United States District Court
Northern District of California

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ANDREA BARTZ, CHARLES GRAEBER, and KIRK WALLACE JOHNSON,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

No. C 24-05417 WHA

ORDER ON FAIR USE

INTRODUCTION

An artificial intelligence firm downloaded for free millions of copyrighted books in digital form from pirate sites on the internet. The firm also purchased copyrighted books (some overlapping with those acquired from the pirate sites), tore off the bindings, scanned every page, and stored them in digitized, searchable files. All the foregoing was done to amass a central library of "all the books in the world" to retain "forever." From this central library, the AI firm selected various sets and subsets of digitized books to train various large language models under development to power its AI services. Some of these books were written by plaintiff authors, who now sue for copyright infringement. On summary judgment, the issue is the extent to which any of the uses of the works in question qualify as "fair uses" under Section 107 of the Copyright Act.

STATEMENT

Defendant Anthropic PBC is an AI software firm founded by former OpenAI employees in January 2021. Its core offering is an AI software service called Claude. When a user

prompts Claude with text, Claude quickly responds with text — mimicking human reading and writing. Claude can do so because Anthropic trained Claude — or rather trained large language models or LLMs underlying various versions of Claude — using books and other texts selected from a central library Anthropic had assembled. Claude was first released publicly in March 2023. Seven successive versions of Claude have been released since. Users may ask Claude some questions for free. Demanding users and corporate clients pay to use Claude, generating over one billion dollars in annual revenue (Opp. Exh. 18).

Plaintiffs Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson are authors of books that Anthropic copied from pirated and purchased sources. Anthropic assembled these copies into a central library of its own, copied further various sets and subsets of those library copies to include in various "data mixes," and used these mixes to train various LLMs. Anthropic kept the library copies in place as a permanent, general-purpose resource even after deciding it would not use certain copies to train LLMs or would never use them again to do so. All of Anthropic's copying was without plaintiffs' authorization.

Author Bartz wrote four novels Anthropic copied and used: *The Lost Night: A Novel*, *The Herd*, *We Were Never Here*, and *The Spare Room*. Author Graeber wrote two non-fiction books likewise at issue: *The Good Nurse: A True Story of Medicine, Madness, and Murder*, and *The Breakthrough: Immunotherapy and the Race to Cure Cancer*. And, Author Johnson penned three non-fiction books also copied and used: *To Be A Friend Is Fatal: The Fight to Save the Iraqis America Left Behind*, *The Feather Thief: Beauty, Obsession, and the Natural History Heist of the Century*, and *The Fishermen and the Dragon: Fear, Greed, and a Fight for Justice on the Gulf Coast*. Plaintiffs Bartz Inc. and MJ + KJ Inc. are corporate entities that Author Bartz and Author Johnson respectively set up to market their works. Between them, these five plaintiffs ("Authors") own all the copyrights in the above-listed works.

From the start, Anthropic "ha[d] many places from which" it could have purchased books, but it preferred to steal them to avoid "legal/practice/business slog," as cofounder and chief executive officer Dario Amodei put it (*see* Opp. Exh. 27). So, in January or February 2021, another Anthropic cofounder, Ben Mann, downloaded Books3, an online library of

196,640 books that he knew had been assembled from unauthorized copies of copyrighted books — that is, pirated. Anthropic's next pirated acquisitions involved downloading distributed, reshared copies of other pirate libraries. In June 2021, Mann downloaded in this way at least five million copies of books from Library Genesis, or LibGen, which he knew had been pirated. And, in July 2022, Anthropic likewise downloaded at least two million copies of books from the Pirate Library Mirror, or PiLiMi, which Anthropic knew had been pirated (Opp. Exh. 6 at 4; Opp. Expert Zhao ¶¶ 17–29; *see* Class Cert. ("CC") Opp. Expert Iyyer ¶¶ 45–46). Although what was downloaded and later duplicated from these sources was sometimes referred to as data or datasets, at bottom they contained full-text "ebooks or scans of books" saved in individual files in formats like .pdf, .txt, and .epub (*see, e.g.*, Opp. Exh. 12 at -0391318). For Books3, most filenames identified the book inside. For LibGen and PiLiMi, Anthropic downloaded a separate catalog of bibliographic metadata for each collection, with fields like title, author, and ISBN (*see, e.g.*, *ibid.*; Opp. Exh. 16 -0533972–73). Anthropic thereby pirated over seven million copies of books, including copies of at least two works at issue for each Author.[1]

As Anthropic trained successive LLMs, it became convinced that using books was the most cost-effective means to achieve a world-class LLM. During this time, however, Anthropic became "not so gung ho about" training on pirated books "for legal reasons" (Opp. Exh. 19). It kept them anyway (*e.g.*, Opp. Exh. 17 at 93–94; CC Opp. Exh. 35 at -0273474).

To find a new way to get books, in February 2024, Anthropic hired the former head of partnerships for Google's book-scanning project, Tom Turvey. He was tasked with obtaining "all the books in the world" while still avoiding as much "legal/practice/business slog" as

[1] Specifically, those works were (*see* Opp. Expert Zhao ¶ 36; CC Br. Expert Zhao ¶ 66):
- Author Bartz's *The Herd* (five copies total) (in LibGen and PiLiMi);
- Author Bartz's *The Lost Night* (three copies total) (in Books3, LibGen, and PiLiMi);
- Author Graeber's *The Breakthrough* (four copies) (in Books3, LibGen, and PiLiMi);
- Author Graeber's *The Good Nurse* (five copies total) (in Books3 and LibGen);
- Author Johnson's *To Be A Friend Is Fatal* (one copy) (in Books3); and
- Author Johnson's *The Feather Thief* (four copies total) (in Books3, LibGen, PiLiMi).

Some evidence suggests Anthropic downloaded still more copies before culling empty files, duplicates, and so on to reach the numbers kept in the central library and counted here.

possible (Opp. Exhs. 21, 27). So, in spring 2024, Turvey sent an email or two to major publishers to inquire into licensing books for training AI. Had Turvey kept up those conversations, he might have reached agreements to license copies for AI training from publishers — just as another major technology company soon did with one major publisher (*e.g.*, Opp. Expert Malackowski ¶¶ 50, 64). But Turvey let those conversations wither.

Instead, Turvey and his team emailed major book distributors and retailers about bulk-purchasing their print copies for the AI firm's "research library" (Opp. Exh. 22 at 145; Opp. Exh. 31 at -035589). Anthropic spent many millions of dollars to purchase millions of print books, often in used condition. Then, its service providers stripped the books from their bindings, cut their pages to size, and scanned the books into digital form — discarding the paper originals. Each print book resulted in a PDF copy containing images of the scanned pages with machine-readable text (including front and back cover scans for softcover books). Anthropic created its own catalog of bibliographic metadata for the books it was acquiring. It acquired copies of millions of books, including of all works at issue for all Authors.[2]

Anthropic may have copied portions of Authors' books on other occasions, too — such as while copying book reviews, academic papers, internet blogposts, or the like for its central library. And, Anthropic's scanning service providers may have copied Authors' print books along the way to delivering the final digital copies to Anthropic. But neither side here specifically raises legal issues implicated by any such copies. Nor will this order.

From all the above sources, Anthropic created a general "research library" or "generalized data area." What was this for? As Turvey said, this was a "way of creating information that would be voluminous and that we would use for research," or otherwise to

---

[2] In other words, within the scanned books were one or more copies of the following works:
- Author Bartz's *The Herd*;
- Author Bartz's *The Lost Night*;
- Author Bartz's *We Were Never Here*;
- Author Bartz's *The Spare Room*;
- Author Graeber's *The Breakthrough*;
- Author Graeber's *The Good Nurse*;
- Author Johnson's *To Be A Friend Is Fatal*;
- Author Johnson's *The Feather Thief*; and,
- Author Johnson's *The Fishermen*.

"inform our — our products" (Opp. Exh. 22 at 145–46, 194). The copies were kept in the original "version of the underlying" book files Anthropic had "obtained or created," that is, pirated or scanned (Opp. Exh. 30 at 3, 4). Anthropic planned to "store everything forever; we might separate out books into categories[, but t]here [wa]s no compelling reason to delete a book" — even if not used for training LLMs. Over time, Anthropic invested in building more tools for searching its "general purpose" library and for accessing books or sets of books for further uses (*see* CC Br. Exh. 12 at -0144509; CC Reply Exh. 45 at -0365931–32, -0365939–42 (reviewing and seeking to improve "[w]hat [ ] researchers do today if they want to search for a book," including improving bibliographic metadata and consolidating varied resources)).

One further use was training LLMs. As a preliminary step towards training, engineers browsed books and bibliographic metadata to learn what languages the books were written in, what subjects they concerned, whether they were by famous authors or not, and so on — sometimes by "open[ing] any of the books" and sometimes using software. From the library copies, engineers copied the sets or subsets of books they believed best for training and "iterate[d]" on those selections over time. For instance, two different subsets of print-sourced books were included in "data mixes" for training two different LLMs. Each was just a fraction of all the print-sourced books. Similarly, different sets or "subsets" or "parts of" or "portions" of the collections sourced from Books3, LibGen, and PiLiMi were used to train different LLMs. Anthropic analyzed the consequences of using more books, fewer books, different books. The goal was to improve the "data mix" to improve each LLM and, ultimately, Claude's performance for paying customers.[3]

[3] (*See, e.g.*, Opp. Exh. 12 at -0391318 (engineers were able to "open any of the books"); CC Reply Exh. 45 at -0365941 (some engineers "want[ed] to search for a book" and get its "scanned book file[ ]"); Opp. Exh. 30 at 3 (made copies of "each such dataset or portions thereof" for training); Opp. Exh. 6 at 3–4 (trained on "portions of datasets," with at least two such portions from LibGen and four from PiLiMi); Opp. Expert Zhao ¶¶ 27–28, 30–31 (plus two more from PiLiMi, and at least three from scanned books); CC Opp. Exh. 35 at -0273477–82 (tested subsets of pirated and purchased-and-scanned books to see consequences for training); CC Br. Exh. 12 at -0144508–09 ("iterate[d]" selections from library and "train[ed] new models on the best data"); Br. Expert Kaplan ¶¶ 42–45 (explained goals of improving data mixes); Br. Expert Peterson ¶ 14 (similar)).

Over time, Anthropic came to value most highly for its data mixes books like the ones Authors had written, and it valued them because of the creative expressions they contained. Claude's customers wanted Claude to write as accurately and as compellingly as Authors. So, it was best to train the LLMs underlying Claude on works just like the ones Authors had written, with well-curated facts, well-organized analyses, and captivating fictional narratives — above all with "good writing" of the kind "an editor would approve of" (Opp. Exh. 3 at -03433). Anthropic could have trained its LLMs without using such books or any books at all. That would have required spending more on, say, staff writers to create competing exemplars of good writing, engineers to revise bad exemplars into better ones, energy bills to power more rounds of training and fine-tuning, and so on. Having canonical texts to draw upon helped (*e.g.*, Opp. Expert Zhao ¶ 81).

Each work selected for training any given LLM was copied in four main ways — and in fact so many times that Anthropic admits it would be impractical even to estimate.

First, each work selected was copied from the central library to create a working copy for the training set.

Second, each work was cleaned to remove a small amount of lower-valued or repeating text (like headers, footers, or page numbers), with a "cleaned" copy resulting. If the same book appeared twice, or if while looking across the entire provisional training set it became clear there was some other reason to cull a book or category, Anthropic had the capability to delete relevant copy(ies) from the set at this step (*see* CC Br. Expert Zhao ¶¶ 71–72).

Third, each cleaned copy was translated into a "tokenized" copy. Some words were "stemmed" or "lemmatized" into simpler forms (e.g., "studying" to "study"). And, all characters were grouped into short sequences and translated into corresponding number sequences or "tokens" according to an Anthropic-made dictionary. The resulting tokenized copies were then copied repeatedly during training. By one account, this process involved the iterative, trial-and-error discovery of contingent statistical relationships between each word fragment and all other word fragments both within any work and across trillions of word

fragments from other copied books, copied websites, and the like. Other steps in training are not at issue here (*id.* ¶¶ 73–76; *see* Opp. Expert Zhao ¶ 38 & n.6).

Fourth, each fully trained LLM itself retained "compressed" copies of the works it had trained upon, or so Authors contend and this order takes for granted. In essence, each LLM's mapping of contingent relationships was so complete it mapped or indeed simply "memorized" the works it trained upon almost verbatim. So, if each completed LLM had been asked to recite works it had trained upon, it could have done so (*e.g.*, Opp. Expert Zhao ¶ 74). Further steps refining the LLM are not at issue here.

However, that was as far as the training copies propagated towards the outside world. When each LLM was put into a public-facing version of Claude, it was complemented by other software that filtered user inputs to the LLM and filtered outputs from the LLM back to the user (*id.* ¶¶ 75–77). As a result, Authors do not allege that *any* infringing copy of their works was or would ever be provided to users by the Claude service. Yes, Claude could help less capable writers create works as well-written as Authors' and competing in the same categories. But Claude created no exact copy, nor any substantial knock-off. Nothing traceable to Authors' works. Such allegations are simply not part of plaintiffs' amended complaint, nor in our record.

Neither side puts directly at issue any copies of any works that might have been used for the filtering software. Nor will this order.

In sum, the copies of books pirated or purchased-and-destructively-scanned were placed into a central "research library" or "generalized data area," sets or subsets were copied again to create training copies for data mixes, the training copies were successively copied to be cleaned, tokenized, and compressed into any given trained LLM, and once trained an LLM did not output through Claude to the public any further copies. Finally, once Anthropic decided a copy of a pirated or scanned book in the library would not be used for training at all or ever again, Anthropic still retained that work as a "hard resource" for other uses or future uses. At least one work from each Author was present in every phase described above.

* * *

In August 2024, the three individual authors brought this putative class action complaining that Anthropic had infringed its federal copyrights by pirating copies for its library and by reproducing them to train its LLMs (Compl. ¶¶ 45–46, 71; *see* Amd. Compl. ¶¶ 47–48, 75). In October 2024, a scheduling order required that any motion for class certification be brought by March 6, 2025 (Dkt. No. 49).

The individual authors soon amended their complaint to include affiliated corporate entities as named plaintiffs, with consent. And, Anthropic chose not to move to dismiss the amended complaint, as it earlier had planned (*see* Dkt. No. 37). Instead, Anthropic moved to allow an early motion for summary judgment on fair use, even before class certification (Dkt. No. 88; *see* Feb. 25, 2025 Tr. 15). Permission was granted.

Anthropic now moves for summary judgment on fair use only. Fair use is a legal question for the judge with underlying fact questions, if any, for the jury. To prevail on summary judgment, Anthropic must rely on undisputed facts and/or factual inferences favoring the opposing side. Anthropic thus bears the burdens of production and persuasion in this motion. *See Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 23–24 (2021); *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 547 n.21 (2023); *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 590 & n.20, 594 (1994); *see also Nissan Fire & Marine Ins. Co. v. Fritz Cos.*, 210 F.3d 1099, 1102–03 (9th Cir. 2000).

Notably, in its motion, Anthropic argues that pirating initial copies of Authors' books and millions of other books was justified because all those copies were at least reasonably necessary for training LLMs — and yet Anthropic has resisted putting into the record what copies or even sets of copies were in fact used for training LLMs. For example, at oral argument, Anthropic asserted that if a purported fair user had retained pirated copies for uses beyond the fair use, then her piracy would not be excused by the fair use (Tr. 53, 56). But when Authors earlier interrogated Anthropic in discovery about what library copies (the original copies "obtained or created" by Anthropic) Anthropic had recopied for further uses, Anthropic responded that providing information about any copies made for uses beyond training commercially released LLMs would be overbroad, and that it could not count up all its

copying even for LLMs in any case (*e.g.*, Opp Exh. 30 at 3). We know that Anthropic has more information about what it in fact copied for training LLMs (or not). Anthropic earlier *produced* a spreadsheet that showed the composition of various data mixes used for training various LLMs — yet it clawed back that spreadsheet in April (Opp. Fredricks Decl. ¶¶ 2–3). A discovery dispute regarding that spreadsheet remains pending. But Anthropic did not need a court order to offer up what it possessed in support of its motion. All deficiencies must be held against Anthropic and not the other way around.

This is the first substantive order in this case. A contemporaneous motion for class certification remains pending. It proposes one class related to works that were pirated (whether or not used to train LLMs), and a second class related to works that were purchased, scanned, and used in training LLMs. This order follows full briefing, a hearing, and supplemental briefing.

To summarize the analysis that now follows, the use of the books at issue to train Claude and its precursors was exceedingly transformative and was a fair use under Section 107 of the Copyright Act. And, the digitization of the books purchased in print form by Anthropic was also a fair use but not for the same reason as applies to the training copies. Instead, it was a fair use because all Anthropic did was replace the print copies it had purchased for its central library with more convenient space-saving and searchable digital copies for its central library — without adding new copies, creating new works, or redistributing existing copies. However, Anthropic had no entitlement to use pirated copies for its central library. Creating a permanent, general-purpose library was not itself a fair use excusing Anthropic's piracy.

ANALYSIS

Section 107 of the Copyright Act identifies four factors for determining whether a given use of a copyrighted work is a fair use:

> [T]he fair use of a copyrighted work . . . for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —

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

These factors presuppose a "use." So, at the threshold, a court must decide whether a "copyrighted [work] has been used in multiple ways," then evaluate each. *Warhol*, 598 U.S. at 533. Uses do not turn on "the subjective intent of the user" but on "an objective inquiry into what use was made, *i.e.*, what the user d[id] with the original work." *Id.* at 544–45. A "use" should be construed narrowly enough to not "swallow" distinguishable infringing uses, much less categories of exclusive rights *in toto*. *Id.* at 541, 543 n.18, 546–48. Sometimes, the challenged copying involves just one use: In *Perfect 10, Inc. v. Amazon.com, Inc.*, Google visited websites having full-sized images, made only reduced-sized copies, and incorporated those directly into its search engine — the sole use of the thumbnails being as "pointer[s]" to the images themselves. 508 F.3d 1146, 1157, 1160, 1165 (9th Cir. 2007). Sometimes, the copying involves many uses: In the *Google Books* cases, Google borrowed books from libraries, made both full-image and text-only copies, and incorporated different copies into different tools — one use being to reveal information "*about those books*," another use being to provide the books to print-disabled patrons, and still another being to back up the print books if lost. *Authors Guild v. Google, Inc.*, 804 F.3d 202, 217 (2d Cir. 2015) (quoted); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 97, 101, 103 (2d Cir. 2014) (other cited uses).

Our parties debate an instructive decision. In *American Geophysical Union v. Texaco Inc.*, Texaco employees used scientific articles in a central library, used copies of them in personal desk libraries, and used selected copies again in the scientific laboratory — the first use paid for, the second infringing, and the third plausibly fair but in fact a rare occurrence. 802 F. Supp. 1, 4–5, 14 (S.D.N.Y. 1992) (Judge Pierre Leval), *aff'd*, 60 F.3d 913, 918–19, 926 (2d Cir. 1994).

Here, our parties contest what use or uses are at issue. Anthropic contends it copied Authors' books only for *one* use: *Only* to train LLMs. By contrast, Authors contend it did so for at least *two* uses: *First* to build a vast, central library of potentially useful content, and *second* to train specific LLMs using shifting sets and subsets of that content — over time selecting the more well-organized and well-expressed works for training. Authors also complain that the print-to-digital format change was itself an infringement not abridged as a fair use (Opp. 15, 25). Authors do not allege, however, that any LLM outputs infringing upon their works ever reached users of the public-facing Claude service.

This order addresses each of the four factors in turn, pointing out how each applies to the training copies and to the purchased and pirated library copies. It concludes with an integrated analysis.

1. THE PURPOSE AND CHARACTER OF THE USE.

For a given use at issue, the first factor addresses "the purpose and character of th[at] use, including whether [it] is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1).

A. THE COPIES USED TO TRAIN SPECIFIC LLMS.

All agree that one use at issue was training LLMs to receive text inputs and return text outputs. More specifically, Anthropic used copies of Authors' copyrighted works to iteratively map statistical relationships between every text-fragment and every sequence of text-fragments so that a completed LLM could receive new text inputs and return new text outputs as if it were a human reading prompts and writing responses. Authors further argue — and this order takes for granted — that such training entailed "memoriz[ing]" works by "compress[ing]" copies of those works into the LLM (Opp. 16–17; *see* Opp. Expert Zhao ¶ 74). The LLMs "memorize[d] A LOT, like A LOT" (Opp. Exh. 35 at -029109). Regardless, the "purpose and character" of using works to train LLMs was transformative — spectacularly so.

To repeat and be clear: Authors do not allege that any LLM output provided to users infringed upon Authors' works. Our record shows the opposite. Users interacted only with the Claude service, which placed additional software between the user and the underlying LLM to

ensure that no infringing output ever reached the users. This was akin to the limits Google imposed on how many snippets of text from any one book could be seen by any one user through its Google Books service, preventing its search tool from devolving into a reading tool. *Google*, 804 F.2d at 222. Here, if the outputs seen by users had been infringing, Authors would have a different case. And, if the outputs were ever to become infringing, Authors could bring such a case. But that is not this case.

Instead, Authors challenge only the inputs, not the outputs, of these LLMs. They point to the fully trained LLMs and the Claude service only to shed light on how training itself uses copies of their works and the ways the Claude service could be used to produce still other works that would compete with their works. This order does the same. Authors' arguments that the training use is not transformative are unavailing.

First, Authors argue that using works to train Claude's underlying LLMs was like using works to train any person to read and write, so Authors should be able to exclude Anthropic from this use (Opp. 16). But Authors cannot rightly exclude anyone from using their works for training or learning as such. Everyone reads texts, too, then writes new texts. They may need to pay for getting their hands on a text in the first instance. But to make anyone pay specifically for the use of a book each time they read it, each time they recall it from memory, each time they later draw upon it when writing new things in new ways would be unthinkable. For centuries, we have read and re-read books. We have admired, memorized, and internalized their sweeping themes, their substantive points, and their stylistic solutions to recurring writing problems.

Second, to that last point, Authors further argue that the training was intended to memorize their works' creative elements — not just their works' non-protectable ones (Opp. 17). But this is the same argument. Again, Anthropic's LLMs have not reproduced to the public a given work's creative elements, nor even one author's identifiable expressive style (assuming arguendo that these are even copyrightable). Yes, Claude has outputted grammar, composition, and style that the underlying LLM distilled from thousands of works. But if someone were to read all the modern-day classics because of their exceptional expression,

memorize them, and then emulate a blend of their best writing, would that violate the Copyright Act? Of course not. Copyright does not extend to "method[s] of operation, concept[s], [or] principle[s]" "illustrated[ ] or embodied in [a] work." 17 U.S.C. § 102(b); *see, e.g.*, *Nichols v. Universal Pictures Corp.*, 45 F.2d 119, 120–22 (2d Cir. 1930) (Judge Learned Hand) (stage properties and storytelling elements); *Apple Comput., Inc. v. Microsoft Corp.*, 35 F.3d 1435, 1445 (9th Cir. 1994) ("user-friendly" design principles and elements); *Swirsky v. Carey*, 376 F.3d 841, 848 (9th Cir. 2004) (music theory principles and chord progressions).

Third, Authors next argue that computers nonetheless should not be allowed to do what people do.

Authors cite a decision seeming to say as much (Opp. 16–17). But the judge there twice emphasized while discussing "purpose and character" of the use that what was trained was "not generative AI (AI that writes new content itself)." Rather, what was trained — using a proprietary system for finding court opinions in response to a given legal topic — was a competing AI tool for finding court opinions in response to a given legal topic. That was not transformative. *Thomson Reuters Enter. Centre GmbH v. Ross Intell. Inc.*, 765 F. Supp. 3d 382, 398 (D. Del. 2025) (Judge Stephanos Bibas), *appeal docketed*, No. 25-8018 (3d Cir. Apr. 14, 2025).

A better analogue to our facts would be an AI tool trained — using court opinions, and briefs, law review articles, and the like — to receive legal prompts and respond with fresh legal writing. And, on facts much like those, a different court came out the other way. It found fair use. *White v. W. Pub. Corp.*, 29 F. Supp. 3d 396, 400 (S.D.N.Y. 2014) (Judge Jed Rakoff).

The latter use stood sufficiently "orthogonal" to anything that any copyright owner rightly could expect to control. *See Warhol*, 598 U.S. at 538–40. It could thus be freed up for the copyist to use, "promot[ing] the progress of science and the arts, *without* diminishing the incentive to create." *Id.* at 531 (emphasis added); *see* U.S. CONST. art. I, § 8, cl. 8.

In short, the purpose and character of using copyrighted works to train LLMs to generate new text was quintessentially transformative. Like any reader aspiring to be a writer, Anthropic's LLMs trained upon works not to race ahead and replicate or supplant them — but

to turn a hard corner and create something different. If this training process reasonably required making copies within the LLM or otherwise, those copies were engaged in a transformative use.

The first factor *favors* fair use for the training copies.

B. THE COPIES USED TO BUILD A CENTRAL LIBRARY.

But that is not the only use at issue. Recall that Anthropic purchased millions of print books for its central library and pirated millions of digital books for its central library, too. It used specific sets and subsets of books for training specific LLMs. And, it then retained all the copies in its central library for other uses that might arise *even after deciding it would not use them to train any LLM (at all or ever again)*. Anthropic seems to believe that because some of the works it copied were sometimes used in training LLMs, Anthropic was entitled to take for free all the works in the world and keep them forever with no further accounting. There is no carveout, however, from the Copyright Act for AI companies.

Because the legal issues differ between the library copies Anthropic purchased and pirated, this order takes them in turn.

(i) The Purchased Library Copies Converted from Print to Digital.

Anthropic *purchased* millions of print copies to "build a research library" (Opp. Exh. 22 at 145, 148). It destroyed each print copy while replacing it with a digital copy for use in its library (not for sharing nor sale outside the company). As to these copies, Authors do not complain that Anthropic failed to pay to acquire a library copy. Authors only complain that Anthropic changed each copy's format from print to digital (*see* Opp. 15, 25 & n.14). On the facts here, that format change itself added no new copies, eased storage and enabled searchability, and was not done for purposes trenching upon the copyright owner's rightful interests — it was transformative.

Anthropic purchased its print copies fair and square. With each purchase came entitlement for Anthropic to "dispose[ ]" each copy as it saw fit. 17 U.S.C. § 109(a). So, Anthropic was entitled to keep the copies in its central library for all the ordinary uses. Yes,

Anthropic changed the format of these library copies from print to digital — giving rise to the issue here.

All agree on the facts of the format change. Anthropic "destructively scan[ned]" the print copies to create the digital ones. Anthropic or its vendors stripped the bindings from the print books, cut the pages to workable dimensions, and scanned those pages — discarding each print copy while creating a digital one in its place. The digital copy was then housed in the "research library" or "generalized data area" in place of the print copy (Opp. Exh. 22 at 145–46, 193–94). Authors do not allege and our record does not show that Anthropic provided its converted digital copies of print books to anyone outside Anthropic.

The parties disagree about the legal consequences of the format change. Was scanning the print copies to create digital replacements transformative? Anthropic argues it was because it was reasonably necessary to training LLMs. Authors argue it was a distinguishable step requiring independent justification.

Here, for reasons narrower than Anthropic offers, the mere format change was a fair use.

Storage and searchability are not creative properties of the copyrighted work itself but physical properties of the frame *around* the work or informational properties *about* the work. *See Texaco*, 802 F. Supp. at 14 (physical), *aff'd*, 60 F.3d at 919; *Google*, 804 F.3d at 225 (informational); *Sony Corp. of Am. v. Universal City Studios, Inc.* ("*Sony Betamax*"), 464 U.S. 417, 447 (1984) (rightful interests). In *Texaco*, the court reasoned that if a purchased scientific journal article had been copied "onto microfilm to conserve space, this might [have been] a persuasive transformative use." 802 F. Supp. at 14 (Judge Pierre Leval), *aff'd*, 60 F.3d at 919 (reducing "bulk[ ]" "might suffice to tilt the first fair use factor in favor of Texaco if these purposes were dominant"). In *Google Books*, the court reasoned that a print-to-digital change to expose information about the work was transformative. *Google*, 804 F.3d at 225 (Judge Pierre Leval). And, in *Sony Betamax*, the Supreme Court held that making a recording of a television show in order to instead watch it at a later time was copying but did not usurp any rightful interest of the copyright owner. 464 U.S. at 447, 455. Important to the Supreme Court's reasoning was the expectation that most such copiers would not distribute the

permanent copies of the work. Finally, in *A&M Records, Inc. v. Napster, Inc.*, our court of appeals recognized the reasoning just explained, and therefore rejected by contrast a digitization effort that was touted as space-shifting but in fact resulted in the multiplication of copies shared with outsiders through a file-sharing service. 239 F.3d 1004, 1019 (9th Cir. 2001), *aff'g in this part* 114 F. Supp. 2d 896, 912–13, 915–16 (N.D. Cal. 2000) (Judge Marilyn Hall Patel) (citing *Sony Betamax* and *Texaco*).

Here, every purchased print copy was copied in order to save storage space and to enable searchability as a digital copy. The print original was destroyed. One replaced the other. And, there is no evidence that the new, digital copy was shown, shared, or sold outside the company. This use was even more clearly transformative than those in *Texaco*, *Google*, and *Sony Betamax* (where the number of copies went up by at least one), and, of course, more transformative than those uses rejected in *Napster* (where the number went up by "millions" of copies shared for free with others).

Yes, Anthropic is a commercial outfit. And, this order takes for granted that Anthropic in fact benefited from the print-to-digital format change — or it would not have gone to all the trouble. But the crux of the first fair use factor's concern for "commercial" use is in protecting the copyright owners and their entitlements to exploit their copyright as they see fit (or not). *See, e.g.*, *Harper & Row, Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 562 (1985). That the accused is a commercial entity is indicative, not dispositive. That the accused stands to benefit is likewise indicative. But what matters most is whether the format change exploits anything the Copyright Act reserves to the copyright owner. Anthropic already had purchased permanent library copies (print ones). It did not create new copies to share or sell outside.

Yes, Authors also might have wished to charge Anthropic more for digital than for print copies. And, this order takes for granted that Authors could have succeeded if Anthropic had been barred from the format change. "But the Constitution's language [in Clause 8] nowhere suggests that [the copyright owner's] limited exclusive right should include a right to divide markets or a concomitant right to charge different purchasers different prices for the same book, [merely] say to increase or to maximize gain." *See Kirtsaeng v. John Wiley & Sons,*

Inc., 568 U.S. 519, 552 (2013); *see also* U.S. CONST. art. I., § 8, cl. 8. Nor does the Copyright Act itself. Section 106 sets out exclusive rights that fair uses under Section 107 abridge. Section 106(1) reserves to the copyright owner the right to make reproductions. But on our facts we face the unusual situation where one copy entirely replaced the another. And, Section 106(2) reserves to the copyright owner the right to make derivative works that add or subtract creative material — as occurs in a "translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, [or] condensation" of a book, 17 U.S.C. § 101 (definitions). For some "other modification[ ]" of a book to constitute a "derivative work," it must itself "represent an original work of authorship." *Ibid.* But on our facts the format was changed but no content was added or subtracted. *See Mirage Editions, Inc. v. Albuquerque A.R.T. Co.*, 856 F.2d 1341, 1342, 1343–44 (9th Cir. 1988) (yes where elements added to create new decorative ceramic).[4] Section 106(3) further reserves to the copyright owner the right to distribute copies. But again, the replacement copy here was kept in the central library, not distributed. *Cf. Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 176–78 (2d Cir. 2018) (enabling searching for "information about the material" can be transformative use, even if some distribution results); *Lewis Galoob Toys, Inc. v. Nintendo of Am., Inc.*, 964 F.2d 965, 968, 971 (9th Cir. 1992) (using nifty converter to "merely enhance[ ]" audiovisual displays emitted from purchased videogame cartridge was fair use of those displays partly because no surplus copies of cartridge or displays were ever created).

As a result, Anthropic's format-change from print library copies to digital library copies was transformative under fair use factor one. Anthropic was entitled to retain a copy of these works in a print format. It retained them instead in a digital format, easing storage and

[4] Even if print-to-digital format change *did* infringe the right to prepare derivative works, Authors have conceded that "Plaintiffs' infringement claims are predicated on Anthropic's unauthorized reproduction (17 U.S.C. § 106(1)); Plaintiffs are not alleging infringement by Anthropic of any right to prepare derivative works (*id.* at § 106(2))" (Dkt. No. 203 at 2 (citations original)). Whether this concession had consequence for copies tokenized and used for training or "compressed" into the trained LLMs is not reached by this order because Anthropic does not rely on Authors' concession and those copies were here used transformatively.

searchability. And, the further copies made therefrom for purposes of training LLMs were themselves transformative for that further reason, as above.

To be clear, this print-to-digital conversion involved a different and narrower form of transformative use than the broader one advanced by Anthropic. Anthropic argues that the central library use was part and parcel of the LLM training use and therefore transformative. This order disagrees. However, this order holds that the mere conversion of a print book to a digital file to save space and enable searchability was transformative for that reason alone. Therefore, the digital copy should be treated just as if the purchased print copy had been placed in the central library.

In sum, the first fair use factor *favors* fair use for the digital library copies converted from purchased print library copies — but these do not excuse the pirated library copies.

(ii) The Pirated Library Copies.

Before buying books for its central library, Anthropic downloaded over seven million pirated copies of books, paid nothing, and kept these pirated copies in its library even after deciding it would not use them to train its AI (at all or ever again). Authors argue Anthropic should have paid for these pirated library copies (*e.g.*, Tr. 24–25, 65; Opp. 7, 12–13). This order agrees.

The basic problem here was well-stated by Anthropic at oral argument: "You can't just bless yourself by saying I have a research purpose and, therefore, go and take any textbook you want. That would destroy the academic publishing market if that were the case" (Tr. 53). Of course, the person who purchases the textbook owes no further accounting for keeping the copy. But the person who copies the textbook from a pirate site has infringed already, full stop. This order further rejects Anthropic's assumption that the use of the copies for a central library can be excused as fair use merely because some will eventually be used to train LLMs.

This order doubts that any accused infringer could ever meet its burden of explaining why downloading source copies from pirate sites *that it could have purchased or otherwise accessed lawfully* was itself reasonably necessary to any subsequent fair use. There is no decision holding or requiring that pirating a book that could have been bought at a bookstore

was reasonably necessary to writing a book review, conducting research on facts in the book, or creating an LLM. Such piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded.

But this order need not decide this case on that rule. Anthropic did not use these copies only for training its LLM. Indeed, it retained pirated copies even after deciding it would not use them or copies from them for training its LLMs ever again. They were acquired and retained, as a central library of all the books in the world.

Building a central library of works to be available for any number of further uses was itself the use for which Anthropic acquired these copies. *One further use* was making further copies for training LLMs. But not every book Anthropic pirated was used to train LLMs. And, every pirated library copy was retained even if it was determined it would not be so used. Pirating copies to build a research library without paying for it, and to retain copies should they prove useful for one thing or another, was its own use — and not a transformative one (*see* Tr. 24–25, 35, 65; Opp. 4–10, 12 n.6; CC Br. Exh. 12 at -0144509 ("everything forever")). *Napster*, 239 F.3d at 1015; *BMG Music v. Gonzalez*, 430 F.3d 888, 890 (7th Cir. 2005).

Anthropic's briefing contains other reasons why it believes its pirated library copies are irrelevant to our fair use analysis, notwithstanding its own statements at our oral argument.

First, Anthropic accepts in this posture that it acted in bad faith but argues that its bad faith in pirating copies cannot "somehow short-circuit[ ]" the fair use analysis (Reply 6 (downplaying *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832, 843 (Fed. Cir. 1992) (applying law of Ninth Circuit))). But its bad faith is not the basis for this decision. Each use of a work must be analyzed objectively. *Warhol*, 598 U.S. at 544–45. The objective analysis here shows the initial copies were pirated to create a central, general-purpose library, as a substitute for paid copies to do the same thing. (Of course, if infringement is found, bad faith would matter for determining willfulness. 17 U.S.C. § 504(c)(2).)

Second, Anthropic argues that its goal to put the copies eventually "to a highly transformative use" requires that each copy and use along the way be justified as having a

transformative use, too (Reply 14). But now Anthropic seeks to take the shortcut Anthropic just said cannot be taken. Again, the Supreme Court tasks us with looking past the "subjective *intent* of the user" to the objective *use* made of each copy. *Warhol*, 598 U.S. at 544–45 (emphasis added). Put another way, what a copyist *says or thinks or feels* matters only to the extent it shows what a copyist in fact *does* with the work. Indeed, the same copy can be used one way, then another, each with a different result. *Id.* at 533. Here, what Anthropic said about its acquisitions at the time — that they were made to "build[ ] a research library" while avoiding a "huge legal/practice/business slog" — are relevant in this regard. And, Anthropic's actual use of these pirated copies was to create its central library of texts that, like any university or corporate library, stored the works' well-organized facts, analyses, and expressive examples for various contingent uses, one being training.[5]

Third, Anthropic argues that *Texaco* — the case involving copies used in a central library, copies used in desk libraries, and copies used in the laboratory — is inapposite. Anthropic argues that the disputed copies in *Texaco* were never used in the laboratory but instead in personal desk libraries for a use "identical to the original purpose and use" of the central library copies, and so not for a transformative use (Reply 8 (summarizing 60 F.3d at 922–23)). By contrast, says Anthropic, here it *did use* copies in the laboratory to train LLMs — a very transformative use. But this is a fast glide over thin ice. Like Texaco, Anthropic possessed copies it did not put into use in the laboratory and it kept those copies in a central library even after its transformative use had been completed. But, unlike Texaco, which bought those copies, Anthropic never paid for the central library copies stolen off the

[5] Our court of appeals has not yet reappraised how bad faith (or good faith) figures in fair use after *Warhol*. Its prior appraisal applied the Supreme Court's statement that "[f]air use presupposes good faith and fair dealing," *Harper & Row*, 471 U.S. at 562 (cleaned up). *See Perfect 10*, 508 F.3d at1164 n.8. Since then, the Supreme Court has renewed its "skepticism about whether bad faith has any role." *Oracle*, 593 U.S. at 32–33 (reiterating doubts of *Campbell*, 510 U.S. at 585 n.18). And, recently, the Supreme Court has held squarely that it is not the "subjective intent" of a copyist that counts, but the "objective . . . use" of the copy. *Warhol*, 598 U.S. at 544–45. This order applies this most recent analysis. *Miller v. Gammie*, 335 F.3d 889, 900 (9th Cir. 2003) (en banc).

internet. *Texaco* also shows why Anthropic is wrong to suppose that so long as you create an exciting end product, every "back-end step, invisible to the public," is excused (Br. 10).

Notably, this is not a case where source copies were unavailable for separate purchase or loan. *See, e.g.*, *NXIVM Corp. v. Ross Inst.*, 364 F.3d 471, 475–76, 478–79 (2d Cir. 2004) (using selections of training manual — otherwise available only to cult's trainees subject to NDAs — to expose cult in critical review); *Time Inc. v. Bernard Geis Assocs.*, 293 F. Supp. 130, 135–36, 138, 146 (S.D.N.Y. 1968) (Judge Inzer Bass Wyatt) (making charcoal drawings of photographs taken of originals otherwise not on sale or loan out to illustrate a history book).[6] Nor were the copies made only incidentally and necessarily from pirated copies. *See, e.g.*, *Perfect 10*, 508 F.3d at 1164 n.8 (copies of images that had been pirated by third-party websites were used to index those same websites while indexing the entire web). Here, piracy was the point: To build a central library that one could have paid for, just as Anthropic later did, but without paying for it.

Nor were the initial copies made immediately transformed into a significantly altered form. In *Perfect 10*, images were copied by the search engine in thumbnail form only and deployed immediately into the transformative use of identifying the full-sized images and the pages from which they came. 508 F.3d at 1160, 1165, 1167. And, in *Kelly v. Arriba Software Corp.*, images were copied at full size and then into thumbnails for immediate use in building a search engine, after which the full-sized copies were immediately deleted. 336 F.3d 811, 815 (9th Cir. 2003). Not here. The *full-text* copies of books were downloaded and maintained "forever."

Nor does the initial copying here even resemble the full-text copying in the *Google Books* cases. There, libraries of authorized copies *already* had been assembled, and *all* copies

[6] Anthropic repeats the misleading characterization of the copyright holder in *Oracle* that the initial copies were there purloined (Reply 5). Not so. "All agree[d] that Google was and remain[ed] free to use the Java language itself. All agree[d] that Google's virtual machine [wa]s free of any copyright issues. All agree[d] that the six-thousand-plus method implementations by Google [we]re free of copyright issues. The copyright issue, rather," was the use of Java for purposes of creating competing software having the same familiar, functional schema. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 978 (N.D. Cal. 2012), *aff'd and rev'd in part*, 750 F.3d 1339 (Fed. Cir. 2014).

therefrom were made for direct employment in a one-to-one further fair use — whether the transformative use of pointing to the works themselves, the use of providing the works in formats for print-disabled patrons, or the use of insuring against going out of print, getting lost, and becoming otherwise unavailable. *HathiTrust*, 755 F.3d at 97, 101, 103; *Google*, 804 F.3d at 206, 216–18, 228 (further distinguishing search and snippet uses, which "test[ed] the boundaries of fair use"). Not so here concerning the pirated copies. No authorized copies existed from which Anthropic made its first copies. No full-text copy therefrom was put immediately into use training LLMs. Not every copy was even necessary nor used for training LLMs. No initial copy was ever deleted, even if never used or no longer used.[7] The university libraries and Google went to exceedingly great lengths to ensure that all copies were secured against unauthorized uses — both through technical measures and through legal agreements among all participants. Not so here. The library copies lacked internal controls limiting access and use.

Nor do the decisions on intermediate copying require anything less than the analysis applied here. Anthropic argues that our court of appeals in *Sega Enterprises Ltd. v. Accolade, Inc.* looked only at the "ultimate use" and "did not analyze a series of atomized acts of 'infringement' distinct from that overall purpose" (Reply 3). To the contrary, the appeals court examined the initial, intermediate, and ultimate copies used by the copyist. *The court explained that the copyist initially purchased commercially available copies of game cartridges* and then made further copies necessarily and "solely in order to discover the functional requirements for compatibility." 977 F.2d 1510, 1522 (9th Cir. 1992). Thus, it reached only one result because on those facts there was only one "overall purpose" for the unauthorized copies. Indeed, the court reaffirmed prior caselaw holding that "intermediate

[7] Training LLMs was not a use where perpetually maintaining a library copy was intrinsic to the proffered fair use (e.g., for a plagiarism-checker service). Nor is this an instance where retaining at least one copy was authorized by contract with the copyright owners (e.g., by agreement to express terms upon submission to a plagiarism-checker service, notwithstanding proposed terms scrawled on a paper prior to submission). *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 635–36 & n.5, 645 n.8 (4th Cir. 2009), *aff'g in relevant parts* 544 F. Supp. 2d 473, 480 (E.D. Va. 2008) (Judge Claude Hilton). Anthropic mischaracterizes this case.

copying of [a work] may infringe the exclusive rights granted to the copyright owner in [S]ection 106 of the Copyright Act regardless of whether the end product of the copying also infringes those rights." *Id.* at 1518–19 (reaffirming *Walker v. Univ. Books*, 602 F.2d 859, 864 (9th Cir. 1979)).

Similarly, in *Sony Computer Entertainment, Inc. v. Connectix Corp.*, our appeals court applied the same law to similarly focused conduct. Another copyist allegedly had purchased an authorized copy and then made further copies solely and necessarily to reverse-engineer compatibility requirements. 203 F.3d 596, 601, 602–03 (9th Cir. 2000).

Both *Sega* and *Sony* avoided imposing an "artificial hurdle" to fair use by generously construing the intermediate copying necessary to the fair use. As one example, *Sega* stated that an engineer should be permitted to reboot her computer while undertaking to reverse-engineer software loaded onto it — even if doing so creates another digital copy of the software and is not strictly necessary to reverse-engineering. *Id.* at 605. But neither *Sega* nor *Sony* fathomed gifting an "artificial head start" to a fair user, either, by treating even the initial copy as an intermediate one.

And, yes, some courts have "not inquire[d]" into intermediate or initial copying at all (Reply 2 (citing *Campbell* as not inquiring into surplus copies in the studio)). But if a "close reading of those cases [ ] reveals that in none of them was the legality of the [initial or] intermediate copying at issue," then it was not raised and not necessarily decided. *Sega*, 977 F.2d at 1519; *see Webster v. Fall*, 266 U.S. 507, 511 (1925). It was expressly decided elsewhere: Our analysis must attend to different uses of different copies, and even to different uses of *the same* copies. *Warhol*, 598 U.S. at 533.

Finally, Anthropic argues that even if the initial copies served a different use than the intermediate and ultimate copies, it was not a use for which Anthropic necessarily would have needed to pay Authors for a copy. In theory, argues Anthropic, it could have done as Google did in *Google Books* — find an existing reference library willing to loan its copies for free as source copies. Or, in theory, it could have done as Anthropic did later — go buy used copies without having to pay Authors at all. *See* 17 U.S.C. § 109(a). But Anthropic did not do those

things — instead it stole the works for its central library by downloading them from pirated libraries.

In sum, the first factor *points against* fair use for the central library copies made from pirated sources — and no damages from pirating copies could be undone by later paying for copies of the same works.

2. THE NATURE OF THE COPYRIGHTED WORK.

The second fair use factor is "the nature of the copyrighted work." 17 U.S.C. § 107(2). This factor "calls for recognition 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. For one thing, less protection is due published works than unpublished ones. For another, less protection is due "factual works than works of fiction or fantasy." *Harper & Row*, 471 U.S. at 563. But less protection is not no protection. Even the arrangement of otherwise unprotectable facts surpasses the low bar for a protectable original work of authorship. *Google*, 804 F.3d at 220.

Here, Anthropic accepts that all of Authors' books — all published, whether non-fiction or fiction — contained expressive elements (Reply 9). And, as set out above, this order accepts Authors' view of the evidence that their works were chosen for their expressive qualities in building a central library and then in training specific LLMs (Opp. 11, 17 (citing, e.g., Opp. Exh. 3 at -03433)).

The main function of the second factor is to help assess the other factors: to reveal differences between the nature of the works at issue and the nature of their secondary use (above), and to reveal any relation between the amount and substantiality of each work taken and the secondary use (next). *E.g.*, *Campbell*, 510 U.S. at 586; *Kelly*, 336 F.3d at 820; *Google*, 804 F.3d at 220; *HathiTrust*, 755 F.3d at 98; *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 612–13 (2d Cir. 2006).

The second factor *points against* fair use for all copies alike.

3. THE AMOUNT AND SUBSTANTIALITY OF THE PORTION USED.

The third fair use factor is "the amount and substantiality of the portion" of the copyrighted work used by the accused. 17 U.S.C. § 107(3). The crux of this factor is whether the amount was "reasonable in relation to the purpose of the copying." *Campbell*, 510 U.S. at 586. Thus, the amount of copying is considered first against the work itself, then more importantly against the proposed transformative purpose. *See Warhol*, 598 U.S. at 543 & n.18.

A. THE COPIES USED TO TRAIN SPECIFIC LLMS.

Copies selected for inclusion in training sets were selected because they were complete and because they contained rich protectible expression, or so this order accepts the record shows for Authors. Was all this copying reasonably necessary to the transformative use?

Yes.

"What matters [ ] is not so much 'the amount and substantiality of the portion used' *in making a copy*, but rather the amount and substantiality of *what is thereby made accessible* to a public [in the purported secondary use] for which it may serve as a competing substitute [for the primary use]." *Google*, 804 F.3d at 222. Here, once again, there is no allegation of any traceable connection between the Claude service's outputs and Authors' works. The copying used to train the LLMs underlying Claude was thus especially reasonable.

In response, Authors object primarily that the copying used in training was both extremely extensive and not strictly necessary.

As to extensive copying, it is true that entire works were copied. And, "copying [ ] entire work[s] 'militate[s] against a finding of fair use.'" *Worldwide Church of God v. Philadelphia Church of God, Inc.*, 227 F.3d 1110, 1118 (9th Cir. 2000) (quoting *Hustler Mag. Inc. v. Moral Majority Inc.*, 796 F.2d 1148, 1155 (9th Cir. 1986)); *see Campbell*, 510 U.S. at 587. But we just addressed why Authors' argument is misdirected. The copies that count for this factor are those that would merely serve the same use as the work's ordinary one. Authors do not allege such copying. The accused use here of the incremental copies is as orthogonal as can be imagined to the ordinary use of a book.

As to strict necessity, Authors make a stronger point. When a productive use is made possible only by borrowing from a specific work, fair use climbs towards its zenith. When a productive use is possible without that borrowing, fair use falls to its nadir — and the borrowing deserves a particularly compelling justification. *See Warhol*, 598 U.S. at 543 & n.18, 547. Here, it is true that Anthropic could have used some other books or no books at all for training its LLMs — or so this order accepts the record shows for Authors. But Anthropic has presented a compelling explanation for why it was reasonably necessary to use them anyway.

For one thing, all agree Anthropic needed billions of words to train any given LLM. If using only books, Anthropic would have needed millions of books per model. If using a set comprising only a small fraction of books and a larger fraction of other texts, Anthropic still would have needed hundreds of thousands of books. Authors contend that because Anthropic showed it could use such smaller sets of books, it surely could have used no books at all — or at least not *their* books (Opp. 23). But Authors forget that "reasonably necessary" does not mean "strictly necessary." Authors do not contest that the volume of text required to train an LLM is monumental. Because using so many works was reasonably necessary, using any one work for *actually training* LLMs was about as reasonable as the next.

For another thing, no output to the public was even alleged to be infringing. So, yes, Authors' works were chosen as the strongest examples of writing. But the compelling benefits of training the LLMs on strong examples were not offset by revelations to the public of any portion of the works themselves. What was copied was therefore especially reasonable and compelling.

The third factor thus *favors* fair use for the training copies.

B. THE COPIES USED TO BUILD A CENTRAL LIBRARY.

But again, there was a separate use — a distinction that makes some difference as to whether the amount and substantiality of the copying was "reasonable in relation to the purpose of the copying" for the library copies. *Campbell*, 510 U.S. at 586.

(i) The Purchased Library Copies Converted from Print to Digital.

For the print library copies that Anthropic purchased and then converted into digital library copies, Anthropic already enjoyed entitlement to keep the copies in its library. The purpose of the copying was to keep them in its library but with more favorable storage and searchability properties. Copying the entire work was exactly what this purpose required. There was no surplus copying. The source copy was destroyed.

The third fair use factor *favors* fair use for the purchased library copies converted from print to digital.

(ii) The Pirated Library Copies.

For the pirated library copies, however, Anthropic lacked any entitlement to hold copies of the books at all. Its purpose, it says, was to train LLMs. But its objective conduct was to seek "all the books in the world" and then retain them even after deciding it would not make further copies from them for training — indicating there were other further uses. Against the purpose of acquiring all the books one could on the chance some might prove useful for training LLMs and maybe other stuff too, almost any unauthorized copying would have been too much. Anthropic copied millions of books *in toto*, Authors' works among them.

The third factor *points against* fair use for the pirated library copies.

4. THE EFFECT OF THE USE UPON THE MARKET FOR OR VALUE OF THE COPYRIGHTED WORK.

The final factor is "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). This factor points against fair use when a copyist makes copies available that displace demand for copies the copyright owner already makes available or readily could. *Texaco*, 60 F.3d at 926–28 (reproduced copies); *Dr. Seuss Enters., L.P. v. ComicMix LLC*, 983 F.3d 443, 461 (9th Cir. 2020) (derivative copies). "While the first factor considers whether and to what extent an original work and secondary use [*in principle* could] have substitutable purposes, the fourth factor focuses on *actual or potential* market substitution." *Warhol*, 598 U.S. at 536 n.12 (emphasis added).

A. THE COPIES USED TO TRAIN SPECIFIC LLMS.

The copies used to train specific LLMs did not and will not displace demand for copies of Authors' works, or not in the way that counts under the Copyright Act.

Again, Authors concede that training LLMs did not result in any exact copies nor even infringing knockoffs of their works being provided to the public. If that were not so, this would be a different case. Authors remain free to bring that case in the future should such facts develop.

Instead, Authors contend generically that training LLMs will result in an explosion of works competing with their works — such as by creating alternative summaries of factual events, alternative examples of compelling writing about fictional events, and so on. This order assumes that is so (Opp. 22–23 (citing, e.g., Opp. Exh. 38)). But Authors' complaint is no different than it would be if they complained that training schoolchildren to write well would result in an explosion of competing works. This is not the kind of competitive or creative displacement that concerns the Copyright Act. The Act seeks to advance original works of authorship, not to protect authors against competition. *Sega*, 977 F.2d at 1523–24.

Authors next contend that training LLMs displaced (or will) an emerging market for licensing their works for the narrow purpose of training LLMs (Opp. 21–22). Anthropic argues that transactional costs would exceed Anthropic's expected benefit from any such bargain, prompting it to cease dealing with any rightsholders or else to cease developing such technology altogether (Br. 22–23). Our record could support either account — so this order must assume Authors are correct. A market could develop (Opp. 19–21 (citing record)). Even so, such a market for that use is not one the Copyright Act entitles Authors to exploit.

None of the cases cited by Authors requires a different result. All contemplated losses of something the Copyright Act properly protected — not the kinds of fair uses for which a copyright owner cannot rightly expect to control. *See TVEyes, Inc.*, 883 F.3d at 181 (use of a right legally reserved to and factually already being licensed by copyright owner); *Texaco*, 60 F.3d 931 (same); *Ringgold v. BET, Inc.*, 126 F.3d 70, 80–81 (2d Cir. 1997) (use of a right legally reserved to and factually likely to be marketable by copyright owner — displaying

images of her artistic work in television shows); *cf. Seltzer v. Green Day, Inc.*, 725 F.3d 1170, 1179 (9th Cir. 2013) (no evidence use could be or "was likely to" be marketable).

The fourth factor thus *favors* fair use for the training copies.

B. THE COPIES USED TO BUILD A CENTRAL LIBRARY.

(i) The Purchased Library Copies Converted from Print to Digital.

For these copies, this order assumes Anthropic's format change from print to digital displaced purchases of new digital copies that Anthropic would have made directly from Authors (had it not been able to purchase print copies in used condition). But for reasons stated under the first factor, such losses did not relate to something the Copyright Act reserves for Authors to exploit. It was a format change.

Authors' next argument, it seems, is that the format change nonetheless exposed it to usurpation of the opportunity to sell rightful copies because Anthropic *might* transmit additional unauthorized digital copies more readily than it could have transmitted additional unauthorized print copies — and that the same would be true for all format converters (*cf.* Opp. 25 n.14; Opp. Expert Malackowski ¶ 52). But after much discovery, there is no inkling in our record of intent to redistribute library copies once acquired nor of inability to secure that valuable library against outside actors. And, if the internal, central library copies did or do in fact lead to further reproduction or distribution, those further copies remain redressable separately by Authors. The format change did not itself usurp the Authors' rightful entitlements.

This factor is thus *neutral* for the purchased library copies converted from print to digital.

(ii) The Pirated Library Copies.

The copies used to build a central library *and* that were obtained from pirated sources plainly displaced demand for Authors' books — copy for copy. Not every person who merely intends to make a fair use of a work is thereby entitled to a full copy in the meantime, nor even to steal a copy so that achieving this fair use is especially simple or cost-effective. Here, the copies employed in training LLMs were one thing, but the copies acquired to assemble a

convenient, general-purpose library of works for various uses for which the company might have of them, if any, was a different use altogether.

Anthropic has almost no rebuttal on these points. *First*, Anthropic argues that "Claude's services do not reduce [or usurp] the value of Plaintiffs' works through substitution in their traditional markets" (*see* Br. Expert Peterson ¶ 33). But stealing pirated copies of Authors' works plainly did. *Second*, Anthropic argues that it may have been able to purchase some books on the open market (and some other texts), but not other texts it copied (*cf. id.* ¶ 48 (re licensing)). But this case does not concern those other texts it could not have purchased. It could have purchased Authors' books (and many others). In fact it later did. *Finally*, Anthropic argues that the effect on these texts from one book foregone was too small to be considered (*see id.* ¶ 77). But the test requires that we contemplate the likely result were the conduct to be condoned as a fair use — namely to steal a work you could otherwise buy (a book, millions of books) so long as you at least loosely intend to make further copies for a purportedly transformative use (writing a book review with excerpts, training LLMs, etc.), without any accountability. As Anthropic itself suggested, "That would destroy the [entire] publishing market if that were the case" (*see* Tr. 53; *see also* Tr. 32, 41; Opp. Expert Malackowski ¶¶ 31–34, 38).

The fourth factor *points against* fair use for the pirated library copies.

5. OVERALL ANALYSIS.

After the four factors and any others deemed relevant are "explored, [ ] the results [are] weighed together, in light of the purposes of copyright." *Campbell*, 510 U.S. at 578.

The copies used to train specific LLMs were justified as a fair use. Every factor but the nature of the copyrighted work favors this result. The technology at issue was among the most transformative many of us will see in our lifetimes.

The copies used to convert purchased print library copies into digital library copies were justified, too, though for a different fair use. The first factor strongly favors this result, and the third favors it, too. The fourth is neutral. Only the second slightly disfavors it. On balance, as

the purchased print copy was destroyed and its digital replacement not redistributed, this was a fair use.

The downloaded pirated copies used to build a central library were not justified by a fair use. Every factor points against fair use. Anthropic employees said copies of works (pirated ones, too) would be retained "forever" for "general purpose" even after Anthropic determined they would never be used for training LLMs. A separate justification was required for each use. None is even offered here except for Anthropic's pocketbook and convenience.

And, as for any copies made from central library copies but not used for training, this order does not grant summary judgment for Anthropic. On this record in this posture, the central library copies were retained even when no longer serving as sources for training copies, "hundreds of engineers" could access them to make copies for other uses, and engineers did make other copies. Anthropic has dodged discovery on these points (*e.g.*, Opp. Exh. 17 at 93–94 (retained); Opp. Exh. 22 at 196 (no limits); Opp. Exh. 30 at 3, 4 (no accounting); *see also* Opp. 15). We cannot determine the right answer concerning such copies because the record is too poorly developed as to them. Anthropic is not entitled to an order blessing all copying "that Anthropic has ever made after obtaining the data," to use its words (Opp. Exh. 30 at 3, 4).

CONCLUSION

With respect to the training copies and the print-to-digital converted copies, this order has drawn all ambiguities and inferences in favor of the opposing side, namely Authors. With respect to the pirated copies, this order has also accepted the Authors' version of the facts. Authors did not move for summary judgment but if they had, then we would have been obligated to accept all reasonable views given the evidence in defendant's favor instead.

This order grants summary judgment for Anthropic that the training use was a fair use. And, it grants that the print-to-digital format change was a fair use for a different reason. But it denies summary judgment for Anthropic that the pirated library copies must be treated as training copies.

We will have a trial on the pirated copies used to create Anthropic's central library and the resulting damages, actual or statutory (including for willfulness). That Anthropic later

bought a copy of a book it earlier stole off the internet will not absolve it of liability for the theft but it may affect the extent of statutory damages. Nothing is foreclosed as to any other copies flowing from library copies for uses other than for training LLMs.

IT IS SO ORDERED.

Dated: June 23, 2025.

WILLIAM ALSUP
UNITED STATES DISTRICT JUDGE

United States District Court
Northern District of California