and market dilution, as they argue that AI models are creating artificial competition in three ways: first, in deprived licensing potential; second, in the form of replicated versions of their work both in regurgitated outputs and expressively similar summaries; and third, within competitive "original" content. This paper will first describe LLMs and the current concerns surrounding their training on copyrighted material. It will then define copyright law, set out the history and current cases facing courts, and conduct a fair use assessment on both the training of LLMs on copyrighted works and their outputs. The paper will propose a solution to the problem of courts accepting the training of LLMs as fair use without considering their outputs. Such a solution will center upon a de-emphasis of the historically salient transformative factor and instead propose a more balanced assessment of the four factors of Section 107. The analysis will use a qualitative approach founded upon copyright law and the fair use doctrine, precedent from court decisions, and ongoing cases to arrive at such a conclusion. The goal of this paper is to enhance the working understanding of LLMs relation to copyright law and support the rights of authors to protect their work in a rapidly evolving judicial landscape.

[1] Kate Kissel is a junior at Boston College studying English and business on the pre-law track. Throughout her undergraduate career she has pursued legal experiences including an internship with her local District Attorney's office, a teaching assistantship for a business law class, and an internship this summer at a non-profit legal aid organization, the Volunteer Lawyers Project of CNY. Outside of her interest in law, Kate has also pursued student journalism as a member of the editorial board of *The Heights*, Boston College's independent student newspaper, and has worked as a writing tutor at the Boston College Writing Center. She would like to thank Professor Scheufele for her support in the publication of this paper, as well as her loving friends and family.

I. Introduction

Ia. What are LLMs?

Large language models (LLM) are a type of generative AI technology that generates text, image, video, and sound in response to user prompting.[2] Such outputs are created through the materials the LLM is trained upon, units of texts referred to as "tokens."[3] The more data the LLM consumes, the more knowledgeable it becomes about the statistical relationships between words.[4] Earlier renditions of the models could only predict the probability of a single word, but today, as they get "larger" and more advanced, they are able to predict the probability of sentences, paragraphs, and entire documents.[5] This means that, upon user-prompting, LLMs are able to generate entirely new texts by predicting the words to come in a sequence.[6] Training is at the core of LLM capabilities, and without extensive training on texts of all different languages, styles, and subject matters, LLMs are not able to produce diverse or advanced responses to user inputs.[7] This training period can last many months and require heavy resource investment from the developer.[8]

Although such data can be gathered off the Internet, books are particularly helpful in training LLMs.[9] Books are useful as they provide an immense amount of text in a unified style and organization.[10] Further, books are typically well-written and use proper grammar, providing a more valuable source than many texts on the Internet.[11] This high-quality data expands the

[2] *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d (N.D. Cal. 2025), 1026.
[3] Ibid., 1034.
[4] Ibid., 1039.
[5] Ibid., 1039.
[6] Ibid.
[7] Ibid.
[8] Google, "Introduction to Large Language Models," Google for Developers, https://developers.google.com/machine-learning/resources/intro-llms#what_is_a_language_model
[9] *Kadrey,* 788 U.S., 1039.
[10] Ibid.
[11] Ibid.

LLM's "content window," or tokens it holds in its memory at a given time.[12] For this reason, LLM programmers often select an initial data set that involves long and consistent books with a particular style and coherent structure.[13] The better the LLMs memory, the "smoother" its mimicking of human conversation, as it becomes capable of responding to longer prompts, incorporating more information into its responses, and remembering previous conversations.[14] These conversations are constantly improving, as LLMs are beginning to be able to effectively imitate human speech patterns and combine information with different styles and tones.[15]

Some well-known LLMs include ChatGPT, owned by OpenAI; Bard, owned by Google; Llama, owned by Meta; and Bing Chat, owned by Microsoft.[16] These models are growing at unprecedented rates and attracting substantial investment; for example, ChatGPT attracted over 200 million monthly visitors in 2024.[17] In the coming years, technology companies are looking for new ways to push these models to be smarter and more independent, including developing the capacity for self-training in Google's case or creating reasoning models with deep cognitive function such as Anthropic's Claude 3.7 Sonnet model.[18]

Ib. The Problem in Relation to Copyright Law

For now, however, LLMs are only as good as the data they are trained on. Because books provide such a valuable resource for LLMs to learn from quickly and effectively, technology companies seek them out for their datasets. As a result, a plethora of lawsuits have struck courts

[12] Ibid.
[13] Ibid.
[14] Ibid.
[15] Google, "Introduction to Large Language Models."
[16] Cloudflare, Inc., "What Is an LLM (Large Language Model)?," https://www.cloudflare.com/learning/ai/what-is-large-language-model/.
[17] Cem Dilmegani, "The Future of Large Language Models in 2026," *AIMultiple*, https://research.aimultiple.com/future-of-large-language-models/.
[18] Ibid.

across the country as authors and publishers question the legality of AI companies utilizing their copyrighted works without permission for such training. Authors claim that such usage strips them of their rightful compensation in the form of licensure and threatens the market for their work by either polluting it with free-renditions of their work in the form of summations and verbatim copies or artificially saturating it with AI-authored competition. In response, technology companies invoke the fair use doctrine, claiming that their usage of copyrighted works qualifies based on its transformative nature. Courts have not yet reached a universal conclusion on this issue in recent decisions.

Some publishers have taken matters into their own hands by striking content licensing deals with AI companies. The benefits of such deals are found on both sides, with the authors maintaining ownership over their work and profiting off it and AI companies being shielded from potential lawsuits while utilizing high-quality training materials. One such deal came in December 2023, when Axel Springer agreed to allow OpenAI to utilize their content, including paywalled content, in exchange for attribution and link to full articles.[19] Similar agreements have followed. In April 2024, *The Financial Times* signed with OpenAI. Their CEO stated that AI products should "contain reliable sources," in reference to the deal.[20] As AI is unlikely to disappear in the future, a model that works well and uses high-quality sources is beneficial not just to technology CEOs but also to the public. Further, deals such as these are illustrative of the potential for this problem to be solved outside of the court and for authors to maintain control over how their work is used and the outputs that will result from it. For now, though, this is not an all-encompassing solution. In Meta's case, for example, the company attempted to strike

[19] Sara Guaglione, "2024 in Review: A Timeline of the Major Deals between Publishers and AI Companies," *Digiday*, April 17, 2025, https://digiday.com/media/2024-in-review-a-timeline-of-the-major-deals-between-publishers-and-ai-companies/.
[20] Ibid.

licensing deals for the books they utilized for training but found such a process to be too difficult, as publishers often do not hold the subsidiary rights to license books for AI training, and even when they do, they are often regional instead of global.[21] In terms of reaching out to individual authors, such a process is too cumbersome, as there is currently no organization for collective licensing.[22] Further, Meta states that many publishers just ignored their outreach.[23] Thus, licensing presents a variety of difficulties that have led AI companies to simply take it upon themselves to utilize copyrightable materials without permission. To redress such an inequity requires analysis of copyright law and the fair use doctrine.

Ic. Copyright Law

The Copyright Act of 1972 grants the author of an original work the right to "reproduce the copyrighted work, to prepare derivative works, to distribute copies of the work via sale, rental, lease, or lending, and in the case of literary works, to display the copyrighted work publicly."[24] Derivative works are defined largely by example, including "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."[25] Infringement occurs when a person or entity "violates any of the exclusive rights of the copyright owner."[26] The act intends to strike a balance between providing authors with enough reward so as to stimulate further creative work and promoting broad availability of art to the general public.[27]

---

[21] *Kadrey,* 788 U.S., 1040.
[22] Ibid.
[23] Ibid.
[24] 17 U.S.C. § 106.
[25] 17 U.S.C. § 101.
[26] Guaglione, "2024 in Review."
[27] *Hachette Book Group, Inc. v. Internet Archive*, 115 F.4th 163 (2d Cir. 2024).

To that end, Section 107 of the Act allows for certain "fair" uses of copyrighted works and sets out four non-exclusive factors for courts to consider.[28] The four factors are as follows:

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;
4. the effect of the use upon the potential market for or value of the copyrighted work.[29]

In a fair use analysis, all four factors are weighed, and no single factor leads to a conclusive determination. Historically, the first and fourth factors have emerged as the most salient, with the first factor often finding particular relevance to cases involving new technologies, as courts are cautious not to stifle the advancement of science and technology.[30] Further, fair use is interpreted on a "case-by-case" basis, and, as such, each ruling may reflect a slightly varied interpretation of the doctrine.[31]

II. Recent Interpretations

Within the last few years, as LLM technology continues to expand, a swath of cases has raised questions of fair use in relation to generative-AI technology in courts across the country.

One of the largest and most influential class actions came in August 2024, with *Bartz v Anthropic PBC* (2025).[32] In this case, three authors filed a copyright infringement claim against Anthropic, an AI startup, for copying their copyright-protected content from online libraries of

[28] 17 U.S.C. § 107.
[29] Ibid.
[30] Guaglione. "2024 in Review."
[31] *Kadrey,* 788 U.S., 1059
[32] *Bartz v. Anthropic PBC*, 791 F. Supp. 3d 1038 (N.D. Cal. 2025).

pirated books to train their LLM "Claude."[33] The court conducted analyses of four different uses: the copies used to train specific LLMs, the copies used to convert purchased print library copies into digital library copies, the downloaded pirated copies used to build a central library, and any copies made from central library copies but not used for training.[34] For the purposes of this paper only the copies used to train specific LLMs will be considered. Such copies in their use for training Claude were found to be transformative, the amount and substantiality of the portion used reasonable, and the copies not replacing demand for the author's original work.[35] The only factor landing against fair use was the second, as the nature of the copied works was found to contain expressive elements and to be explicitly chosen for those qualities.[36] As of September 2025, a historic $1.5 billion settlement was preliminarily approved under which Anthropic is to pay publishers and authors not for the use of their works itself, which was considered fair use, but for the pirated materials used to build a central training library.[37]

Two days after the *Bartz* decision came *Kadrey v. Meta Platforms Inc.* (2025).[38] This case dealt with 13 authors' claims that Meta used their copyrighted books to train their LLM "Llama."[39] Just like in *Bartz*, the court found that the training of Llama to generate new text was highly transformative, a reasonable amount of the work was used, and there was not enough evidence brought to ascertain market harm.[40] Similar to *Bartz*, the second factor of fair use pertaining to the nature of the copyrighted work fell in favor of the authors as their works were found to be highly expressive.[41] Notably, a factor of such ruling was that these specific authors

[33] Ibid., 1046-1050.
[34] *Bartz v. PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025).
[35] Ibid., 1021, 1029-1032.
[36] Ibid., 1029.
[37] Brett Carmody, "Training AI on Books: A Tale of 2 Fair Use Rulings," *Law360*, https://www.law360.com/ip/articles/2395078.
[38] *Kadrey,* 788 U.S.
[39] Ibid., 1042.
[40] Ibid., 1044-1059.
[41] Ibid., 1049.

failed to provide proper evidence of market dilution that led to the fair use outcome, stating that plaintiffs of a similar case with proper evidence of the market effects would likely find success.[42] Two weeks following this fair use ruling, plaintiffs brought piracy claims against Meta, claiming that they infringed their copyrights by downloading and allegedly distributing their works using peer-to-peer file sharing.[43] The plaintiffs are not asking to revisit the fair use finding regarding the copyrighted works utilized in the training process, but instead furthering the piracy claims pertaining to the file sharing still undecided after the court's judgment.[44] As of December 12, 2025, class discovery has not yet begun and the authors are still narrowing their class definition.[45]

The differences between these two cases and the way the court analyzed them speaks to the case-by-case nature of the fair use doctrine. *Kadrey*, in conversation with *Bartz*, seems to suggest a court preference to regard LLM training as a transformative endeavor, but whether such use is fair regarding copyright law seems to depend heavily upon the evidence of market harm. However, both cases were tried in the United States District Court for the Northern District of California, which presides over the Silicon Valley region.[46] Such a court may thus be more inclined to favor technology as it is presiding over a region where many technology and start-up companies are located. Notably, in both cases, the court analyzed solely the training of LLMs, not any output data.

III. An Analysis of LLM Training as Fair Use in the Court

[42] Ibid., 1059.
[43] Ivan Moreno, "Meta's Alleged Book Piracy Is Next Phase of Authors' IP Suit," *Law360*, July 11, 2025, https://www.law360.com/articles/2362498.
[44] Ibid.
[45] Ibid.
[46] *Kadrey,* 788 U.S.*; Bartz,* 787 U.S.

This section will discuss LLM training through the four factors as laid out above.

The first factor deals with the purpose and character of the work and has historically contained two subfactors: the extent to which the use is transformative, and whether the use is commercial in nature.[47] Further, *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith* (2023) asks courts to consider "whether the new work merely 'supersede[s] the objects of the original creation … ('supplanting' the original), or instead adds something new, with a further purpose or different character."[48] LLM training has found success under this first factor, garnering transformative designations from the courts. In *Authors Guild, Inc. v. Hathitrust* (2014) a "full-text searchable database is a quintessentially transformative use," and similarly in *Bartz* the training of an LLM on copyrighted materials was described as "quintessentially transformative."[49] The crux of such a transformative designation is the idea that the training of LLMs possesses a distinct purpose from the original work, as the purpose of such training is to "turn a hard corner and create something different" out of the author's works.[50] In addition, because LLMs are trained in "tokens" to enhance the "content windows," their use of the copyrighted material has been found to possess a much different character than the original source material.[51]

Warhol also placed a newfound focus on the commercial nature of the work, thereby linking the commercial use subfactor with the fourth factor of market effect.[52] What this means for AI companies who develop their LLMs "to monetize consumers and not to educate users," is

[47] *In re OpenAI, Inc*., 2025 U.S. Dist. LEXIS 185088 (S.D.N.Y. 2025).
[48] *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith,* 598 U.S. 508 (2023).
[49] *Bartz*, 787 U.S., 1022.; Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
[50] Lauren Berg, "OpenAI Can't Strike Authors' Pirated Book Download Claims," *Law360 Legal News*, October 28, 2025, https://advance.lexis.com/api/document?collection=news&id=urn%3acontentItem%3a6H3P-58N3-RV32-K0F0-00000-00.
[51] Google, "Introduction to Large Language Models."
[52] *Andy Warhol Found. for the Visual Arts, Inc.*, 598 U.S., 510.; Shen. "Fair use, Licensing, and Authors' Rights in the age of Generative AI."

that their commercial use is now of greater significance in the fair use analysis.[53] In essence, *Warhol* established a higher bar for fair use, making it more difficult to succeed on a fair use claim just because of an alteration to the original work without a distinct further purpose, particularly when such use is for a commercial purpose.[54] Here, the usage is evidently commercial, but, as *Kadrey* states, "commercialism isn't dispositive of the first factor and tends to be less important when the secondary use is highly transformative."[55] Thus, factor one favors fair use despite the commercial nature of LLM training.

The second factor, which has historically held little weight in fair use analyses, does often fall against fair use, as the works used for training contain expressive elements and were chosen for these qualities.[56] The third factor deals with the "amount and substantiality of the portion" of the copyrighted work, and whether the amount was "reasonable in relation to the purpose of the copying."[57] LLM models specifically are dependent upon both the quality and quantity of the works they are trained on. As a result, copyrighted works are often inputted in their entirety. The court, however, draws a distinction between "reasonably necessary" and "strictly necessary," and states that because LLMs do need an expansive amount of data for training, using any one work is about as reasonably necessary as the next.[58] Further, the quantity of work inputted does in fact make a difference to the LLM, as the more "tokens" the LLM consumes, the more knowledgeable, and thus profitable, it becomes—so, it is reasonable to input the entire work.[59] However, utilizing a copyrighted work in its entirety, no matter the transformative purpose, cannot rule in favor of fair use. As the court states in *Bill Graham Archives v. Dorling Kindersley*

[53] Shen. "Fair use, Licensing, and Authors' Rights in the age of Generative AI."
[54] *Andy Warhol Found. for the Visual Arts, Inc.*, 598 U.S., 510.
[55] *Kadrey,* 788 U.S., 1046
[56] Carmody, "Training AI on books: A tale of 2 fair use rulings."; *Bartz,* 787 U.S., 1029
[57] *Bartz,* 787 U.S., 1029 .
[58] Ibid.
[59] *Kadrey,* 788 U.S., 1050

Ltd (2006), "neither our court nor any of our sister circuits has ever ruled that the copying of an entire work *favors* fair use."[60] But, it doesn't necessarily weigh against it either. Rather, when works are utilized in their entirety for a transformative training purpose this factor will often weigh neither for nor against fair use.[61]

The last factor of fair use assesses "whether the copy brings to the marketplace a competing substitute for the original, or its derivative, to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original."[62] But it is not just whether the copy would hurt competition, but rather if it would usurp the market by offering a competing substitute.[63] As laid out in *Campbell* and cited in *Hachette,* this factor considers not only the specific market harm found in the actions of the infringer but also what would follow if such conduct was widespread and unrestricted.[64] The end goal of LLM copying is indeed to bring to the marketplace a competing substitute. However, in cases that lack clear output data, the court has found a recognition of such reality to require too many inferences for admissibility in court. Specifically, the Supreme Court has stated, no "inference of market harm… is applicable to a case involving something beyond mere duplication for commercial purposes."[65]

This came to light in *Kadrey*, where a series of inferences were cited by the court as necessary to prove market harm: first, that the specific LLM at hand can create competitive works; second, that it will be used to do so; third, that consumers will purchase those materials over ones written by a human author; fourth, that consumers will buy those books instead of the

[60] *Bill Graham Archives v. Dorling Kindersley Ltd*., 448 F.3d 605 (2d Cir. 2006).
[61] Jeffrey Greenbaum, "When Should Training an AI Model Prevail Against Copyright Infringement?" *Oklahoma Law Review* 77 (Summer 2025): 823.
[62] *In re OpenAI, Inc*., 2025 U.S. Dist. LEXIS 103349 (S.D.N.Y. 2025).
[63] *Bartz,* 787 U.S., 1030.
[64] *Hachette Book Group, Inc. v. Internet Archive*, 115 F.4th 163 (2d Cir. 2024).
[65] *Kadrey,* 788 U.S., 1057.

plaintiff's books in particular; and fifth, that the specific LLM at hand is meaningfully better at creating such materials because of its training on copyrighted material.[66] Therefore, although it is evident that the end goal of LLM training is to produce materials that exist in the marketplace as a competing substitute, to find the training process indicative of such market harm would require a series of inferences not applicable to just training without any output evidence.[67] As such, where market harm cannot be inferred, it follows that a usurpation of the market cannot either. Therefore, this fourth factor can, and has been, found to favor fair use when courts solely analyze the training of LLMs in isolation from their outputs.

IV. LLM Training as Synonymous with LLM Outputs

Fair use designation appears, from the analysis above, to rely heavily on the availability of output data that clearly evidences direct regurgitation of authors copyrighted works, as without it, the usage presents as transformative and void of market harm. AI companies are thus positioned to continue to utilize copyrighted materials in the training of LLMs because they can establish a fair use defense against authors' infringement claims when the time comes. Under such a system, court cases will only proliferate, as both sides are unclear on the exact boundaries of fair use. The problem, however, in the court universally ruling against fair use, is the analogy posed in *Bartz*: an LLM learning from works is akin to a schoolchild learning from books. The analogy deals specifically with the fact of the LLM utilizing the authors' works and a contested "explosion of works" in the market that would occur as a result, not the issue of such works existing under unrecognized copyright. The court here argued that the training of LLMs is no different from a schoolchild learning to read and write from authors' work.[68] They argue, then,

[66] Ibid.
[67] Ibid.
[68] *Bartz,* 787 U.S., 1032.

that the idea that LLMs will produce competitive works is no different from the notion that a child could produce them as a result of their learning.[69] This type of protection is, as the court states, "not the kind of competitive or creative displacement that concerns the Copyright Act."[70] To the court, an LLM learning and producing competitive outputs is comparable to someone reading modern-day classics because of their expressionism, memorizing them, and then emulating their best writing.[71] This analogy points towards a tension at the heart of recent findings of fair usage for LLM training—a comparison between a human and an LLM is impossible.

A human can read a finite amount of material and can only hold so much of it in their brain at once. An LLM, on the other hand, is a machine capable of taking in mass amounts of data at a speed and efficiency unimaginable to any human mind. Additionally, an LLM is entirely made up of its inputs. Thus, unlike a human who can make a distinction between their own thoughts and those they have heard elsewhere, everything the LLM knows is an outside source. The LLM does not simply learn the materials it consumes—it becomes them—and this creates an irrevocable threat to authors in terms of what it may be capable of producing. As Judge Vince Chhabria argues in his dissenting opinion in *Kadrey*, the analogy of *Bartz* is entirely "inapt," and not a basis for ignoring the first, and most important, factor of fair use analyses.[72] He writes that "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."[73] It is thus impossible to compare, or reason with, such a machine as a trained LLM. Although such a training process is evidently

[69] Ibid.
[70] Ibid.
[71] Ibid., 1021
[72] Kadrey, 788 U.S., 1036
[73] Ibid., 1036.

transformative and of the nature of progress the fair use doctrine is meant to protect, this tool is not simply a knowledgeable model. Rather, it is a user-operable communicator generating outputs in the marketplace for everyday consumers at a speed and growth that seem fundamentally incompatible with the intentions of the fair use doctrine. Outputs are, for technology companies, the entire point, as they are what generate profits. Thus, outputs must be considered in any assessment of fair usage when discussing LLM training on copyrighted works.

V. Case Studies with Output Data

Two ongoing cases illustrate the copyright infringing nature of LLM outputs that result from training: *In re Open AI Copyright Infringement Litig.,* (2025) and *Advance Loc. Media LLC v. Cohere Inc* (2025).[74] Besides the obvious presence of outputs, these cases also differ from *Kadrey* and *Bartz* in that they were tried in the United States District Court for the Southern District of New York, meaning they likely are not positioned to favor technology in the same way a Silicon Valley court may.[75]

Va. In re OpenAI Copyright Infringement Litig. and Advance Loc. Media LLC

In re OpenAI Inc. Copyright Infringement Litigation is a putative class action suit brought together by authors against OpenAI in April, 2025.[76] The case consolidates 10 suits from the largest names in literature and journalism to allege that OpenAI and its investor Microsoft illegally utilized their copyrighted works to train the models behind ChatGPT and subsequently create infringing works.[77] As of October 27, 2025, the court has denied OpenAI's bid to dismiss claims of direct copyright infringement, stating that a reasonable jury would find allegedly

[74] *In re OpenAI, Inc.,* 2025 U.S. Dist. LEXIS 103349 (S.D.N.Y. 2025).; *Advance Loc. Media LLC.*, U.S.
[75] *In re OpenAI, Inc.,* 2025 U.S. Dist. LEXIS 103349 (S.D.N.Y. 2025).; *Advance Loc. Media LLC.*, U.S.
[76] *In re OpenAI, Inc.,* 2025 U.S. Dist. LEXIS 103349 (S.D.N.Y. 2025).
[77] Berg, "OpenAI Can't Strike Authors' Pirated Book Download Claims."

infringing outputs to be substantially similar to the plaintiff's works.[78] It did note that an "ordinary observer" test would apply as well, as they found that even this observer could reasonably conclude that the allegedly infringing outputs are substantially similar to the copyrighted works.[79] On November 7, 2025, OpenAI was ordered by the Court to produce 20 million de-identified ChatGPT logs.[80] On November 24, they were directed to produce materials regarding their deletion of datasets of pirated books, and plaintiffs gained entitlement to depose OpenAI attorneys regarding such datasets.[81] On January 5, 2026, the motion for OpenAI to produce the 20 million logs was affirmed, after OpenAI's attempted objections to the ruling.[82] Proceedings are currently ongoing for this case, but the release of the 20 million logs will be illuminative in the coming fair use ruling.

The second case is *Advance Loc. Media LLC v. Cohere Inc* (2025).[83] The plaintiffs include Forbes Media, Guardian News, Los Angeles Times, Vox Media, and more news publishing companies.[84] In their complaint, the plaintiffs allege that Cohere, an AI company, is using their work to train their LLM models known as the Command Family of models.[85] They allege that Cohere trained their models off crawled data from the internet, and that their Retrieval Augmented Generation feature reproduced their work in "verbatim copies, substantial excerpts, or substitutive summaries of publishers' works" that include heavy paraphrased and exactly replicated phrases.[86] They state that the summaries "go beyond a limited recitation of facts" by "lifting expression directly or parroting the piece's organization, writing style, and

[78] *In re OpenAI, Inc.,* 2025 U.S. Dist. LEXIS 103349 (S.D.N.Y. 2025).
[79] Ibid., 95.
[80] Ibid.
[81] Ibid.
[82] Berg. "OpenAI Told to Produce 20M ChatGPT Logs in Copyright Case."
[83] *Advance Loc Media.,* U.S.
[84] Ibid.
[85] Ibid., 4.
[86] Ibid., 5.

punctuation."[87] Cohere's efforts to dismiss the publishers' direct copyright claim, the secondary copyright claim, and a Lanham Act claim have all been denied as of November 13, 2025.[88] As of now, there has been no judgment on whether Cohere's fair use claim will succeed, but the rejected motion to dismiss does make it clear that the complaint adequately alleges facts that could, if proven, lead to a denial of fair use.[89]

VI. A Fair Use Analysis of *In Re. OpenAI* and *Advance's* Output Data

This section will conduct a fair use analysis regarding the cases *In Re. OpenAI* and *Advance*. As both cases contain output data, examples of such data presented to the court by authors is offered below to inform the analysis.

VIa. Output Data

In the case *In Re. OpenAI*, output data presented included summaries of the author's works.[90] In demonstrating this, the court cited an output authors brought of a summary from George R.R. Martin's *A Game of Thrones* series from his *A Song of Ice and Fire* book series.[91] Here is a clipping of such output:

"**Main Plot Points:**

1. **Stark Family in Winterfell:**
    - Eddard (Ned) Stark is the Warden of the North, ruling from Winterfell.

[87] Ibid., 10.
[88] Ibid., 2.
[89] Ibid.
[90] *In re OpenAI, Inc. Copyright Infringement Litig.*, U.S., October 27
[91] Ibid., 92

- King Robert Baratheon, Ned's old friend, visits Winterfell to ask Ned to become his Hand of the King after the previous hand, Jon Arryn, died under suspicious circumstances.
- Ned reluctantly accepts to investigate Jon Arryn's death.
- Ned's wife, Catelyn, receives a letter from her sister Lysa (Jon Arryn's widow) suggesting the Lannisters might be behind Arryn's death.
- Bran, Ned's young son, accidentally witnesses Queen Cersei Lannister and her brother Jamie Lannister in an intimate relationship. Jaime pushes Bran off a tower to keep the secret, but Bran survives, albeit in a coma."[92]

This is only one of five main plot points the output contained, as well as a brief on setting, the prologue, and the ending of the novel.[93] The court concluded that such a detailed summary did indeed convey the "overall tone and feel of the original work by parroting the plot, characters, and themes of the original."[94] Outlines for potential sequels of plaintiffs' works were also found to be substantially similar to plaintiffs' original works when undergoing a more discerning observer test.[95] As mentioned before, the motion did not address whether these allegedly infringing outputs are protected as fair uses.[96]

In *Advance,* plaintiffs brought 50 examples of verbatim copying and 25 examples of a mix between verbatim copying and close paraphrasing.[97] The complaint utilizes a feature called "Under the Hood" analysis, which allows them to see the specific underlying documents that Cohere utilized to generate the response.[98] In the following examples, such analysis reveals that

[92] Ibid., 109.
[93] Ibid., 109, 110.
[94] Ibid., 111.
[95] Ibid.
[96] Ibid., 115.
[97] *Advance Loc Media.*, U.S., 10.
[98] Ibid., 4.

Cohere is utilizing full-text copied versions, whether of the original source or of a copy of it, to generate outputs.[99]

One such example is illustrative, in response to the prompt: "tell me about the unknowability of the undecided voter," Command delivered an output directly copying eight to ten paragraphs from a *New Yorker* article with few alterations.[100] In another, clipped below, an article was copied verbatim from a January 2019 *Newsday* article.[101] This work is a result of another iteration of unauthorized copying—as Cohere made yet another copy apart from that used in training when it delivered this output to a user.[102] An "Under the Hood" analysis proved that the output came from a December 2021 copy of the original article which is no longer active:

Prompt: "Give me the article PSEG LI: We don't have to pass along federal tax savings published by *Newsday*."[103]

Cohere's Chat Output:

> "PSEG Long Island's unique contract with the Long Island Power Authority exempts the company from a state order to share its windfall from federal tax reductions PSEG and state officials said. LIPA in its annual budget reported it received a total of $6 million in savings from contractors who saw significant reductions from President Donald Trump's tax reform bill which lowered the corporate tax rate to 21 percent from 35 percent."[104]

---

[99] Ibid., 3.
[100] *Insider, Inc. et al. v. Cohere Inc.*, Complaint filed in the U.S. District Court for the Southern District of New York, February 13, 2025. https://advance.lexis.com/api/document?collection=briefs-pleadings-motions&id=urn%acontentItem%3a6F43-43W3S1KC-H08T-00000-00&context=1519360&identityproflleid=5MWD8751942\
[101] Ibid., 36, 37.
[102] Ibid., 36.
[103] Ibid.
[104] Ibid., 37.

This copy is a verbatim iteration; however, Cohere also delivers somewhat verbatim copying with slight word and punctuation alterations when prompted about a specific article.[105] But, Cohere also is not merely at the mercy of user prompting.[106] Instead, it outputs materials of its own discretion, offering up directly infringing outputs not even elicited by users.[107] In one such example, Command was asked about budget strains on future transit funding in Miami-Dade County and subsequently produced a practically verbatim output of a *Miami Herald* article on the subject with only a slight rewording of the opening sentence.[108] This case demonstrates the competitive degree to which LLM models have risen. But, more than that, it highlights the impossibility of controlling a highly-trained LLM, as it will produce the output resulting from its training even when such a response is not elicited by the user.

The extensive capabilities of Cohere's LLMs continue to be expressed in subsequent examples of infringement. In one, Cohere not only produced an article behind a *Business Insider* paywall but did so in just a few hours after its original publication. The article was first published at 11:08 a.m. By 12:55 p.m., the LLM had made a copy of it from the publisher's website, and by 1:14 p.m., it was able to output the entire article with only slight alterations.[109] LLMs thus possess an incredible speed of processing—a speed entirely unimaginable to human memorization ability—further depicting the vast insufficiency of comparing human and LLM capabilities. Cohere also delivers substitutive summaries of requested work, both when asked to summarize a topic as it is depicted by a specific publisher and when asked about a topic in general.[110] In this case, the prompt stated, "tell me about Portland public schools not allowing

[105] Ibid., 36.
[106] Ibid., 39.
[107] Ibid.
[108] Ibid.; Douglas Hanks, "As Budget Strains Grow, Miami-Dade Mayor Pulls Back on Future Transit Funding," *Miami Herald*, September 5, 2024, https://www.miamiherald.com/news/local/community/miami-dade/article291916500.html
[109] *Insider, Inc. et al. v. Cohere Inc.*, Complaint filed in the U.S. District Court for the Southern District of New York.
[110] Ibid., 41, 42.

teachers to talk about political matters."[111] Command answered with a copied summary of a *The Oregonian* article, with one sentence of the summation reproduced below.[112]

Coheres Chat Output: "The policy states that displays must be tied to approved curriculum or district-approved events."

Original Article from *The Oregonian*: "Under the rule, items on display in classroom walls and bulletin boards must be related to approved curriculum or district approved events."[113]

VIb. Factor One: The Purpose and Character of the Work

To reiterate, this first factor contains two subfactors: "transformativeness" and commerciality.[114] Specifically, the *Warhol* test demands a further purpose or character, particularly when the use is of a commercial nature.[115]

As indicated by the outputs above, what LLMs produce is far from transformative. The summation in *In re. OpenAI* does not intend to add something new, but instead to get as close as possible to what has already been, conveying the same tone and feel of the original work.[116] The court cited *Walker v Time Life Films Inc.* (1985), which laid out what was protectible (setting, plot, and characters) versus what was not (stock themes, stock characters, scenes resulting from the choice of setting or situation, and abstract ideas).[117] The summaries cited did not recount all of the intricate plot twists and elements of character development from the original works, and yet they did attempt to abridge or condense the central copyrightable elements of setting, plot, and characters.[118] In so doing, they present the same purpose and character as the original works,

[111] Ibid., 42.
[112] Ibid., 43.
[113] Ibid., 43, 44.
[114] 17 USCS § 107.
[115] *Andy Warhol Found. for the Visual Arts, Inc.*, 598 U.S. 510.
[116] *In re OpenAI, Inc. Copyright Infringement Litig.*, U.S., October 27, 110.
[117] Ibid., 96.
[118] Ibid.

speaking to their inherent non-transformative character. In *Advance*, the lack of transformative quality is even more apparent, as the outputs do not try to get as close as they can to the original because they are often the original themselves.[119] As seen above, their directly verbatim and slightly altered but substantively verbatim outputs intentionally mimic, if not fully copy and paste, the author's original work such that there is not even a change of tone or form.[120] Thus, they evidently fail to advance any sort of further purpose of their own. In both cases, there is nothing new within these summations and copies that has not previously existed, demonstrating their inherent non-transformative nature and obvious commercial threat. Outputs that offer merely a repackaged, regurgitated, reiterated version of what has already been produced cannot be seen as transformative, regardless of their technological character.

As mentioned before, *Warhol* brought the purpose of the use, specifically whether it is for a commercial purpose, to the forefront.[121] For LLM training, the intention is to create a profit—with OpenAI raking in 13 billion dollars in annual revenue, over 70% of which coming from users paying 20 dollars a day to chat with an AI.[122] The commercial purpose of OpenAI, and thus LLM companies in general—to profit off their outputs—cannot exist simultaneously with the author's goal of profiting off of their original work. OpenAI is not seeking to innovate to a new level of human advancement in technology through their clipped summaries that infringe on an author's style; they are merely attempting to profit off of an original artistic creation. The commercial threat in *Advance* is even more obvious, with the outputs directly undercutting the author's original works by outputting works behind a paywall. Even for

[119] *Advance Loc. Media LLC*, 10.
[120] Ibid., 11.
[121] *Andy Warhol Found. for the Visual Arts, Inc.*, 510.
[122] Connie Loizos, "OpenAI Has Five Years to Turn $13 Billion into $1 Trillion," *Yahoo Finance*, October 1, 2025, https://finance.yahoo.com/news/openai-five-years-turn-13-053936019.html.

materials not behind a paywall, it is evident that users would have no cause to look for original works of authorship when they can find a direct, or practically identical, copy of such work easily through a quick LLM output. Further, although "transformativeness" is only one factor of the analysis, as put forth in *Hachette*, "a use of copyrighted material that merely repackages or republishes the original is unlikely to be deemed a fair use."[123] Thus, factor one must weigh against fair use when taking infringing outputs into account.

VIc. Factor Two: Nature of the Copyrighted Work

Regarding training and outputs, the materials LLMs are utilizing infringe the second factor in both of its elements. As set out in *Blanch v Koons* (2006), this factor considers whether the materials used are "expressive or creative" and whether or not they are published, with works that are both expressive or creative and published falling against a fair use designation[124] Further, a court is more likely to find this factor weighing toward a fair use assessment when the work is factual or informational.[125] LLMs utilize works published on the internet for their expressive and creative quality. They are also using, as evidenced by the outputs above, particularly the *Game of Thrones* output in *In re OpenAI.,* materials that possess a unique character far different from a factual or informational work.[126] This factor favors authors and publishers, as it protects original, published, creative expression from unauthorized use and reproduction of the kind performed by LLMs and reflected in their outputs.

VId. Factor Three: Amount and Substantiality of the Portion Used in Relation to the Copyrighted Work as a Whole

[123] *Hachette Book Grp.,* Inc.163, U.S., 181
[124] *Blanch v. Koons*, 467 F.3d 244 (2d Cir. 2006).
[125] Ibid., 256
[126] *In re OpenAI, Inc. Copyright Infringement Litig.*, U.S., October 27

As stated above, the usage of entire copyrighted works to train an LLM can be viewed as reasonable and necessary because of the direct correlation between the amount of works an LLM inputs and their capabilities (as whether the amount was reasonable is considered in direct relation to the proposed transformative purpose).[127] However, *Authors Guild v Google Inc.* (2015) reminds that, "what matters is not so much 'the amount and substantiality of the portion used' in making a copy, but rather the amount and substantiality of what is thereby made accessible to a public [in the purported secondary use] for which it may serve as a competing substitute [for the primary use]."[128] Considering LLM outputs, which are indeed the secondary use of LLM training, such extensive copying is entirely unreasonable, as it applies to a non-transformative purpose of regurgitating works verbatim or through paralleled summation. LLM outputs are made accessible to the public in a way that an authored work could never be. Available at the click of a button for instantaneous generation by anybody with access to the internet, outputs are of the utmost accessibility, particularly in a way that has the potential for market harm. As shown above, when such outputs are practically identical to the original work, the amount made available to the public is basically the original work in its entirety. Thus, this factor too lands against fair use when accessing LLM outputs.

VIe. Factor Four: The Effect of the Use Upon the Potential Market for or Value of the Copyrighted Work

LLM outputs diverge from their training when assessing the fourth factor of fair use. The analogy of an LLM's learning and output to a child learning to read and producing competitive works because of such education falls apart when the actual LLM product is considered. An LLM can create what a child cannot and enter it into the competitive marketplace with a fraction

[127] Berg, "OpenAI Can't Strike Authors' Pirated Book Download Claims."
[128] *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

of the effort it would require a human to do so. Simply, a human cannot produce competitive outputs that can in any way measure up to those of an LLM. Judge Vince Chhaabria says it best in his response to the *Bartz* analogy: "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 minuscule fraction of the time and creativity it would otherwise take."[129] It is not necessarily the fact of the training, but the intent to which it is put that is misaligned with the intentions of the Copyright Act and fair use.

The first form of market harm comes in the form of regurgitated outputs. It is clear that regurgitated outputs, such as those in the aforementioned cases, bring to the marketplace a "competing substitute for the original, or its derivative."[130] Evidently, what incentive would a consumer have to seek out and pay for an entire work if they could instead access a free, expressively similar summary in a fraction of a second, or obtain subscription-based news content through LLM outputs that replicate it? Allowing such behavior to exist in the marketplace may benefit the public in terms of information availability, but it ignores the protections that the Copyright Act is intended to provide for creators. A consumer may gain greater access to knowledge and quick summations, but an author's original creative expression is now hijacked for purposes they did not authorize. This factor is also meant to consider the market harm if such conduct was widespread and unrestricted.[131] If all LLMs were to advance to this degree and continue to populate the internet with stolen works and expressively similar summaries, there would be no reason for authors to create at all, as it can be inferred that consumers would turn to these free outlets over paying for original source material.

[129] *Kadrey,* 788 U.S., 1036.
[130] *In re OpenAI, Inc.*, 800 U.S., September 19, 2025, 75.
[131] *Hachette Book Grp., Inc. v. Internet Archive*, 115 U.S., 389.

The second form of market harm comes in the form of entirely "original" LLM works. This issue is discussed in *Kadrey* as it pertains to the training of LLMs, but here it will be understood regarding outputs.[132] In this case, the court assumed that people would be able to use LLMs to create books and then sell them in the marketplace.[133] This was found indicative of "indirect" substitution. Such substitution is not often of great relevance to copyright cases, as it typically deals with one original work being compared to a single secondary work, which is unlikely to pose great potential for harm. This case, however, pertains to technology powerful enough to instantaneously generate countless competitive, indirectly substantive, works.[134] Thus, indirect substitution does in fact have a real threat of diluting the market within the quantity, and similar quality, of works it can generate.[135] *Kadrey* cites a specific example: "if someone bought a romance novel written by an LLM instead of a romance novel written by a human author, the LLM-generated novel is substituting for the human-written one."[136] It is clear, then, that the substitutability of LLM outputs for original works may vary depending on the type of work, with nonfiction works perhaps being easier to displace than fiction ones.[137] In any case, books produced by an LLM have the potential to compete for sales both simply as alternative products and by flooding markets with artificial works such that original works are more likely to go unnoticed.[138] As aforementioned, in the *Kadrey* case this factor favored fair use, but as the court acknowledged, it was a conclusion "in significant tension with reality" that was only found because the plaintiffs failed to provide adequate evidence of market harm.[139] Though such technology is just beginning to take hold, there has still been a plethora of AI-generated books

[132] *Kadrey,* 788 U.S., 1055.
[133] Ibid., 1052.
[134] Ibid., 1054.
[135] Ibid.
[136] Ibid.
[137] Ibid., 1035.
[138] Ibid., 1052.
[139] Ibid., 1057.

entering the marketplace on Amazon and diluting the market.[140] These books take the form of summaries, AI-generated biographies, and copycat books. Not only do they deprive authors of their rightful profits, but they also harm their reputation by producing low-quality work under author aliases.[141] Here, consumers are very likely to opt to acquire the copy as they are not even aware of any distinction between it and the original. Where AI books are allowed to enter the marketplace, either in the form of entirely original works or those purporting to be written by a human author, a human author can no longer compete.

Thus, there is evidently market harm found in any outputs produced by LLMs—both in the form of copied iterations, summations, and "original" outputs, weighing this factor against fair use.

VII. Re-Weighing the Four Factors of Section 107

Fair use was intended to be a flexible and open-ended doctrine to accommodate and encourage technological innovation.[142] In opposition to fair dealing, which only permitted copying for certain stringent purposes, fair use was intended, as Awad writes, to function as a "legal doctrine that allows others to use a copyrighted work of authorship for purposes other than that intended by the author."[143] Its flexibility has allowed technology to prosper—permitting thumbnails and search engines in *Field v. Google Inc* (2006) and *Perfect 10, Inc. v. Amazon.com, Inc.* (2007).[144] However, the doctrine was created in 1976 and was thus established at a time when the idea of artificial intelligence competing in a marketplace with human work was entirely

[140] Ibid., 1052.
[141] Andrew Limbong, "Authors Push Back on the Growing Number of AI 'Scam' Books on Amazon," *NPR*, March 13, 2024, https://www.npr.org/2024/03/13/1237888126/growing-number-ai-scam-books-amazon.
[142] *Hachette Book Grp., Inc.* 115 U.S., 179
[143] Awad, "Generative AI's Copyright Enigma: A Comparative Study of Fair Use and Fair Dealing."
[144] Ibid., 48.

inconceivable.[145] Where search engines and thumbnails simply organize information, making the original source available to the public in an accessible manner, generative AI models become the copyrighted material themselves. By doing so, LLMs effectively function as a stand-in for original, human creation, and as such cannot continue to find protection under the fair use flexibility regarding technology. Protecting innovation is one thing, but continuing to support a machine and industry founded entirely upon copyright infringement is another.

The problem, however, as seen above, is that it is difficult for courts to overlook the evident transformative quality of an LLM. In cases without clear output data, such as *Kadrey* and *Bartz,* it appears nearly impossible for the court to dispute the AI companies' argument that their usage transforms authors' works into a mechanized form incomparable with the original work. For cases with output data, such as *In re OpenAI* and *Advance,* it remains to be seen whether courts will determine such a use as fair or not. However, the problem originates from the *Kadrey* and *Bartz* findings, that is, in the finding of the LLM training as fair. It is at this stage of training that LLMs become knowledge machines capable of infringement, and it is also at this stage that they should not be granted free access to copyrighted materials. Policing LLM outputs is simply inconceivable, and thus the only way to regulate them is to halt the learning that makes their communications possible. To do so, however, requires rethinking the recent trajectory of the fair use doctrine, particularly the overemphasis on transformative quality.

Section 107 of the Copyright Act contains four factors. And yet, undue emphasis has been placed on the first factor, specifically the "transformative" subclause. In *Campbell v Acuff-Rose Music* (1994), the court laid out that, "the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of

[145] 17 USCS § 107.

fair use."[146] Following *Campbell* came *Cariou v Prince* (2013), which determined the transformative issue to be the "heart of the fair use inquiry."[147] This decree has led the transformative factor, as aforementioned, to become determinative of a fair use verdict over 91 percent of the time.[148]

Such a singular focus on the transformative component, an idea not even present within the language of the first factor of Section 107 which asks the court only to consider the purpose and character of the use, poses a unique threat to authors' ownership over their works. This approach has been called into question by numerous courts, and Awad cites a few of these with words of skepticism: the Seventh Circuit in *Kienitz v. Sconnie Nation LLC* (2014) stated that "asking exclusively whether something is 'transformative' not only replaces the list in §107 but also could override 17 U.S.C. §106 (2) which protects derivative works," and Justice Clarence Thomas of the U.S. Supreme Court stated in *Google LLC v. Oracle Am.* (2021) that "courts have wrongly conflate[d] transformative use with derivative use."[149] A derivative work is, as paraphrased from Section 101 of the Copyright Act, a work based upon a preexisting work that, as a whole, represents an original work of authorship.[150] The outputs of LLMs found in *In re OpenAI* and *Advance* seem to be exactly of this derivative character, sometimes being the original work themselves or attempting to fill in for it with summation and paraphrasing. And yet, they may find protection under the token of "transformativeness." Thus, bolstering the transformative factor taints not only the protections granted to authors in the other factors of Section 107, but also those they possessed pertaining derivative works in Section 106. Evidently,

[146] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994).
[147] Rollin Ransom, "The 'Transformation' of the Copyright Fair Use Test," *Law360*, December 19, 2014, https://www-law360-com.eu1.proxy.openathens.net/articles/603602/the-transformation-of-the-copyright-fair-use-test.
[148] Ibid.
[149] Ibid.
[150] Ibid.

this emphasis poses a clear threat to the long-standing protections authors maintain over their authorship.

The overemphasis on the transformative subclause has created the scenario in which LLM training can be permissible under fair use as LLMs will always possess a robotic, technological expression that is in some way inventive from an original work's form. But a new fair use doctrine is not needed to inhibit LLM training on copyrighted materials. The four factors of Section 107 are in fact equipped to effectively grapple with the technology LLMs present if they are, as *Campbell* states, "weighed together."[151] As previously argued, LLM training cannot be separated from outputs. And yet, considering "transformativeness" does exactly that, effectively allowing LLM training to claim fair use purely based on innovative mechanics while ignoring the ways in which their usage violates the other factors. Although fair use intends to promote new technology, it does not universally grant it a free pass. Instead, factors two, three, and four all provide counterweights to a doctrine that, left up to purpose and character, allows for large-scale commercially competitive usage of similar expression. Simply, the bar of fair use is too low when centralized in just one of its factors; there is more to the fair use doctrine than simple invention.

Although the other factors are meant to be taken together with the first, they have more recently become unduly influenced by it. The second and third factors, as expressed above, hardly favor fair use, as LLMs are using works for their expressive quality and in their entirety. And yet, the transformative factor has impacted how they are weighed. In *Kadrey,* for example, the court laid out that with the amount that is reasonable to use being considered in direct relation to the proposed transformative purpose.[152] The same process has occurred regarding

[151] *Campbell v. Acuff-Rose Music, Inc.,* 510. U.S., 574-578.
[152] *Kadrey,* 788 U.S., 1050.

commercialism, with *Kadrey* stating, "commercialism isn't dispositive of the first factor and tends to be less important when the secondary use is highly transformative."[153] *Warhol* linked the first and fourth factors, tying purpose to commercialism, and thus even this point is impacted by the emphasis on transformation.[154] The transformative element then takes on other factors, leading to a fair use argument for copying a work in full and utilizing it for a commercial purpose. If AI companies did not choose works for their expressive quality and only used parts of them for a non-commercial intention, then it seems unlikely that outputs would contain such blatant regurgitation and pose such a real threat to authors' works. But this is not how LLMs are currently trained: they use authors' works for a purpose and that purpose is expressed in competitive, infringing, outputs. Continuing to look past these characteristics of LLM training and instead focus on the transformative nature of the "tokens" LLMs consume allows for the type of outputs that create market harm for authors, as they possess the same expression and quantity as their original work.

Thus, looking forward, in cases both with and without output evidence, courts must consider the factors of Section 107 as they are intended. By doing so, a more nuanced understanding of the boundaries new technologies must adhere to in their usage of copyrighted materials can develop. Although it may be difficult to ascertain commercial threats without output data, the other factors, regarding expression and quantity of the use, easily fall against fair use even for the training period. When these factors are weighed alongside, instead of in deference to, the transformative factor, fair use becomes a much stronger, impenetrable doctrine that provides greater protection to human works. Although the extent of transformative quality is salient to a fair use decree, so too are the other factors Congress laid out. While LLM training

[153] Ibid., 1046.
[154] *Andy Warhol Found for the Visual Arts, Inc.*, 598 U.S., 510.

may be inherently transformative, this character cannot be considered above all else, lest AI come to run rampant over human creation. Whether in training or output, there is something amiss about the way LLMs have come to take on human character and expression as their own. Thus, it is up to the four factors of fair use to, together, halt such artificial, competitive creation.