

UNITED STATES DISTRICT COURT  
NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADREY, et al.,  
Plaintiffs,  
v.  
META PLATFORMS, INC.,  
Defendant.

---

Case No. 23-cv-03417-VC

**ORDER DENYING THE PLAINTIFFS’  
MOTION FOR PARTIAL SUMMARY  
JUDGMENT AND GRANTING  
META’S CROSS-MOTION FOR  
PARTIAL SUMMARY JUDGMENT**

Re: Dkt. Nos. 482, 501

Companies are presently racing to develop generative artificial intelligence models—software products that are capable of generating text, images, videos, or sound based on materials they’ve previously been “trained” on. Because the performance of a generative AI model depends on the amount and quality of data it absorbs as part of its training, companies have been unable to resist the temptation to feed copyright-protected materials into their models—without getting permission from the copyright holders or paying them for the right to use their works for this purpose. This case presents the question whether such conduct is illegal.

Although the devil is in the details, in most cases the answer will likely be yes. What copyright law cares about, above all else, is preserving the incentive for human beings to create artistic and scientific works. Therefore, it is generally illegal to copy protected works without permission. And the doctrine of “fair use,” which provides a defense to certain claims of copyright infringement, typically doesn’t apply to copying that will significantly diminish the ability of copyright holders to make money from their works (thus significantly diminishing the incentive to create in the future). Generative AI has the potential to flood the market with endless

amounts of images, songs, articles, books, and more. People can prompt generative AI models to produce these outputs using a tiny fraction of the time and creativity that would otherwise be required. So by training generative AI models with copyrighted works, companies are creating something that often will dramatically undermine the market for those works, and thus dramatically undermine the incentive for human beings to create things the old-fashioned way.

Take, for example, biographies. If a company uses copyrighted biographies to train a model, and if the model is thus capable of generating endless amounts of biographies, the market for many of the copied biographies could be severely harmed. Perhaps not the market for Robert Caro's *Master of the Senate*, because that book is at the top of so many people's lists of biographies to read. But you can bet that the market for lesser-known biographies of Lyndon B. Johnson will be affected. And this, in turn, will diminish the incentive to write biographies in the future.

Or take magazine articles. If a company uses copyrighted magazine articles to train a model capable of generating similar articles, it's easy to imagine the market for the copied articles diminishing substantially. Especially if the AI-generated articles are made available for free. And again, how will this affect the incentive for human beings to put in the effort necessary to produce high-quality magazine articles?

With some types of works, the picture is a bit murkier. For example, it's not clear how generative AI would affect the market for memoirs or autobiographies, since by definition people read those works because of who wrote them. With fiction, it might depend on the type of book. Perhaps classic works of literature like *The Catcher in the Rye* would not see their markets diminished. But the market for the typical human-created romance or spy novel could be diminished substantially by the proliferation of similar AI-created works. And again, the proliferation of such works would presumably diminish the incentive for human beings to write romance or spy novels in the first place.

Some students of copyright law respond that none of this matters because when companies use copyrighted works to train generative AI models, they are using the works in a

way that's highly creative in its own right. In the language of copyright law, the companies' use of the works is "transformative." As a factual matter, there's no disputing that. And as a legal matter, it's true that you're less likely to be liable for copyright infringement if you're copying the work for a transformative purpose. In that situation, you're more likely to be protected by the fair use doctrine. But as the Supreme Court has emphasized, the fair use inquiry is highly fact dependent, and there are few bright-line rules. There is certainly no rule that when your use of a protected work is "transformative," this automatically inoculates you from a claim of copyright infringement. And here, copying the protected works, however transformative, involves the creation of a product with the ability to severely harm the market for the works being copied, and thus severely undermine the incentive for human beings to create. Under the fair use doctrine, harm to the market for the copyrighted work is more important than the purpose for which the copies are made.

Speaking of which, in a recent ruling on this topic, Judge Alsup focused heavily on the transformative nature of generative AI while brushing aside concerns about the harm it can inflict on the market for the works it gets trained on. Such harm would be no different, he reasoned, than the harm caused by using the works for "training schoolchildren to write well," which could "result in an explosion of competing works." Order on Fair Use at 28, *Bartz v. Anthropic PBC*, No. 24-cv-5417 (N.D. Cal. June 23, 2025), Dkt. No. 231. According to Judge Alsup, this "is not the kind of competitive or creative displacement that concerns the Copyright Act." *Id.* But when it comes to market effects, using books to teach children to write is not remotely like using books to create a product that a single individual could employ to generate countless competing works with a miniscule fraction of the time and creativity it would otherwise take. This inapt analogy is not a basis for blowing off the most important factor in the fair use analysis.

Another argument offered in support of the companies is more rhetorical than legal: Don't rule against them, or you'll stop the development of this groundbreaking technology. The technology is certainly groundbreaking. But the suggestion that adverse copyright rulings would

stop this technology in its tracks is ridiculous. These products are expected to generate billions, even trillions, of dollars for the companies that are developing them. If using copyrighted works to train the models is as necessary as the companies say, they will figure out a way to compensate copyright holders for it.

The upshot is that in many circumstances it will be illegal to copy copyright-protected works to train generative AI models without permission. Which means that the companies, to avoid liability for copyright infringement, will generally need to pay copyright holders for the right to use their materials.

But that brings us to this particular case. The above discussion is based in significant part on this Court's general understanding of generative AI models and their capabilities. Courts can't decide cases based on general understandings. They must decide cases based on the evidence presented by the parties.

In this case, thirteen authors—mostly famous fiction writers—have sued Meta for downloading their books from online “shadow libraries” and using the books to train Meta's generative AI models (specifically, its large language models, called Llama). The parties have filed cross-motions for partial summary judgment, with the plaintiffs arguing that Meta's conduct cannot possibly be fair use, and with Meta responding that its conduct must be considered fair use as a matter of law. In connection with these fair use arguments, the plaintiffs offer two primary theories for how the markets for their works are affected by Meta's copying. They contend that Llama is capable of reproducing small snippets of text from their books. And they contend that Meta, by using their works for training without permission, has diminished the authors' ability to license their works for the purpose of training large language models. As explained below, both of these arguments are clear losers. Llama is not capable of generating enough text from the plaintiffs' books to matter, and the plaintiffs are not entitled to the market for licensing their works as AI training data. As for the potentially winning argument—that Meta has copied their works to create a product that will likely flood the market with similar works, causing market dilution—the plaintiffs barely give this issue lip service, and they present no

evidence about how the current or expected outputs from Meta’s models would dilute the market for their own works.

Given the state of the record, the Court has no choice but to grant summary judgment to Meta on the plaintiffs’ claim that the company violated copyright law by training its models with their books. But in the grand scheme of things, the consequences of this ruling are limited. This is not a class action, so the ruling only affects the rights of these thirteen authors—not the countless others whose works Meta used to train its models. And, as should now be clear, this ruling does not stand for the proposition that Meta’s use of copyrighted materials to train its language models is lawful. It stands only for the proposition that these plaintiffs made the wrong arguments and failed to develop a record in support of the right one.

## **I. COPYRIGHT LAW AND FAIR USE**

The goal of copyright law is to promote “broad public availability of literature, music, and the other arts.” *Twentieth Century Music Corp. v. Aiken*, 422 U.S. 151, 156 (1975). To this end, copyright law incentivizes creativity by giving authors of original works a bundle of exclusive rights—for instance, the rights to prevent others from reproducing or distributing the works. 17 U.S.C. § 106. At the same time, however, copyright law “trades off the benefits of incentives to create against the costs of restrictions on copying.” *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 526 (2023). For example, copyright only protects expression, not underlying ideas, and the duration of copyright protection is limited. *See id.* (citing 17 U.S.C. §§ 102, 302–305).

One major way the Copyright Act strikes a balance between protecting ownership and leaving room for innovation is through the affirmative defense of fair use. Under this doctrine, “the fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching . . . , scholarship, or research, is not an infringement of copyright.” 17 U.S.C. § 107. Fair use “permits courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster.” *Google LLC v. Oracle America, Inc.*, 593 U.S. 1, 18 (2021) (quoting *Stewart v. Abend*, 495 U.S. 207, 236 (1990)).

The Copyright Act lists four factors to be considered in determining whether a given use is fair:

1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
2. the nature of the copyrighted work;
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
4. the effect of the use upon the potential market for or value of the copyrighted work.

17 U.S.C. § 107.

While the statute lists these four factors, fair use is a “flexible concept.” *Warhol*, 598 U.S. at 527 (quotation marks omitted) (quoting *Oracle*, 593 U.S. at 20). The list is not exhaustive. A particular factor “may prove more important in some contexts than in others.” *Oracle*, 593 U.S. at 19. And application of the factors “requires judicial balancing, depending upon relevant circumstances, including ‘significant changes in technology.’” *Id.* (quoting *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417, 430 (1984)). The factors may also overlap such that facts relevant to one factor are also relevant to others. *See A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 642 (4th Cir. 2009). Overall, the factors are not meant to be applied mechanically, but to contribute “to a holistic inquiry”: whether the secondary work is likely to substitute for the original work in the marketplace and therefore undermine the incentive to create. *See Romanova v. Amilus Inc.*, 138 F.4th 104, 117 n.9 (2d Cir. 2025) (Leval, J.); *see also Warhol*, 598 U.S. at 528 (referring to substitution as “copyright’s *bête noire*”).

Because it “focuses on actual or potential market substitution,” *Warhol*, 598 U.S. at 536 n.12, the fourth factor is “undoubtedly the single most important element of fair use,” *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 566 (1985). If the law allowed people to copy your creations in a way that would diminish the market for your works, this would diminish your incentive to create more in the future. Thus, the key question in virtually any case where a defendant has copied someone’s original work without permission is whether allowing people to engage in that sort of conduct would substantially diminish the market for the original

work. *See Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 (1994).

Because fair use is an affirmative defense, the burden of proof is on the party invoking it. *Dr. Seuss Enterprises, L.P. v. ComicMix LLC*, 983 F.3d 443, 459 (9th Cir. 2020), *abrogated on other grounds by Jack Daniel’s Properties, Inc. v. VIP Products LLC*, 599 U.S. 140 (2023). In particular, because the fourth factor is the most important, the secondary user (generally, the defendant) will “have difficulty carrying the burden of demonstrating fair use without favorable evidence about relevant markets.” *Campbell*, 510 U.S. at 590. But while the rightsholder need not prove or present evidence of market harm, they “may bear some initial burden of *identifying* relevant markets.” *Hachette Book Group, Inc. v. Internet Archive*, 115 F.4th 163, 194 (2d Cir. 2024); *see also Newegg Inc. v. Ezra Sutton, P.A.*, No. CV 15-01395, 2016 WL 6747629, at \*2 (C.D. Cal. Sep. 13, 2016). Moreover, because fair use is a holistic inquiry, the party invoking it “bears the burden on the defense as a whole,” not as to each individual factor. William F. Patry, *Patry on Fair Use* § 2:5 (May 2025 ed.).

Fair use is a mixed question of law and fact, but the “question primarily involves legal work.” *Oracle*, 593 U.S. at 24. Therefore, fair use can be addressed at summary judgment where there are no genuine issues of material fact relevant to fair use. *Leadsinger, Inc. v. BMG Music Publishing*, 512 F.3d 522, 530 (9th Cir. 2008). By contrast, where there are genuine factual disputes that might affect whether the defendant’s use was fair, those disputes must be resolved by a jury. *See Oracle*, 593 U.S. at 23–25. Once a jury finds the facts, whether those facts support fair use “is a legal question for judges to decide.” *Id.* at 23–24.

It bears emphasis that where a fair use defense fails, the consequence isn’t necessarily that the defendant must stop whatever they were doing. The consequence will often be that the defendant needs to pay the copyright owner for a license that grants them permission to do whatever they were doing. This way, the defendant compensates the copyright owner for the fact that the defendant’s conduct will otherwise harm the market for the original work. The defendant will only be forced to stop what they’re doing if they’re unwilling or unable to pay for the right to do it.

## II. FACTS AND PROCEDURAL HISTORY

### A

“Generative AI” is a type of artificial intelligence that creates new content, such as text, images, videos, or sound.<sup>1</sup> Generative AI models do this, as Meta describes it, by extracting “increasingly complex mathematical patterns from training data, enabling the network to output a prediction or decision based on the patterns derived.” To put it more simply, generative AI models are “trained” to identify common patterns across large training datasets. They can then create, in response to user prompts, new content based on the patterns they have recognized in that training data. By the same token, a model’s outputs are limited based on the patterns that existed in its training data. For instance, if the only bridge in an image-generating model’s training data was the Golden Gate Bridge, and a user told that model to generate an image of a bridge, it would likely generate an orange-red suspension bridge, because that is the pattern of a bridge that would emerge from its training data.

A large language model, or LLM, is a particular type of generative AI model designed to understand and generate text. Users can prompt LLMs to do a wide range of things, such as draft emails, summarize documents, or write computer code. Well-known LLMs include OpenAI’s ChatGPT models and Google’s Gemini models.

LLMs learn to understand language by analyzing relationships among words and punctuation marks in their training data. The units of text—words and punctuation marks—on which LLMs are trained are often referred to as “tokens.” LLMs are trained on an immense amount of text and thereby learn an immense amount about the statistical relationships among words. Based on what they learned from their training data, LLMs can create new text by predicting what words are most likely to come next in sequences. This allows them to generate text responses to basically any user prompt. Model developers may also “post-train” or “finetune” their models to improve their performance at specific tasks or otherwise adjust their outputs, such as to prevent generation of offensive statements. Therefore, as with other

---

<sup>1</sup> Except as noted, the parties do not dispute the facts described in this section.



generative AI models, LLMs’ outputs are limited by their training data. To be able to generate a wide range of text—in different languages or styles, or regarding different subject matter—an LLM’s training dataset must be large and diverse. As one Meta witness put it, “If a model only saw social media posts, for example, it would not do well in generating source computer code.”

But while a variety of text is necessary for training, books make for especially valuable training data. This is because they provide very high-quality data for training an LLM’s “memory” and allowing it to work with larger amounts of text at once. (The technical term for how many tokens an LLM can hold in its memory at once is its “context window.”) For instance, an LLM with a better memory will be able to process and respond to longer prompts, incorporate more information into outputs, and remember things from earlier in an exchange, resulting in smoother “conversations.” Books are good data for training LLMs’ memories because, in the words of one of Meta’s expert witnesses, they are “long but consistent,” maintaining a particular style and coherent structure. They are also high quality in the sense that they generally are well written and use proper grammar (especially compared to text from the internet, which varies widely on these metrics).

## **B**

Meta Platforms owns and operates social media services including Facebook, Instagram, and WhatsApp. It is also the developer of a series of LLMs named “Llama.” Meta released Llama 1 in February 2023 and Llama 2 that July. Llama 3—along with Meta AI, an easily accessible AI chatbot (analogous to ChatGPT) that incorporates Llama 3—was released in April 2024. Llama 4 is planned for release later in 2025. As Meta explains, each new Llama edition improved in certain ways over its predecessor: Llama 2 was finetuned “to improve the safety, quality, and consistency” of its outputs; Llama 3 made “significant improvements in performance and efficiency”; and Llama 4 is generally “larger” and “more advanced.” Subject to certain restrictions, members of the public can download all of the Llama models for free for noncommercial use; Llama 2 and 3 are also free to download for commercial use. While the Llama models are free to download, Meta estimates that its total revenue from generative AI will

range from \$2 to \$3 billion in 2025, and from \$460 billion to \$1.4 trillion over the next ten years. *See* Pls. MSJ Ex. 8 at 12.

To get the varied and extensive text necessary to train its models, Meta cast a wide net. Approximately two-thirds of the data used to train Llama 1 and 2 came from Common Crawl, a nonprofit organization that collects and provides free access to website data, metadata, and text. The remainder came from websites and databases including Wikipedia, GitHub, ArXiv, Stack Exchange, and a combination of Project Gutenberg and Books3 (two book databases).<sup>2</sup> With the exception of Books3, none of the sources contained any copyrighted material at issue in this case.

Although Meta needed (and acquired and used) a wide range of training data, it especially needed books because, as discussed above, books make for high-quality data. Meta AI researchers and engineers repeatedly discussed the benefits of using books as training data, as well as the need to acquire more books for this use. One Meta employee said that the “best resources we can think of are definitely books.” Pls. MSJ Ex. 18 at 2. Another said it was “really important for us to get books data ASAP.” *Id.* Ex. 40 at 2. So as Meta expanded its datasets generally, it also continued to look for more books in particular.

At first, Meta wanted to license books and so tried to negotiate licensing deals with several major publishers. Meta’s head of generative AI discussed spending up to \$100 million on licensing. But as negotiations proceeded, Meta realized that licensing would be more difficult than anticipated. For one thing, publishers generally do not hold the subsidiary rights to license books for AI training. These rights are instead held by individual authors, and there is no organization for collective licensing of such rights. Sinkinson Decl. ISO Meta MSJ ¶¶ 58–59, 62. Even where publishers do hold AI training licensing rights, they do so regionally rather than globally. Meta MSJ Ex. 34 at 22:22–25:15. For another thing, some publishers apparently

---

<sup>2</sup> According to Meta, GitHub is “a leading cloud-based platform where coders store and share code, frequently on an open-source basis.” ArXiv is “a free online archive of math, science, and economics papers.” Stack Exchange is “a network of question-and-answer websites for sharing technical knowledge, geared toward the programming community.”

ignored Meta’s outreach, and only one gave Meta a pricing proposal. *Id.* at 23:11–14, 24:2–10.

Eventually, Meta began investigating the possibility of procuring the books (and other text) needed for training by downloading them from “shadow libraries.” A shadow library is an online repository that provides things like books, academic journal articles, music, or films for free download, regardless of whether that media is copyrighted. Meta first used a shadow library in October 2022, when it downloaded the Library Genesis (“LibGen”) database to investigate whether there was value in training Llama on the works it contained. Pls. MSJ Ex. 32 at 3. If the answer was yes, the plan was to then set up licensing agreements for those or similar works. *Id.* But in spring 2023, after failing to acquire licenses and following escalation to CEO Mark Zuckerberg, Meta decided to just use the works acquired from LibGen as training data. *Id.* Ex. 61 at 5. And after confirming that LibGen contained most of the works available for license from certain publishers with which it had been negotiating, Meta abandoned its licensing efforts. *Id.* Ex. 50 at 131:1–132:10, 383:5–384:12; *id.* Ex. 57 at 2; *id.* Ex. 58 at 3; *see also id.* Ex. 92 at 12. In early 2024, Meta also downloaded Anna’s Archive, a compilation of shadow libraries including LibGen, Z-Library, and others. *See id.* Ex. 66 at 2–3.

To download these large datasets more quickly and without unnecessarily slowing down its networks, Meta torrented them. “Torrenting” is a filesharing technique that entails the simultaneous distribution of small portions of a larger file from many different sources. To be more precise, those sources are many other computer systems that also contain that file. So, for instance, one who torrented LibGen would download small pieces of each book LibGen contains from other users who had copies of LibGen on their computer and who were participating in the torrenting network. The torrenting software would then take those pieces and reassemble them into the original files on the downloader’s computer.<sup>3</sup>

Certain torrenting protocols—including the one used by Meta, called BitTorrent—are, by

---

<sup>3</sup> *See generally* David Gerwitz, *What Is Torrenting?*, ZDNET (Aug. 6, 2024), <https://www.zdnet.com/article/what-is-torrenting-and-how-does-it-work> [<https://perma.cc/8PG5-H7UW>].

default, configured so that files downloaded via torrenting may also be reuploaded to other computer systems. This reuploading can occur both while files are still being downloaded (which the parties refer to as “leeching”) and after those files have been fully downloaded (which the parties refer to as “seeding”). Some torrenting protocols—including BitTorrent—are designed to prioritize downloads to users who are also uploading.

There is no dispute that Meta torrented LibGen and Anna’s Archive, but the parties dispute whether and to what extent Meta uploaded (via leeching or seeding) the data it torrented. A Meta engineer involved in the torrenting wrote a script to prevent seeding, but apparently not leeching. *See* Pls. MSJ at 13; *id.* Ex. 71 ¶¶ 16–17, 19; *id.* Ex. 67 at 3, 6–7, 13–16, 24–26; *see also* Meta MSJ Ex. 38 at 4–5. Therefore, say the plaintiffs, because BitTorrent’s default settings allow for leeching, and because Meta did nothing to change those default settings, Meta must have reuploaded “at least some” of the data Meta downloaded via torrent. The plaintiffs assert further that Meta chose not to take any steps to prevent leeching because that would have slowed its download speeds. Meta responds that, even if it reuploaded some of what it downloaded, that doesn’t mean it reuploaded any of the plaintiffs’ books. It also notes that leeching was not clearly an issue in the case until recently, and so it has not yet had a chance to fully develop evidence to address the plaintiffs’ assertions.

Either way, Meta added the books it downloaded to the datasets it used to train the Llama models. It also post-trained its models to prevent them from “memorizing” and outputting certain text from their training data, including copyrighted material. These training efforts, which Meta calls “mitigations,” appear to have been successful. Meta’s expert witness tested them using a method designed to get LLMs to regurgitate material from its training data (which Meta calls “adversarial prompting”). Even using that method, the expert could get no model to generate more than 50 words and punctuation marks (that is, “tokens”) from the plaintiffs’ books. And the plaintiffs’ expert could only get the Llama model best at regurgitation to generate 50 words and punctuation marks from the plaintiffs’ books in 60% of tests. She also testified that Llama was not able to reproduce “any significant percentage” of them. Meta MSJ Ex. 24 at 237:16–19; *see*

also Pls. Ex. 79 ¶¶ 70–72, 79, 82–83, 92; Meta MSJ Ex. 23 at 179:22–25, 180:17–181:16. In short, Llama cannot currently be used to read or otherwise meaningfully access the plaintiffs’ books.

### C

The plaintiffs are thirteen published authors who have written, and who hold copyright in, various works. Those works are mostly novels, but also include plays, short stories, memoirs, essays, and nonfiction books. Examples include Sarah Silverman’s *The Bedwetter*, a comic memoir; Rachel Louise Snyder’s *No Visible Bruises: What We Don’t Know About Domestic Violence Can Kill Us*, a nonfiction book about domestic violence and how to combat it; Junot Díaz’s Pulitzer Prize–winning novel, *The Brief Wondrous Life of Oscar Wao*; and Andrew Sean Greer’s *Less*, also a Pulitzer Prize–winning novel. All of the books in which the plaintiffs hold copyright can be found in the datasets Meta downloaded, including both Books3 and the Anna’s Archive databases. In total, Meta downloaded at least 666 copies of books whose copyrights the plaintiffs hold.

Each plaintiff says that they would be open to licensing their books for use as generative AI training data, but that Meta did not approach them about this licensing. No plaintiff has licensed a book to any company for use as LLM training data or been asked by any company to license a book for that purpose.

The plaintiffs filed this lawsuit seeking to represent a class of all owners of copyrighted works used as training data for Llama. They brought claims for direct copyright infringement (based on Meta’s reproduction of their books), vicarious copyright infringement, removal of copyright management information in violation of the Digital Millennium Copyright Act (DMCA), unfair competition, unjust enrichment, and negligence. The plaintiffs seek damages, restitution, and injunctive and declaratory relief, although it is not entirely clear what exactly they seek to enjoin. They did not, for instance, seek a preliminary injunction preventing Meta from using their works as training data or requiring Meta to retrain the existing Llama models on data excluding their books. *Cf. Concord Music Group, Inc. v. Anthropic PBC*, No. 24-cv-3811,

2025 WL 904333, at \*3–4 (N.D. Cal. Mar. 25, 2025) (discussing music publishers’ motion for a preliminary injunction seeking relief based on future training of AI models).

All of the claims except the direct copyright infringement claim were dismissed early on. The plaintiffs were later granted leave to amend to expand their copyright claim to encompass a theory of infringement by distribution (based on the allegation that Meta was reuploading the data it torrented), and to add both a different DMCA claim and a claim under the California Comprehensive Computer Data Access and Fraud Act (CDAFA). Meta moved to dismiss the new claims, and its motion was granted as to the CDAFA claim but denied as to the DMCA claim.

Often, the next step in a case like this is a motion for class certification by the named plaintiffs. But sometimes the parties will first move for summary judgment regarding the individual claims of the named plaintiffs. For defendants in such cases, there is a trade-off. On the one hand, a defendant could benefit by getting a favorable ruling and disposing of the case before being subjected to expensive and burdensome class-related discovery and motion practice. On the other hand, this favorable ruling for the defendant binds only the individual named plaintiffs, leaving all other members of the proposed class free to sue on the same claims. In this case, Meta proposed doing summary judgment regarding the individual claims of the named plaintiffs first, and the Court accepted this approach.

Thus, after the close of discovery relating to the merits of the named plaintiffs’ claims, the plaintiffs moved for partial summary judgment, arguing that they had made out a facial claim for copyright infringement and that Meta’s fair use defense could not possibly apply to negate that claim. Meta did not dispute that the plaintiffs established a facial case of infringement of their rights of reproduction. But Meta opposed the plaintiffs’ motion—and indeed filed its own cross-motion—on the ground that its reproduction was fair use as a matter of law.

Meta also moved for summary judgment as to the plaintiffs’ DMCA claim; that motion will be granted in a separate order. With respect to the plaintiffs’ claim that Meta infringed their copyrights by distributing their works (via leeching or seeding), neither side moved for summary

judgment, so this will remain a live issue in the case.<sup>4</sup>

Six friend-of-the-court briefs (that is, amicus briefs) were also filed. An assortment of intellectual property law professors and the Electronic Frontier Foundation (a civil liberties organization) filed briefs in support of Meta. An assortment of copyright professors; the Copyright Alliance (an organization of creators); the Association of American Publishers; and the International Association of Scientific, Technical and Medical Publishers filed briefs in support of the plaintiffs.

### III. FACTOR ONE: THE PURPOSE AND CHARACTER OF THE USE

The first factor “considers the reasons for, and nature of, the copier’s use of an original work.” *Warhol*, 598 U.S. at 528. Several things can be relevant to the “purpose and character” of a use. One is whether that use is “of a commercial nature or is for nonprofit educational purposes.” 17 U.S.C. § 107(1). Another might be whether it was made in good or bad faith (although whether this is relevant is unclear under current law). *See Oracle*, 593 U.S. at 32–33.

Primarily, however, the first factor focuses on whether the secondary use is “transformative”—that is, on whether and to what extent “the new work merely supersedes the

---

<sup>4</sup> The plaintiffs moved for summary judgment only on the grounds that “Meta copied [their] copyrighted books without permission” and that its “reproduction . . . without permission . . . is not fair use.” Pls. MSJ at vii, 19. The plaintiffs did at times suggest that their motion encompassed their distribution claim. *See, e.g., id.* at 22 (“Meta’s initial reproduction” was not fair use because it “result[ed] in distributing copyrighted material.”). But reproduction and distribution are separate rights that must be considered separately. *See* 17 U.S.C. § 106(1), (3); *Columbia Pictures Industries, Inc. v. Fun*, 710 F.3d 1020, 1034 (9th Cir. 2013) (“Both uploading and downloading copyrighted material are infringing acts. The former violates the copyright holder’s right to distribution, the latter the right to reproduction.”).

As discussed below, the specific manner of Meta’s reproduction (that is, torrenting the plaintiffs’ books from shadow libraries) is still relevant to whether that reproduction was fair use. But Meta’s alleged distribution must be addressed independently (unless, maybe, its acquisition necessarily involved distribution, which does not appear to be the case, *see* Pls. Ex. 67 at 106:14–108:25, 246:13–248:23, 270:12–16 (explaining that default settings could be changed such that leeching was not always occurring)). Even if the plaintiffs had moved for summary judgment as to whether any distribution was fair use, the record on Meta’s alleged distribution is incomplete, making summary judgment on that issue improper at this point in the case. *See* Order Granting as Modified Meta’s Request for Leave to File a Rebuttal Expert Report, Dkt. No. 499 (giving Meta leave to serve supplemental expert report on distribution, with deadline after Meta’s deadline to oppose the plaintiffs’ motion for summary judgment).



objects of the original creation (supplanting the original), or instead adds something new, with a further purpose or different character.” *Warhol*, 598 U.S. at 528 (cleaned up). Allowing a use with a “distinct purpose” is often consistent with the goals of copyright because it encourages the development of new expression “without diminishing the incentive to create.” *Id.* at 531. On the other hand, a secondary use with the same purpose as the original work is “more likely to provide the public with a substantial substitute for” the original. *Id.* at 531–32 (cleaned up).

This factor favors Meta. There is no serious question that Meta’s use of the plaintiffs’ books had a “further purpose” and “different character” than the books—that it was highly transformative. The purpose of Meta’s copying was to train its LLMs, which are innovative tools that can be used to generate diverse text and perform a wide range of functions. *Cf. Oracle*, 593 U.S. at 30 (transformative to use copyrighted computer code “to create a new platform that could be readily used by programmers”). Users can ask Llama to edit an email they have written, translate an excerpt from or into a foreign language, write a skit based on a hypothetical scenario, or do any number of other tasks. The purpose of the plaintiffs’ books, by contrast, is to be read for entertainment or education.

The plaintiffs do not meaningfully disagree about Llama’s purpose. To the contrary, they acknowledge that LLMs have “end uses” including serving “as a personal tutor,” assisting “with creative ideation,” and helping users “generate business reports.” And several of the plaintiffs testified to using LLMs for various purposes, all distinct from creating or reading an expressive work like a novel or biography—for instance, to find recipes, get tax or medical advice, translate documents, or conduct research. All of these functions are different from the use to which the plaintiffs’ books are generally put. So copying the books to develop a tool that can perform those functions is a use with a different purpose and character than the books themselves.

## A

The plaintiffs’ law professor *amici* argue that Meta’s use has the same purpose and character as the books because an LLM training on a book is akin to a human reading one. One might also analogize Meta’s copying of the books to train Llama to a situation in which a



professor copies a book and gives it to a student so that the student can use the knowledge from the book (along with knowledge they get from other books) to go do great things. But there are a few important differences.

First, an LLM’s consumption of a book is different than a person’s. An LLM ingests text to learn “statistical patterns” of how words are used together in different contexts. It does so by taking a piece of text from its training data, removing a word from that text, predicting what that word will be, and updating its general understanding of language based on whether it was right or wrong—and then repeating this exercise billions or trillions of times with different text. This is not how a human reads a book.

Second, unlike the hypothetical professor, Meta did not just give the plaintiffs’ books to one person. Meta copied the plaintiffs’ books as part of an effort to create a tool that can generate a wide range of text. Any person can use that tool to help them create further expression, whether by having it help them brainstorm or research for a creative writing project (like plaintiff David Henry Hwang, a playwright and screenwriter) or by having it write code to develop new software programs (like Lockheed Martin). By creating a tool that anyone can use, Meta’s copying has the potential to exponentially multiply creative expression in a way that teaching individual people does not. *Cf. Oracle*, 593 U.S. at 30.

In contrast to the copyright professors, the plaintiffs make different (and much weaker) arguments for why Meta’s use is not transformative. For example, the plaintiffs suggest that Llama has “no critical bearing” on their books, the way criticism or parody would. But “critique or commentary on the original” are not “the only uses that will furnish a justification ultimately qualifying as fair use.” *Romanova*, 138 F.4th at 115. To the contrary, a use that enables “the furnishing of valuable information on any subject of public interest” or renders “a valuable service to the public” might be justified, especially where that benefit is “provided without allowing public access to the copy.” *Id.*

In addition, the plaintiffs argue that Meta’s use is not transformative because Llama will output material that “mimics” the plaintiffs’ work or writing styles if prompted to do so.

Therefore, the plaintiffs say, Meta’s use “merely amounts to a ‘repackaging’” of their books. The plaintiffs point to evidence that they say shows that Meta trained Llama to be able to emulate certain writers’ styles. Pls. Reply Exs. 111–14. But this evidence does not show that Meta trained Llama to repackaging the plaintiffs’ *works*. To the contrary, as noted above, even using “adversarial” prompts designed to get Llama to regurgitate its training data, Llama will not produce more than 50 words of any of the plaintiffs’ books. Pls. Reply Ex. 79 ¶¶ 79, 82–83, 92. And there is no indication that it will generate longer portions of text that would function as “repackaging” of those books. Nor is there even any indication that, as the plaintiffs’ *amici* claim, Meta developed Llama with the purpose of enabling it to create books that compete with the plaintiffs’ (without rising to the level of repackaging them).<sup>5</sup> So at most, this evidence shows that Meta wanted Llama to be able to generate text in certain styles. But style is not copyrightable—only expression is. *See* 17 U.S.C. § 102(b); *cf. Mattel, Inc. v. MGA Entertainment, Inc.*, 616 F.3d 904, 916 (9th Cir. 2010). Even if one possible use of Llama is to generate text with similarities to unprotectable aspects of the plaintiffs’ books, that does not mean Meta’s copying had the same purpose as those books.<sup>6</sup>

## B

As noted earlier, whether the secondary use is transformative doesn’t dictate the outcome of the first factor analysis (let alone of the entire fair use inquiry). Also relevant is the commercial nature of Meta’s use. Although Llama is available under a free license, it was ultimately developed for commercial reasons, and Meta expects it to generate 460 billion to 1.4 trillion dollars in revenue over the next ten years. Pls. MSJ Ex. 8 at 2. That a use is commercial

---

<sup>5</sup> This sort of competition—from AI-generated books that are like the plaintiffs’ but not similar enough to be infringing—is also discussed at length with respect to the fourth factor.

<sup>6</sup> By contrast, consider an LLM that was designed to be used to create works substantially similar to those on which it was trained, or to create works that competed with the originals without being substantially similar. Using copyrighted works to train such an LLM could be less transformative than using them to train a general-purpose LLM, because that use would have the purpose and character of enabling an LLM to develop substitute works. That said, even then, training the LLM would still likely be at least somewhat transformative; transformativeness isn’t an on-off switch.

“tends to weigh against a finding of fair use” because, all else equal, commercial copying is less justified than noncommercial copying. *Warhol*, 598 U.S. at 537 & n.13 (quoting *Harper & Row*, 471 U.S. at 562). So the fact that Llama may make Meta many billions of dollars is relevant and shouldn’t be completely brushed aside, as Meta tries to do. As discussed later, if copying would result in market harm to the protected works, it could matter a great deal whether the copying was part of a for-profit endeavor as opposed to, say, an academic endeavor. Nevertheless, commercialism isn’t dispositive of the first factor and tends to be less important when the secondary use is highly transformative. *See Oracle*, 593 U.S. at 32; *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2003). Thus, while the profit Meta stands to gain from its development of a product trained on the plaintiffs’ works is relevant to the fair use analysis overall, it does not tilt the first factor in the plaintiffs’ favor.

The same is true of the manner in which Meta acquired the plaintiffs’ books. The plaintiffs are wrong that the fact that Meta downloaded the books from shadow libraries and did not start with an “authorized copy” of each book gives them an automatic win. To say that Meta’s downloading was “piracy” and thus cannot be fair use begs the question because the whole point of fair use analysis is to determine whether a given act of copying was unlawful. *See generally* Amicus Br. of Electronic Frontier Foundation. Although the Federal Circuit once suggested the contrary in *Atari Games Corp. v. Nintendo of America Inc.*, 975 F.2d 832, 843 (Fed. Cir. 1992), that opinion overread the cases on which it relied for its statement about the need to start with an authorized copy, *see, e.g., Religious Technology Center v. Netcom On-Line Communication Services, Inc.*, 923 F. Supp. 1231, 1244 n.14 (N.D. Cal. 1995) (discussing *Atari*’s reasoning and concluding that the cases it relied on misread *Harper & Row*).

But Meta is also wrong to suggest that its use of shadow libraries is irrelevant to whether its copying was fair use. It’s relevant—or at least potentially relevant—in a few different ways.

First, Meta’s use of shadow libraries is relevant to the issue of bad faith, which is “often taken up under the first factor.” *Oracle*, 593 U.S. at 32. The law is in flux about whether bad faith is relevant to fair use. *Compare, e.g., id.* at 32 (noting that “skepticism about whether bad

faith has any role in a fair use analysis” is “justifiable”), with *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1164 n.8 (9th Cir. 2007) (“[A] party claiming fair use must act in a manner generally compatible with principles of good faith and fair dealing.”), and *Triller Fight Club II LLC v. H3 Podcast*, No. CV21-3942, 2023 WL 11877604, at \*8 (C.D. Cal. Sep. 15, 2023) (determining that *Perfect 10*’s language regarding good faith was still binding despite *Oracle*’s “skepticism”). It seems like good faith versus bad faith shouldn’t be especially relevant: The purpose of fair use is to allow new expression that won’t substitute for the original work, and whether a given use was made in good or bad faith wouldn’t seem to affect the likelihood of that use substituting for the original.<sup>7</sup> But even if bad faith is relevant, it doesn’t move the needle here, given the rest of the summary judgment record. *See Oracle*, 593 U.S. at 33 (describing bad faith as a “factbound consideration” that was “not determinative” in that case).

Second, downloading copyrighted material from shadow libraries would be relevant if it benefitted those who created the libraries and thus supported and perpetuated their unauthorized copying and distribution of copyrighted works. In the vast majority of cases, this sort of peer-to-peer file-sharing will constitute copyright infringement. Some of the libraries Meta used have themselves been found liable for infringement. *See, e.g., Elsevier Inc. v. Sci-Hub*, No. 15-cv-4282, 2017 WL 3868800, at \*1–2 (S.D.N.Y. June 21, 2017) (entering default judgment and finding that LibGen was liable for willful copyright infringement). Some of their operators have even been indicted for criminal copyright infringement. *See* Indictment, *United States v. Napolsky*, No. 22-cr-525 (E.D.N.Y. Nov. 16, 2022), Dkt. No. 4 (indictment against founders of

---

<sup>7</sup> It may seem inconsistent to say that commercialism is relevant but bad faith shouldn’t be. After all, commercial uses can still entail the creation of new expression for public consumption. The difference between commercialism and good faith, however, is that the former has to do with the secondary use while the latter mostly has to do with the secondary user. The goal of copyright is to encourage “activity that is useful to the public education.” Pierre N. Leval, *Toward a Fair Use Standard*, 103 Harv. L. Rev. 1105, 1126 (1990). Although not dispositive, commercialism is relevant because nonprofit uses are (at least theoretically) more likely to be aimed at benefiting the public than are for-profit uses. *Cf.* 17 U.S.C. § 107(1) (juxtaposing “commercial nature” with “nonprofit educational purposes”); *Sony*, 464 U.S. at 448–51. Good faith, by contrast, focuses on “the morality of the secondary user”—not on “whether her *creation* . . . is of the type” that benefits the public and thus should be protected by copyright law. Leval, 103 Harv. L. Rev. at 1126.

Z-Library). So if Meta’s act of downloading propped up these libraries or perpetuated their unlawful activities—for instance, if they got ad revenue from Meta’s visits to their websites—then that could affect the “character” of Meta’s use. But the plaintiffs have not submitted any evidence about this. In any event, because any such effects would be even more relevant to the fourth factor (insofar as they could contribute to use of the libraries and thus infringement by others), this issue is discussed below as part of the analysis of that factor.<sup>8</sup>

### C

The last issue relating to the character of Meta’s use (and thus the first factor) is the relationship between Meta’s downloading of the plaintiffs’ books and Meta’s use of the books to train Llama. To the extent the plaintiffs suggest that the former must be considered wholly separately from the latter, they are wrong. To be sure, Meta’s downloading is a different use from any copying done in the course of LLM training. But that downloading must still be considered in light of its ultimate, highly transformative purpose: training Llama. *See Authors Guild v. Google, Inc. (Google Books)*, 804 F.3d 202, 216–18 (2d Cir. 2015) (considering the creation of digital copies of books in light of the secondary user’s overall purpose of creating a searchable database); *cf. Warhol*, 598 U.S. at 533 (noting that different uses must be considered separately, but that “the same copying may be fair when used for one purpose but not another”); *contra* Order on Fair Use at 18, *Bartz*, No. 24-cv-5417. Because Meta’s ultimate use of the plaintiffs’ books was transformative, so too was Meta’s downloading of those books.

The plaintiffs also assert that Meta downloaded multiple copies of the databases containing their books, and that only some of these copies were used for LLM training, so the downloading of the ones that were not used for training cannot be fair use. But all of the

---

<sup>8</sup> Meta’s use of shadow libraries is also clearly relevant to the plaintiffs’ distribution claim. But as discussed above, reproduction and distribution present separate issues. So even if Meta’s torrenting from shadow libraries did entail distribution, that wouldn’t be dispositive of whether its reproduction was fair use.

Separately, if Meta’s downloading materially contributed to the shadow libraries’ own infringement, Meta could potentially be liable as a contributory infringer. *See Perfect 10*, 508 F.3d at 1170–72. But the plaintiffs did not bring a contributory infringement claim or develop any evidence in support of one.

downloads the plaintiffs identify had the ultimate purpose of LLM training. The plaintiffs say that Meta only used its initial October 2022 download of LibGen to see whether the books in the database made for good training data. Pls. Reply at 12. This is a reasonable first step towards training an LLM. *See* Pls. MSJ Ex. 32 at 3. The plaintiffs say that Meta cross-referenced its next download of LibGen and its first download of Anna’s Archive with publisher catalogues to see whether it was still worth pursuing licensing (or whether all the books available for licensing were already included in those databases). But the plaintiffs concede that these downloads were also used as training data. *See* Pls. Reply at 13–14. And there is no indication that comparing the books in those databases to the books in another entailed any additional copying. So that cross-referencing alone cannot create infringement liability and does not need to separately constitute fair use. *Cf. Warhol*, 598 U.S. at 534 & n.10 (discussing application of fair use test to different uses).

Finally, the plaintiffs say that after Meta abandoned licensing and decided to use the books it downloaded from shadow libraries as training data, it also downloaded several other “copies of pirated datasets, only some of which ever made it into an LLM for training.” But they provide no evidence that this was actually the case. They point only to deposition testimony in which a Meta employee said she didn’t know whether Meta used every LibGen copy it downloaded for training. Pls. Reply Ex. 109 at 66:17–20. Two other Meta AI employees, meanwhile, said that they weren’t aware of any downloads that weren’t used as training data or for related efforts like the experiments mentioned above. Pineau Decl. ISO Meta Reply ¶ 6; Kambadur Decl. ISO Meta Reply ¶ 7.<sup>9</sup> In any event, even if Meta did download some copies that weren’t ultimately used for training, fair use doesn’t require that the secondary user make the lowest number of copies possible. *Cf. Sony Computer Entertainment, Inc. v. Connectix Corp.*, 203 F.3d 596, 601, 605 (9th Cir. 2000).

---

<sup>9</sup> The plaintiffs’ objections to these declarations are overruled because a party may attach to a reply brief declarations that are a “reasonable response to the opposition,” and these declarations were not inconsistent with the declarants’ deposition testimony. *Hodges v. Hertz Corp.*, 351 F. Supp. 3d 1227, 1249 (N.D. Cal. 2018); *see also* Civil L.R. 7-3(c).

#### IV. FACTOR TWO: THE NATURE OF THE COPYRIGHTED WORK

The second factor recognizes that “some works are closer to the core of intended copyright protection than others, with the consequence that fair use is more difficult to establish when the former works are copied.” *Campbell*, 510 U.S. at 586. Works receiving greater copyright protection include creative ones like books and movies; works receiving lesser protection include computer code. *Oracle*, 593 U.S. at 29.

This factor favors the plaintiffs. Their books—mostly novels, memoirs, and plays—are highly expressive works “of the type that the copyright laws value and seek to protect.” *Hachette*, 115 F.4th at 187 (quoting *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 98 (2d Cir. 2014)). That some of their works may be factual (like an autobiography) as opposed to fictional does not meaningfully change this conclusion, because copyright still protects an author’s “manner of expressing” facts. *Google Books*, 804 F.3d at 220.

Meta argues that this factor favors it anyway because Meta only used the plaintiffs’ books to gain access to their “functional elements,” not to capitalize on their creative expression. Meta primarily relies on two Ninth Circuit cases involving “intermediate copying.” In both of those cases, a video game company copied a video game console manufacturer’s copyrighted code and reverse-engineered it to understand certain functional elements of that code. This allowed the game companies to build their own products that would work with the plaintiffs’. In each case, the Ninth Circuit held that the defendant’s fair use defense would likely succeed because, although the defendants copied expressive elements of the plaintiffs’ code, they only did so to access the code’s unprotected, functional elements. *See Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520–26 (9th Cir. 1992); *Connectix*, 203 F.3d at 602.

But unlike the uses in those cases, Meta’s use of the plaintiffs’ books does depend on the books’ creative expression. As Meta itself notes, LLMs are trained through learning about “statistical relationships between words and concepts” and collecting “statistical data regarding word order, frequencies [what words are used and how often], grammar, and syntax.” Word order, word choice, grammar, and syntax are how people express their ideas. *See Harper & Row*,



471 U.S. at 548 (discussing how “ordering and choice of words” are expression under even a narrow interpretation of what counts as expression). So even though LLMs may only learn about “statistical relationships,” those relationships are the product of creative expression. This is true even though, as discussed earlier, Llama consumes that expression in a different way than a person would.

To support its argument that it copied the plaintiffs’ books to extract non-expressive information (such that the intermediate copying cases should apply), Meta cites *Google Books*. But that case is distinguishable. There, the plaintiffs were authors who alleged that Google committed copyright infringement by making digital copies of their books and creating a database that users could search to see what books in the database contained the search terms. 804 F.3d at 207–10. Unlike here, the technology at issue in *Google Books* was content agnostic: The database wouldn’t work any better or worse if it contained books full of complete gibberish or written in unknown languages. If someone searched for that text, those books would appear. Here, by contrast, if Meta’s LLMs are to generate high-quality text, they need coherent, reasonably high-quality training data. In other words, they need high-quality expression. Therefore, the “intermediate copying” cases don’t apply. *See Disney Enterprises, Inc. v. VidAngel, Inc.*, 869 F.3d 848, 862 n.12 (9th Cir. 2017).

The second factor, however, “has rarely played a significant role in the determination of a fair use dispute.” *Google Books*, 804 F.3d at 220. And it applies “with less force” when the copied works have already been published and the secondary user therefore cannot interfere with the creator’s right to control the first public appearance of their work. *VHT, Inc. v. Zillow Group, Inc.*, 918 F.3d 723, 744 (9th Cir. 2019) (quoting *Kelly*, 336 F.3d at 820). So the fact that the second factor favors the plaintiffs doesn’t mean much for the analysis as a whole.

## **V. FACTOR THREE: THE AMOUNT AND SUBSTANTIALITY OF THE PORTION USED IN RELATION TO THE COPYRIGHTED WORK AS A WHOLE**

This factor “asks whether ‘the amount and substantiality of the portion used’” are “reasonable in relation to the purpose of the copying.” *Campbell*, 510 U.S. at 586 (quoting 17



U.S.C. § 107(3)). This factor is therefore related to the first, because “the extent of permissible copying varies with the purpose and character of the use.” *Id.* at 586–87.

As an initial matter, the amount copied doesn’t seem especially relevant in this case. In a case involving, for instance, a musical parody, copying large portions of the original song might increase the parody’s “potential for market substitution.” *See id.* at 589. But given that Meta’s LLMs won’t output any meaningful amount of the plaintiffs’ books, it’s not clear how or why Meta’s copying would be less likely to lead to the creation of direct substitutes for the books if Meta had copied less of them. *Cf. Hachette*, 115 F.4th at 188–89 (“The relevant consideration . . . is not the amount of copyrighted material *used by the copier*, but ‘the amount of copyrighted material *made available to the public*.’” (quoting *Fox News Network, LLC v. TVEyes*, 883 F.3d 169, 179 (2d Cir. 2018))).

In any event, this factor favors Meta, even though it copied the plaintiffs’ books in their entirety. The amount that Meta copied was reasonable given its relationship to Meta’s transformative purpose. *See Oracle*, 593 U.S. at 34. Everyone agrees that LLMs work better if they are trained on more high-quality material. *See* Ungar Decl. ISO Meta MSJ ¶¶ 42–48; Pls. Reply Ex. 115 ¶¶ 79–80. So feeding a whole book to an LLM does more to train it than would feeding it only half of that book. With this in mind, it was “reasonably necessary” for Meta to “make use of the entirety of the works.” *HathiTrust*, 755 F.3d at 98.<sup>10</sup>

## **VI. FACTOR FOUR: THE EFFECT OF THE USE UPON THE POTENTIAL MARKET FOR OR VALUE OF THE COPYRIGHTED WORK**

This factor looks to both the “extent of market harm caused by the particular actions of the alleged infringer” and to “whether unrestricted and widespread conduct of the sort engaged in by the defendant . . . would result in a substantially adverse impact on the potential market’ for the original.” *Campbell*, 510 U.S. at 590 (quoting 3 M. Nimmer & D. Nimmer, *Nimmer on Copyright* § 13.05 (1993)). The “only harm” relevant to this factor “is the harm of market

---

<sup>10</sup> Meta’s argument here does cut against it on the fourth factor, however. As discussed below, an LLM trained on copyrighted books is more likely to be capable of generating books that can compete with the ones on which it was trained.

substitution.” *Id.* at 593. When, by contrast, the secondary work kills demand for the first through criticism or parody, the harm is not cognizable under the Copyright Act. *Id.* at 591–92. Also relevant to this factor are “the public benefits the copying will likely produce.” *Oracle*, 593 U.S. at 35.

As noted previously, the fourth factor is “undoubtedly the single most important element of fair use.” *Harper & Row*, 471 U.S. at 566. Meta is therefore wrong to suggest that, because the first factor strongly favors it, the inquiry should basically end there. To the contrary, given the fourth factor’s importance, it’s easy to imagine a situation in which a secondary use is highly transformative but the secondary user nonetheless loses on fair use because allowing people to engage in that kind of use would have too great an effect on the market for the original work. But by the same token, in a case where the first factor cuts strongly in favor of the defendant, generally the plaintiff’s only chance to defeat fair use will be to win decisively on factor four.

In a case involving the use of copyrighted works to train generative AI models, there are at least three ways a plaintiff might try to argue that the defendant’s copying harmed the market for the works (or that the market would be harmed if that copying were widespread). First, the plaintiff might claim that the model will regurgitate their works (or outputs that are substantially similar), thereby allowing users to access those works or substitutes for them for free via the model. Second, the plaintiff might point to the market for licensing their works for AI training and contend that unauthorized copying for training harms that market (or precludes the development of that market). Third, the plaintiff might argue that, even if the model can’t regurgitate their own works or generate substantially similar ones, it can generate works that are similar enough (in subject matter or genre) that they will compete with the originals and thereby indirectly substitute for them. In this case, the first two arguments fail. The third argument is far more promising, but the plaintiffs’ presentation is so weak that it does not move the needle, or even raise a dispute of fact sufficient to defeat summary judgment.

## A

If Llama could be used to generate significant portions of the plaintiffs’ books—or text so

similar to their books as to be infringing in its own right—that would threaten the market for the books because people would read those outputs instead. But that theory of harm is not viable in this particular case because, as discussed above, Llama does not allow users to generate any meaningful portion of the plaintiffs’ books. Neither party’s expert opined that Llama was able to regurgitate more than 50 words from any of the plaintiffs’ books, even in response to “adversarial” prompting designed specifically to make LLMs regurgitate. *See* Pls. Ex. 79 ¶¶ 71–72, 82–84, 92. And the plaintiffs’ expert conceded that Llama would not generate “any significant percentage” of their books. Meta MSJ Ex. 24 at 237:16–19. In *Google Books*, by way of comparison, the Second Circuit held that the secondary use did “not threaten the rights holders with any significant harm to the value of their copyrights or diminish their harvest of copyright revenue” despite allowing users to see snippets adding up to as much as 16% of a book.<sup>11</sup> 804 F.3d at 224. Llama’s ability to regurgitate miniscule portions of the plaintiffs’ books if manipulated into doing so does not threaten to have a “meaningful or significant effect ‘upon the potential market for or value of’” the plaintiffs’ books. *Id.* (quoting 17 U.S.C. § 107(4)).

## B

The plaintiffs’ primary theory of market harm is that Meta’s unauthorized use of their books for LLM training harms the market for licensing their books for that purpose. The plaintiffs devote nearly all of their discussion of the fourth factor to this theory. The parties therefore go back and forth at length about whether a market for licensing general trade books exists or is likely to develop.

But whether such a market exists or is likely to develop is irrelevant, because this market is not one that the plaintiffs are legally entitled to monopolize. In every fair use case, the

---

<sup>11</sup> As *Google Books* noted, it isn’t the case that allowing someone to see 16% of a book could *never* threaten substantial harm. The tool there allowed users to see snippets that “were usually not sequential but scattered randomly throughout the book.” 804 F.3d at 222. If it “could be used to reveal a coherent block amounting to 16% of a book, that would raise a very different question.” *Id.* at 223. A portion of a work that is small in quantitative terms may also still be significant if it is “the heart of” the work or otherwise qualitatively important. *Cf. Harper & Row*, 471 U.S. at 565 (quoting *Harper & Row, Publishers, Inc. v. Nation Enterprises*, 557 F. Supp. 1067, 1072 (S.D.N.Y. 1983)).

“plaintiff suffers a loss of a potential market if that potential [market] is defined as the theoretical market for licensing” the use at issue in the case. *Tresóna Multimedia, LLC v. Burbank High School Vocal Music Association*, 953 F.3d 638, 652 (9th Cir. 2020) (emphasis omitted) (quoting 4 Melville B. Nimmer & David Nimmer, *Nimmer on Copyright* § 13.05 (2019)). Therefore, to prevent the fourth factor analysis from becoming circular and favoring the rightsholder in every case, harm from the loss of fees paid to license a work for a transformative purpose is not cognizable. *Id.*; *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 614–15 (2d Cir. 2006); *see also Oracle*, 593 U.S. at 38 (“cautioning against the ‘danger of circularity’” (quoting 4 Nimmer § 13.05)).

### C

The third way that using copyrighted books to train an LLM might harm the market for those works is by helping to enable the rapid generation of countless works that compete with the originals, even if those works aren’t themselves infringing. Assume for this discussion that people can (or will soon be able to) use LLMs to generate massive amounts of text in significantly less time than it would take to write that text, and using a fraction of the creativity. People could thus use LLMs to create books and then sell them, competing with books written by human authors for sales and attention. Indeed, to some extent, this appears to be occurring already—one expert for the plaintiffs briefly discusses reports of AI-generated books “flooding Amazon.” Pls. MSJ Ex. 76 ¶ 199; *see id.* ¶¶ 193–207. People might even be motivated to make those books available for free, given how easily it will presumably be to prompt an LLM to create them. Harm from this form of competition is the harm of market dilution. Or as one commentator describes it, the harm of “indirect” substitution, rather than “direct” substitution (which would be the first form of harm described). *See Matthew Sag, Fairness and Fair Use in Generative AI*, 92 Fordham L. Rev. 1887, 1916–20 (2024).

Of course, not all copyrighted works would have their markets diluted equally by AI-generated competitors. It seems unlikely, for instance, that AI-generated books would meaningfully siphon sales away from well-known authors who sell books to people looking for

books by those particular authors. But it's easy to imagine that AI-generated books could successfully crowd out lesser-known works or works by up-and-coming authors. While AI-generated books probably wouldn't have much of an effect on the market for the works of Agatha Christie, they could very well prevent the next Agatha Christie from getting noticed or selling enough books to keep writing.<sup>12</sup>

This effect also seems likely to be more pronounced with respect to certain types of works. For instance, an AI model that can generate high-quality images at will might be expected to greatly affect the market for such images, diminishing the incentive for humans to create them. An LLM that could generate accurate information about current events might be expected to greatly harm the print news market. The market for certain nonfiction works—for example, books about how to take care of your garden—could be greatly diminished by the ability of LLMs to produce books on that topic. For fiction works, it might be more dependent on the author or the genre in which that author operates.

The difference might be in part because some works are relatively functional and generally less dependent on the author's creativity. When picking a news article, readers want something that will tell them about a current (or past) event clearly, accurately, and concisely. When picking a novel, by contrast, readers may care about a much longer list of characteristics. They may care, for instance, about tone, thematic depth, writing style, plot, or characters; they may want a book that contains a number of plot twists or depicts a certain type of character development. These elements of a novel depend greatly on the creativity of the author. While a news article is also a product of its author's creativity (especially with respect to things like

---

<sup>12</sup> To be clear, the point is not that authors are entitled to more or less copyright protection based on how famous or popular they are. *Cf. Warhol*, 598 U.S. at 544 & n.19. The point is that different works may have different markets that will be affected differently by floods of AI-generated competitors. *See, e.g., Cariou v. Prince*, 714 F.3d 694, 709 (2d Cir. 2013) (“Prince’s work appeals to an entirely different sort of collector than Cariou’s.”); *Andy Warhol Foundation for Visual Arts, Inc. v. Goldsmith*, 11 F.4th 26, 48 (2d Cir. 2021) (“We cannot . . . endorse the district court’s implicit rationale that the market for Warhol’s works is the market for ‘Warhols,’ . . . [but] we see no reason to disturb the district court’s overall conclusion that the two works occupy distinct markets[.]”), *aff’d*, *Warhol* 598 U.S. 508.

structure and diction), there are many more creative choices in the average novel than the average news article, and those creative choices are more important to the average novel's quality. Relatedly, one could imagine people caring more about whether a novel is AI-generated (as opposed to the product of human creativity) than whether a news article is AI-generated.<sup>13</sup>

It also should be noted that, when considering market dilution, the proper comparison isn't to a world with no LLMs, but to a world where LLMs weren't trained on copyrighted books. Perhaps an LLM trained only on public domain works could still be capable of quickly generating large numbers of books that could compete for sales with copyrighted books. But there is plenty of evidence in the record that training on books substantially benefits LLMs' creativity and ability to generate long pieces of text. *E.g.*, Pls. MSJ Ex. 25 at 2; *id.* Ex. 27 ¶ 183. And because LLMs perform better the more text they are trained on, an LLM trained only on public domain books would presumably, all else equal, lag significantly behind a book trained also on copyrighted ones. *See* Ungar Decl. ISO Meta MSJ ¶ 45. So training an LLM on copyrighted books would seem, in most circumstances, to make that LLM better able to generate works that could dilute the market for the books in its training data.

Meta and its law professor *amici*, as well as the Matthew Sag article cited above, argue that market dilution does not count under the fourth factor. They argue that harm caused by an LLM's outputs is only relevant if the outputs are themselves infringing—that is, if the LLM regurgitates copyrighted material (or generates text that is substantially similar to copyrighted material). *See* May 1 Hr'g Tr. at 22:7–24:21; 108–09; Amicus Br. of Intellectual Property Law Professors at 9–10; *see also* Sag, 92 Fordham L. Rev. at 1919–20. But that can't be right. To be sure, it would be easier to conclude that the market for copied books would be harmed by an LLM that is capable of regurgitating those books or generating substantially similar text. But less

---

<sup>13</sup> This is not to suggest that news articles or other works that may be less dependent on their author's creativity are thus less deserving of protection, or that it would therefore be more appropriate to use those works to train an LLM. To the contrary, as noted with respect to the second factor, nonfiction works are still protected by copyright because the law protects their authors' choices as to how to express facts. *See Google Books*, 804 F.3d at 220.

similar outputs, such as books on the same topics or in the same genres, can still compete for sales with the books in the training data. And by taking sales from those books, or by flooding stores and online marketplaces so that some of those books don't get noticed and purchased, those outputs would reduce the incentive for authors to create—the harm that copyright aims to prevent.

The Supreme Court has said that the “only harm” that matters under the fourth factor “is the harm of market substitution.” *Campbell*, 510 U.S. at 593. But indirect substitution is still substitution: 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. This is different from the (non-cognizable) harm caused by criticism or commentary, which can harm demand for an original work without serving as a replacement for it.

Relatedly, Meta argues that “legitimate” competition from noninfringing secondary works is not cognizable under the fourth factor. It cites the intermediate copying cases for this proposition. *See Sega*, 977 F.2d at 1523–24; *Connectix*, 203 F.3d at 607. But key to those cases’ reasoning was the fact that the secondary users’ competing products did not benefit from the creative expression in the works they copied. By contrast, as discussed, LLMs are better able to generate text (including competing works) because they are trained on the creative expression in copyrighted books. So this competition is not “legitimate” within the meaning of those cases.

It’s true that, in many copyright cases, this concept of market dilution or indirect substitution is not particularly important. That’s because, in a more typical case, an original work is being compared to a single secondary work. If the secondary work is somewhat similar, but not so similar as to effectively be a copy, it still might have a small indirect effect on the market for the original work. But that likely won’t matter. Recall that the fourth factor looks to whether “conduct of the sort engaged in by the defendant” would have a “*substantially* adverse impact on the potential market for the original.” *Campbell*, 510 U.S. at 590 (emphasis added) (quoting 3 Nimmer § 13.05). The existence of *some* harm from indirect substitution isn’t dispositive of the fourth factor or the fair use inquiry. Where, for instance, the first factor cuts in favor of the



secondary user, the law might tolerate a little bit of competition. *See Google Books*, 804 F.3d at 224. In cases involving a single secondary work that’s similar-but-not-too-similar, it’s unlikely that harm from market dilution would be significant enough to matter. Even considering the effect of “widespread conduct of the sort engaged in by the defendant,” *Oracle*, 593 U.S. at 38 (quoting 4 Nimmer § 13.05), creating one indirectly substitutional work at a time could only have so great an effect on the market for the original.

This case is different. This is not a case where an original work is being compared to one secondary work. Nor is this case like the previous fair use cases involving creation of a digital tool. In those cases, like *Google Books* and *Perfect 10*, the tool could at most be used to access part or all of the original works. This case, unlike any of those cases, involves a technology that can generate literally millions of secondary works, with a miniscule fraction of the time and creativity used to create the original works it was trained on. No other use—whether it’s the creation of a single secondary work or the creation of other digital tools—has anything near the potential to flood the market with competing works the way that LLM training does. And so the concept of market dilution becomes highly relevant.

In arguing that this sort of harm doesn’t count just because it’s never made a difference in a case before, Meta makes the mistake the Supreme Court instructs parties and courts to avoid: robotically applying concepts from previous cases without stepping back to consider context. Fair use is meant to be a flexible doctrine that takes account of “significant changes in technology.” *Oracle*, 593 U.S. at 19 (quoting *Sony*, 464 U.S. at 430). Courts can’t stick their heads in the sand to an obvious way that a new technology might severely harm the incentive to create, just because the issue has not come up before. Indeed, it seems likely that market dilution will often cause plaintiffs to decisively win the fourth factor—and thus win the fair use question overall—in cases like this.

But courts can’t decide cases based on what they think will or should happen in other cases. They must decide cases based on the arguments presented and the evidence submitted by the parties. The question, then, is whether these particular thirteen plaintiffs in this particular



case have presented enough evidence to win on this factor. Or, to put it more precisely given the procedural posture of this case, whether these plaintiffs have presented enough evidence to raise a genuine dispute of material fact sufficient to give the question of market dilution to a jury. The answer is no.

In their complaint, the plaintiffs asserted only two types of market harm—that users of Llama can reproduce text from their books, and that Meta’s copying harmed the market for licensing copyrighted materials to companies for AI training. As for market dilution—the notion that allowing companies like Meta to copy their works to train products like Llama would inevitably cause the market for the plaintiffs’ works to be flooded with similar works—the plaintiffs never so much as mentioned it in their complaint. Nor did they mention it in their own summary judgment motion.

Naturally, given the allegations in the complaint, Meta’s cross-motion for summary judgment focused on defeating the first two theories. But Meta also noted in its motion that the plaintiffs hadn’t presented any evidence that Meta’s use of their books to train Llama had harmed book sales. *See* Meta MSJ Exs. 8–9. And Meta presented its own expert testimony explaining that Llama 3’s release did not have any discernible effect on the plaintiffs’ sales (or those of other books in Llama’s training data), at least in the period shortly after the release. *See* Sinkinson Decl. ISO Meta MSJ ¶¶ 18–35.

In opposition, the plaintiffs’ primary response was that this was beside the point because of their first two theories. They did make fleeting reference to a report by one of their experts, who briefly discussed the concept of indirect substitution and mentioned articles discussing how AI-created books are starting to flood Amazon. *See* Pls. Reply Ex. 126 ¶¶ 193–207. But this discussion generates more questions than answers.

First, is Llama capable of generating such books? If it isn’t currently, will it be capable of doing so in the near future? Presumably the answer is yes, but that’s not a foregone conclusion. An LLM could, for instance, be configured to be unable to produce book-length or book-style outputs. So the fact that books are being created by *some* LLM does not automatically mean that

Llama can create them or will be able to do so soon.

Second, what are these AI-generated books? Do they compete with Sarah Silverman’s memoir? With plaintiff Matthew Klam’s book of short stories? With Rachel Louise Snyder’s nonfiction work on domestic violence? The plaintiffs provide no analysis of the markets for their books, no discussion of whether these markets are or could be affected by AI-generated books, and no explanation of whether the existing AI-generated books referenced in the expert report compete in these markets.

Third, what impact does this competition actually have on sales of the books it competes with? Does it drown out those books entirely? Does it just chisel at their sales at the margins? Or, as discussed above and seems likely, does it depend on the book—are readers of romance novels happy to buy AI-generated ones, while all the people who want to read Sarah Silverman’s memoir still want to read it over AI-generated comic memoirs? Whatever the effects have been thus far, are they likely to increase in the future, as more and more AI-generated books are written, and as LLMs get better and better at writing human-like text?

Fourth, how does the threat to the market for the plaintiffs’ books in a world where LLM developers can copy those books compare to the threat to the market for the plaintiffs’ books in a world where the developers can’t copy them? There is no hint of that in the briefs or evidence presented by the plaintiffs.

The analysis is complicated somewhat by the fact that fair use is an affirmative defense and that Meta moved for summary judgment on it. For those reasons, Meta had the burden of presenting evidence that its copying doesn’t threaten to substantially harm the market for the plaintiffs’ books. It didn’t conclusively establish that its copying couldn’t do so in the future—potentially because its copying did in fact make Llama better able to generate countless works that will dilute the market for the plaintiffs’ books. But where a defendant introduces evidence of a lack of market harm, “and the plaintiff fails to introduce empirical evidence countering such a showing, the fourth factor should be weighed in the defendant’s favor.” *Patry on Fair Use* § 6:13; *see also Seltzer v. Green Day, Inc.*, 725 F.3d 1170, 1179 (9th Cir. 2013); *cf. Perfect 10*,

508 F.3d at 1168. That is exactly what happened here. Meta introduced evidence that its copying hasn't caused market harm. The plaintiffs presented no empirical evidence to the contrary—no evidence that the copying has already caused market harm, and no evidence that the copying is likely to cause market harm in the future. All the plaintiffs presented is speculation, and speculation is insufficient to raise a genuine issue of fact and defeat summary judgment. *E.g.*, *Anheuser-Busch, Inc. v. Natural Beverage Distributors*, 69 F.3d 337, 345 (9th Cir. 1995).

The plaintiffs argue that they didn't need to present empirical evidence because market harm can be inferred. For this argument, they cite to *Hachette*, in which the Second Circuit inferred market harm—even though the plaintiffs had not provided “empirical data” showing any and the secondary user presented expert testimony that there was none—because it was “self-evident” that the secondary use would cause such harm if widespread. 115 F.4th at 192–93. In *Hachette*, the secondary user maintained a database that let internet users “download an identical copy of” the plaintiffs' books for free. *Id.* at 194. The secondary use therefore offered a directly “competing substitute” for the original books. *Id.* at 195.

While it made sense to infer market harm in *Hachette*, it doesn't make sense to do so here. First, the Supreme Court has stated that no “inference of market harm . . . is applicable to a case involving something beyond mere duplication for commercial purposes.” *Campbell*, 510 U.S. at 591. In *Hachette*, the secondary use was basically “mere duplication.” Here, by contrast, Meta's use is highly transformative and has a purpose well beyond that. Second, unlike in *Hachette*, Meta's use does not let users access any significant portion of the plaintiffs' books, so it isn't self-evident that Meta's use would create harm via direct substitution. Nor is it self-evident that Llama will harm the book sale market by enabling users to create a flood of competing books. It's possible, even likely, that Llama will harm the book sale market. But to conclude that it will requires inferring that Llama (and not just any LLM) can be used to create such books, that it will be used to create such books, that consumers will purchase those books instead of books written by human authors, that consumers will buy those books instead of the plaintiffs' books in particular, and that Llama is meaningfully better at creating those books

because it was trained on copyrighted material. In *Hachette*, on the other hand, the only necessary inference was that readers might choose to download the plaintiffs' books for free instead of paying for them—a much shorter (and more obvious) inferential leap. *Cf. American Society for Testing & Materials v. Public.Resource.Org*, 82 F.4th 1262, 1271–72 (D.C. Cir. 2023).

On this record, then, Meta has defeated the plaintiffs' half-hearted argument that its copying causes or threatens significant market harm. That conclusion may be in significant tension with reality, but it's dictated by the choice the plaintiffs made to put forward two flawed theories of market harm while failing to present meaningful evidence on the effect of training LLMs like Llama with their books on the market for those books.<sup>14</sup>

## D

Two other issues are relevant to the fourth factor. First, as noted above, is whether Meta's use of shadow libraries benefited those libraries or their other users. If it did, then this would be relevant to the fourth factor. It would mean that Meta's copying helped others acquire copyrighted works, potentially including the plaintiffs' works, without paying for them (and without any indication that those other people were acquiring the works for fair use purposes). But although the plaintiffs discussed Meta's use of shadow libraries at length, they did not argue that it had these effects or was relevant to the fourth factor beyond allowing Meta to get the books without paying. At the hearing, the plaintiffs' counsel did suggest that, by using shadow libraries, Meta (and other companies like it) would reduce the stigma associated with shadow

---

<sup>14</sup> The plaintiffs also assert that the market for their works was harmed in the more narrow sense that, if Meta had not downloaded the books from a shadow library, it would have been required to buy the books. But as already discussed, even though that downloading is a separate use, it must be considered in light of its overall purpose. For instance, imagine a researcher who downloaded books from a shadow library in the process of writing an article on shadow libraries, and only did so for their research. That downloading would almost certainly be a fair use. Of course, in that example, the downloader has less ability to procure the books elsewhere than Meta did. But the point is that downloading from a shadow library, which the plaintiffs refer to as “unmitigated piracy,” must be viewed in light of its ultimate end. Because Meta's purpose of LLM training is so transformative, the plaintiffs needed to win decisively on the fourth factor. The loss of isolated sales to AI developers is not the kind of market harm that could tip the scales for the plaintiffs.

libraries and encourage more people to use them. May 1 Hr’g Tr. at 92–93. It’s not clear whether this would matter in the overall analysis. But in any event, counsel conceded that the record contains no evidence of this dynamic playing out. *Id.* at 93–94.<sup>15</sup>

Second is the public benefit associated with Meta’s copying. Neither side’s presentation on this front does much to move the needle. The plaintiffs say that sanctioning Meta’s conduct would encourage piracy by incentivizing other LLM companies to pirate and to “support and defend” shadow libraries “that make stolen works available for free.” There is no evidence in the record that Meta (or any other LLM developer) is actively supporting or otherwise encouraging widespread use of shadow libraries. As for incentivizing other LLM developers to use shadow libraries, the plaintiffs again beg the question—whether LLM developers should have to pay for the books they use as training data is the issue addressed in this opinion (and, obviously, a fact-specific one that can’t be answered uniformly across the board). Meta, for its part, mostly discusses the various ways that LLMs can be useful. But the public benefits most relevant to the fourth factor are those “related to copyright’s concern for the creative production of new expression.” *Oracle*, 593 U.S. at 35. So the fact that Llama can help someone do their taxes, for example, is not especially relevant here. Nevertheless, Meta’s use of copyrighted works as training data will likely help Llama create new expression, whether by making it better at helping users generate creative text or by improving its “memory” and thereby making it more useful to the researchers who use it to develop software. Public benefit considerations thus slightly favor

---

<sup>15</sup> One of Meta’s expert witnesses did testify at her deposition that, depending on how it configured its torrenting software, it was more likely than not that Meta contributed to the BitTorrent network’s “bandwidth, content, storage, and processing power.” Pls. MSJ Ex. 67 at 103:3–104:5. But there is no evidence of whether Meta actually had the right settings to do so or of how much it might have contributed to this network. More importantly, there is no indication that any computing power Meta contributed to the *BitTorrent network* would have assisted the *shadow libraries* from which Meta torrented (or otherwise contributed to infringement of the plaintiffs’ copyrights). To the contrary, the plaintiffs cite a source indicating that the vast majority of torrented files are movies, TV shows, video games, and music—which are generally copyrighted but are not at issue in this case—and that books comprise less than one percent of torrented material. Jacqui Cheng, *BitTorrent Census: About 99% of Files Copyright Infringing*, *Ars Technica* (Jan. 29, 2010), <https://arstechnica.com/information-technology/2010/01/bittorrent-census-about-99-of-files-copyright-infringing> [<https://perma.cc/KZ7N-R9BN>].

Meta, confirming that it wins on factor four.

## E

Relatedly, Meta argues that the “public interest” would be “badly disserved” by preventing Meta (and other AI developers) from using copyrighted text as training data without paying to do so. Meta seems to imply that such a ruling would stop the development of LLMs and other generative AI technologies in its tracks. This is nonsense.

As mentioned earlier, a ruling that certain copying isn’t fair use doesn’t necessarily mean the copier has to stop their copying—it means that they have to get permission for it. So where copying for LLM training isn’t fair use, LLM developers (including Meta) won’t need to stop using copyrighted works to train their models. They will need only to pay rightsholders for licenses for that training.

Presumably, where copying for AI training isn’t fair use, AI developers will simply figure out a way to license the works they wish to use as training data. Meta’s contention that markets for this licensing can’t or won’t develop is hard to believe. If books are as good for LLM training as Meta says they are, then it seems nearly certain that LLM developers would be willing to pay for licenses. (Indeed, Meta itself was willing to pay to license books—it just found licensing too logistically difficult.) Even if the value of any particular book as training data is too low to justify negotiating licensing deals book by book, LLM developers would still presumably be interested in licensing large numbers of books at once. Publishers may not currently hold the subsidiary rights necessary to make group licensing possible. But it’s hard to believe that they won’t soon start negotiating those rights with their authors so that they can engage in large-scale negotiation and licensing with LLM developers—assuming they haven’t already started to do so. It seems especially likely that these licensing markets will arise if LLM developers’ only choices are to get licenses or forgo the use of copyrighted books as training data. If they instead choose to use only public domain works as training data (instead of licensing copyrighted works), that would indicate that they don’t actually need the copyrighted works as badly as they say they do.

So if it isn’t fair use for Meta and other LLM developers to use copyrighted books as

training data without permission, they won't have to stop working on their LLMs altogether. They'll just have to pay for licenses or use books that aren't copyrighted. Either way, it may be that LLM companies move somewhat more slowly or make somewhat less money. But the suggestion that the growth of LLM technology would come to a halt (or anything close) doesn't pass the straight face test.

## VII. CONCLUSION

Fair use is a fact-specific doctrine that requires case-by-case analysis that is sensitive to new technologies and their potential consequences. No previous case has involved a use that is both as transformative and as capable of diluting the market for the original works as LLM training is. So no previous case answers the question whether Meta's copying was fair use. That question must be answered by flexibly applying the fair use factors and considering Meta's copying in light of the purpose of copyright and fair use: protecting the incentive to create by preventing copiers from creating works that substitute for the originals in the marketplace.


In cases involving uses like Meta's, it seems like the plaintiffs will often win, at least where those cases have better-developed records on the market effects of the defendant's use. No matter how transformative LLM training may be, it's hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books. And some cases might present even stronger arguments against fair use. For instance, as discussed above, it seems that markets for certain types of works (like news articles) might be even more vulnerable to indirect competition from AI outputs. On the other hand, though, tweak some facts and defendants might win. For example, using copyrighted books to train an LLM for nonprofit purposes, like national security or medical research, might be fair use even in the face of some amount of market dilution. *See Oracle*, 593 U.S. at 32 (“[A] finding that copying was not commercial in nature tips the scales in favor of fair use.”). Or plaintiffs whose works are unlikely to face meaningful competition from AI-generated ones may be unable to defeat a fair use defense.

In this case, because Meta’s use of the works of these thirteen authors is highly transformative, the plaintiffs needed to win decisively on the fourth factor to win on fair use. *See, e.g., Perfect 10*, 508 F.3d at 1168 (fair use where secondary use was “significant[ly] transformative” and fourth factor “favor[ed] neither party”). And to stave off summary judgment, they needed to create a genuine issue of material fact as to that factor. Because the issue of market dilution is so important in this context, had the plaintiffs presented any evidence that a jury could use to find in their favor on the issue, factor four would have needed to go to a jury. Or perhaps the plaintiffs could even have made a strong enough showing to win on the fair use issue at summary judgment. But the plaintiffs presented no meaningful evidence on market dilution at all. Absent such evidence and in light of Meta’s evidence, the fourth factor can only favor Meta. Therefore, on this record, Meta is entitled to summary judgment on its fair use defense to the claim that copying these plaintiffs’ books for use as LLM training data was infringement.

As previously noted, summary judgment will be granted for Meta in a separate ruling on the plaintiffs’ DMCA claim. A Zoom case management conference is scheduled for July 11, 2025, at 10:00 a.m. to discuss how to proceed on the plaintiffs’ separate claim that Meta unlawfully distributed their protected works during the torrenting process.

**IT IS SO ORDERED.**

Dated: June 25, 2025

  
\_\_\_\_\_  
VINCE CHHABRIA  
United States District Judge