UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: *Authors Guild, et al. v. OpenAI, Inc., et al.*, No. 1:23-cv-08292-SHS-OTW	25-md-3143 (SHS) (OTW) Hon. Sidney H. Stein Hon. Ona T. Wang ORAL ARGUMENT REQUESTED REDACTED — PUBLIC VERSION

OPENAI'S MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT OF MOTION FOR SUMMARY JUDGMENT

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

Federal Cases

Federal Statutes

Rules

Other Authorities

I. INTRODUCTION[1]

Plaintiffs are authors who allege that OpenAI infringed their copyrights by using their books (among billions of other works) as part of the pretraining of large language models ("LLMs"). That alleged use was fair because it was highly transformative, did not display any meaningful portion of the books to ChatGPT users, and caused no cognizable harm.

OpenAI is a pioneer. Years before "generative AI" became a household term, OpenAI, which was founded in 2015 as a research organization, began a series of studies designed to advance the science of artificial intelligence in a way that would benefit humanity as a whole. After many years of discoveries made and obstacles overcome, that research resulted in ChatGPT—a groundbreaking new technology used by over a billion people to improve their daily lives, advance scientific and medical research, and empower human creativity.

ChatGPT is powered by OpenAI's revolutionary LLMs. Like other LLMs, OpenAI's models are "pretrained" on an enormous volume of text from the Internet. The undisputed facts show that the purpose of OpenAI's pretraining process was to derive broad, unprotectable statistical patterns related to language that can be used to create new text, not reproduce protected expression from copyrighted works. Indeed, OpenAI undertakes extensive efforts to prevent this from happening. That purpose is entirely new and different from the purpose for which Plaintiffs wrote their books, which is why the only two courts to address the use of copyrighted books to pretrain LLMs held it was "highly," "quintessentially," and "spectacularly" transformative. *Kadrey v. Meta Platforms, Inc.,* 788 F. Supp. 3d 1026, 1044 (N.D. Cal. 2025); *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1021–22 (N.D. Cal. 2025). This use is "among

[1] In all quotations appearing herein, all internal quotation marks, citations, alterations, and the like have been omitted, and all emphases added, unless otherwise noted.

the most transformative many of us will see in our lifetimes." *Bartz*, 787 F. Supp. 3d at 1033.

OpenAI's transformative use has not harmed Plaintiffs. ChatGPT does not display copies of Plaintiffs' books to users. The undisputed evidence in this case shows that, at most, 0.00007% of ChatGPT conversation logs "regurgitated" content from Plaintiffs' books and when this rare event does happen, the amount of text regurgitated was typically a few dozen words. So Plaintiffs complain that ChatGPT can provide "summaries" of their books, but this content is no different from commonplace excerpts or summaries that anyone can read for free on the Internet, and has had no negative impact on the market for Plaintiffs' books. Plaintiffs also complain about a purported scourge of new books written by or with the assistance of "AI," but competitive harm from new, non-infringing works is not harm that copyright protects against, and Plaintiffs have not been harmed in any event. They continue to write and successfully sell their books to a wide audience, just as they did before ChatGPT. In short, ChatGPT is simply not a substitute for reading the copyrighted expression in Plaintiffs' books. *See Authors Guild v. Google, Inc.* (*Google Books*), 804 F.3d 202, 225 (2d Cir. 2015). It is a fair use as a matter of law.

II. BACKGROUND

A. OpenAI Before ChatGPT

OpenAI was founded in 2015 as a research lab. SUF[2] ¶ 1. Since then, OpenAI has published numerous academic papers advancing the science of machine learning and "natural language processing" ("NLP"). SUF ¶ 2.

NLP researchers had long hypothesized that language follows statistical patterns that could be used to revise or generate new text. SUF ¶ 3. For decades, researchers ran experiments

[2] OpenAI's Statement of Undisputed Material Facts in Support of Its Motions for Summary Judgment ("SUF").

to evaluate different architectures for capturing linguistic patterns in statistical systems called "language models," using unlicensed books, news articles, and other texts to "train" those models. SUF ¶ 4. And while these efforts advanced considerably—yielding consumer products like Google Translate and Microsoft Word's grammar checker—researchers struggled to achieve the kind of broad and flexible language understanding required to create "more general systems which can robustly perform many tasks." SUF ¶¶ 5–6.

OpenAI ran a series of similar experiments using a new architecture it developed called the "generative pretrained transformer" (or "GPT"). SUF ¶ 7. Like previous researchers, OpenAI trained its early models (GPT-1 and GPT-2) on self-published books and other text from the Internet. SUF ¶ 8. It soon became clear that a deeper scientific understanding of language required training on a larger and more diverse corpus of data. SUF ¶ 9. So to advance its research, OpenAI used broader mixtures of data from the public Internet, including compilations of books from Library Genesis used in training GPT-3 (Spring 2020) and GPT-3.5 (late 2021), long before ChatGPT was released in November 2022. SUF ¶ 10. No other LLMs at issue in this case[3] were trained on material downloaded from Library Genesis. SUF ¶ 11.

OpenAI's GPT architecture was a major innovation, demonstrating that advanced "pretraining" methods using an enormous volume of text could create LLMs capable of comprehending the nuances and breadth of human language. To make those LLMs useful, OpenAI developed another major innovation: a suite of "post-training" techniques designed to train LLMs to answer questions, distinguish safe conversation topics from harmful ones, and behave appropriately. SUF ¶ 12.

OpenAI combined these two scientific breakthroughs (and many others) to create an

[3] The models at issue in this case are GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4o mini. ECF 707 (Order) at 1.

LLM that could safely engage in complex conversations with users about a wide range of topics. To share its progress and advance its research, OpenAI created an interface called "ChatGPT" and published a blog post on November 30, 2022 inviting people to try it out as a free "research preview." SUF ¶ 13. OpenAI did so to "get users' feedback and learn about its strengths and weaknesses." SUF ¶ 14. OpenAI's "original intent was to wind down the ChatGPT demo after [it] had the necessary learning and feedback." SUF ¶ 15.

B. OpenAI After ChatGPT

Within days of its release, more than a million people signed up to use ChatGPT. SUF ¶ 16. The Harvard Business Review hailed ChatGPT as a "tipping point for AI" and noted the model had "crossed a threshold" to become "genuinely useful for a wide range of tasks, from creating software to generating business ideas to writing a wedding toast." SUF ¶ 17. Commenters also hailed OpenAI's post-training efforts. The New York Times proclaimed that "ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public," and praised OpenAI's "commendable steps to avoid the kinds of racist, sexist and offensive outputs that have plagued other chatbots." SUF ¶ 18.

ChatGPT was just the beginning. OpenAI would go on to advance scientific progress and revolutionize the field of generative artificial intelligence it created many times over. OpenAI developed techniques to enable language models to understand audio, images, and video, including by engaging in a verbal dialogue with users. SUF ¶ 19. It also created a mechanism for language models to browse the Internet for factual information relevant to a user's query. SUF ¶ 20. Each of these innovations built upon the foundational discoveries and techniques discussed above and further below.

Over a billion people around the world use OpenAI's LLMs. SUF ¶ 21. Those LLMs are

used every second of every day for a nearly infinite variety of personal and business purposes, including education, healthcare, and scientific research.[4] SUF ¶ 22.

C. How OpenAI Transforms Training Data to Build New Products

This section provides a high-level overview of the multi-layered process OpenAI uses to construct its LLMs. **First**, OpenAI builds and processes a training dataset. This is a challenge: training a useful LLM requires hundreds of billions to tens of trillions of words[5] from a multitude of diverse sources. SUF ¶ 23. For that reason, model developers utilize content from the Internet—a singularly diverse, global source of text—which they gather using the same techniques that search engines like Google have used for decades to gather broad swaths of content from the Internet. SUF ¶¶ 30–31. Once collected, the data is de-duplicated, processed, tokenized (*i.e.*, translated into numbers corresponding to individual words or sub-word parts), and then split into excerpts large enough for the model to derive the semantic and linguistic meaning of the words therein based on surrounding context. SUF ¶¶ 35–37.

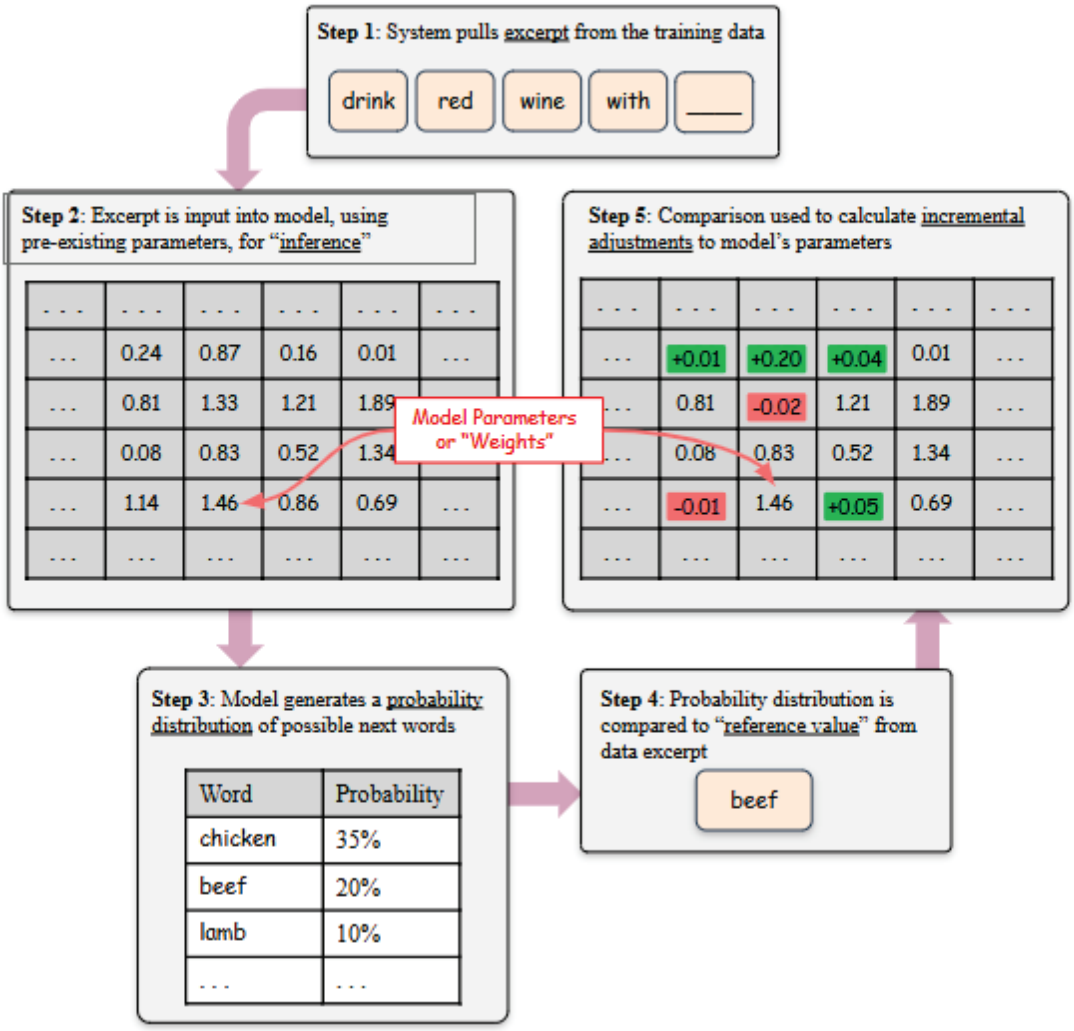
Second, OpenAI then uses those text excerpts for "pretraining," an iterative process whereby the model learns broad linguistic patterns that it can use to comprehend and generate new text—*i.e.*, to *generalize* broader ideas and principles of language. SUF ¶ 38. The model "learns" these patterns by encoding them in a network of up to trillions of *parameters* or *weights*—*i.e.*, numerical values that function like coefficients in a highly complex math equation that generates predictions. SUF ¶ 39. These parameters, represented below as a grid for illustration purposes, begin the training process as a set of random numbers. SUF ¶ 40.

[4] OpenAI continues to innovate with new LLMs and related products not at issue in this case, including breakthroughs in agentic AI, reasoning models, and video and image-generation models.

[5] Technically, the model evaluates "tokens," not words. SUF ¶¶ 24–29. However, for ease of use, OpenAI refers to "words" rather than "tokens" throughout this brief.

...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	0.24	0.87	0.16	0.01	1.10	1.95	0.05	0.24	0.87	0.16	0.01	1.10	1.95	0.05	0.24	0.87	0.16	0.01	1.10	...
...	0.81	1.33	1.21	1.89	0.91	1.18	0.20	0.81	1.33	1.21	1.89	0.91	1.18	0.20	0.81	1.33	1.21	1.89	0.91	...
...	0.08	0.83	0.52	1.34	0.12	0.65	0.71	0.08	0.83	0.52	1.34	0.12	0.65	0.71	0.08	0.83	0.52	1.34	0.12	...
...	1.14	1.46	0.86	0.69	1.25	0.04	0.62	1.14	1.46	0.86	0.69	1.25	0.04	0.62	1.14	1.46	0.86	0.69	1.25	...
...	0.67	0.31	1.56	0.68	1.85	0.05	0.56	0.67	0.31	1.56	0.68	1.85	0.05	0.56	0.67	0.31	1.56	0.68	1.85	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

As illustrated in the next figure, during each step of pretraining, the model takes a randomly selected excerpt (**Step 1, below)** from the training corpus (*e.g.*, "*drink red wine with* ") and uses the model's parameters **(Step 2)** to generate a *probability distribution* **(Step 3)** reflecting predictions of different words that might come next, along with scores indicating the model's level of confidence as to each prediction (*e.g.*, 35% - "chicken", 20% - "*beef*"). SUF ¶ 41. The algorithm compares **(Step 4)** that probability distribution against the *reference value*—*i.e.*, the actual next word in the excerpt—to measure the gap between the two (*train loss*) and uses the results of that comparison to update the model's parameters **(Step 5)**. SUF ¶ 42.



The model does not store any particular excerpt of training data during pretraining. Instead, this cycle (Steps 1-5) repeats millions of times across as many different excerpts, and the model uses those excerpts to make an enormous series of incremental adjustments to its parameters, as illustrated below. SUF ¶ 43.

As training progresses at a massive scale, the parameters eventually discern and encode basic linguistic rules that improve the quality of its predictions—leading to a steadily narrowing gap and a lower train loss. SUF ¶ 44. This is roughly analogous to a student who completes a practice exam (*training excerpts*), uses an answer key to score her performance (*train loss*),

thinks about what she can learn from the questions she got wrong (*parameter update*), and then achieves a better score on the next practice exam (improving *train loss*).

Third, during pretraining, OpenAI periodically uses a subset of collected data set aside for testing purposes (the "*test set*") to measure how accurately the model makes predictions on data it has never seen (and never used to update its parameters). SUF ¶ 45. These measurements result in a metric called *test* or *validation loss*. SUF ¶ 46. Test loss, not train loss, is the core metric OpenAI uses to evaluate pretraining. SUF ¶ 47. Test loss is a direct measure of the model's ability to *generalize*: it evaluates the degree to which the model has learned general patterns that it can apply to new, unseen data. SUF ¶ 48. Test loss is roughly analogous to the student's score on a real exam, which evaluates true subject matter understanding by presenting questions the student has never seen before.

Once pretraining is complete, OpenAI can then use that model to generate entirely new text. SUF ¶ 49. This occurs through a process called *inference*, whereby the model processes an input sequence and then uses a *probability distribution* provided by the model to generate a new word in the sequence. SUF ¶ 50. This next-word prediction mechanism is a simple way to model intelligence. For example, showing that a model can process a long judicial analysis and then write the concluding language ("*For the foregoing reasons, defendants' motion is granted*") suggests that the model has internalized not only reading comprehension, but basic legal reasoning. Next-word prediction can also be used iteratively to generate entirely new text. After a model processes an input sequence ("*she served beef and*") and generates a new word ("*carrots*"), the model will append the generation onto the input sequence ("*she served beef and carrots*") and repeat the process over and over again until the sequence is complete ("*she served beef and carrots with a glass of red wine*"). In this way, the models create sentences or even

paragraphs of entirely new text simply by predicting one word at a time—not by reproducing training data.

Fourth, OpenAI *post-trains* its models to engage in question-and-answer dialogue, to distinguish between safe and unsafe requests, and to provide balanced, non-biased responses. SUF ¶ 52. This is how OpenAI converts a pretrained model into something that can assist users by drawing on its generalized knowledge from pretraining to provide helpful and safe responses to a virtually infinite array of user queries—all via iterative next-word prediction. SUF ¶¶ 51–52.

Fifth, OpenAI conducts a comprehensive battery of evaluations to ensure safe behavior and engages human "red teams" to test the model's resilience to attempts to circumvent its safety features. SUF ¶ 51. These safety features include rigorous training of the model to minimize potential bias from its training data—efforts The New York Times has praised from ChatGPT's launch. SUF ¶ 53 (noting ChatGPT's appropriate handling of a deliberately provocative question involving Nazis). If a model passes these evaluations, it is deployed via one of two interfaces. The first is ChatGPT: a consumer-facing chatbot interface, available for free via website or mobile app. Users submit "prompts" to ChatGPT, and ChatGPT generates new text in response via the *inference* process described above. SUF ¶ 56. The second is OpenAI's application programming interface ("API"). SUF ¶ 57. The API generates text via the same *inference* process, but is primarily used by software developers to power new applications or features in their products. SUF ¶ 57.

D. How OpenAI Prevents Regurgitation

The purpose of pretraining is to teach the model generalizable rules of language and facts about the world. SUF ¶ 58. The models learn these by analyzing vast amounts of existing texts and identifying the *commonalities* between them. SUF ¶ 59. The models encode these observed

commonalities as patterns corresponding to linguistic rules (*e.g.*, subject-verb agreement) or common facts (*e.g.*, the name of our second President) after seeing those patterns in hundreds of thousands of training examples. So just as a model might learn that "john adams" was the second President after seeing thousands of texts discussing the American Revolution and its aftermath during training, a model that encounters thousands of reviews, blog posts, and short excerpts of the works of J.R.R. Tolkien might associate "frodo baggins" with the name of a fictional hobbit who bears a magic ring.

For similar reasons, a model may encode not only facts but common idioms or turns of phrase (*e.g.*, "*the grass is always greener . . .*") or even public-domain texts (*e.g.*, "*four score and seven years ago, our fathers . . .*")—again, because those sequences appear in countless variations in the training corpus. Researchers sometimes refer to this encoding of specific text sequences as *memorization*. SUF ¶ 62. For the same reason a model might "memorize" passages from the Gettysburg Address after seeing them quoted in thousands of texts, a model may encounter the same oft-repeated book passage or news article excerpt hundreds of times from different sources during training, and as a result, might believe the words therein constitute a common pattern. SUF ¶ 63. Thus, in that highly atypical scenario, a model prompted with an initial passage of a memorized book passage or news article and tasked with completing it (typically using customized settings for that purpose) may do so by generating the next few sentences verbatim. Researchers refer to this as *regurgitation*. SUF ¶ 64.

OpenAI has invested significant time and resources in designing models to learn general linguistic rules and common facts, but to not regurgitate copyrighted material. SUF ¶ 65. This is no easy task, for the simple reason that nothing in the pretraining process allows a model to distinguish between the linguistic commonalities that arise from language itself (*i.e.*, idioms,

public-domain texts) and the linguistic commonalities that stem from the popularity of specific copyrighted works. SUF ¶¶ 59-63. Nonetheless, from the very earliest days of OpenAI's work on the GPT models, it has devised and implemented a number of innovative solutions to this issue—reflecting the fact that OpenAI's intention is for the models to generate new text, not to regurgitate training data. SUF ¶¶ 65–66. Those efforts are broad and multifaceted, and largely fall into four categories: (1) data de-duplication; (2) pretraining optimizations; (3) copyright post-training; and (4) output filtering.

De-duplication. One of the most significant drivers of memorization is duplication in the training data. SUF ¶ 67. When a text sequence occurs multiple times in the training set, the model is far more likely to interpret the repeated sequence as a common idiom and memorize it, rather than simply learning the broad patterns the sequence shares with other sequences. SUF ¶ 68. As such, de-duplicating training data—*i.e.*, searching for repeated passages and removing them—is one way to minimize memorization and, by extension, regurgitation of copyrighted material. SUF ¶ 69.

OpenAI began de-duplicating its training data at least as early as 2019 via its work on the GPT-2 model. SUF ¶ 70. These efforts grew more sophisticated over time as the research evolved. For its work on GPT-3, OpenAI developed "fuzzy deduplication" techniques to capture not only exact but also near-verbatim matches. SUF ¶ 71. For later models, OpenAI also implemented a technique to identify repeated phrases in the training data—*e.g.*, longer passages frequently quoted in book reviews—to avoid training disproportionately on the same text sequence. SUF ¶ 72.

Training Optimizations**.** OpenAI also designs its training algorithm to avoid memorization that might lead to unwanted regurgitation. SUF ¶ 73. For example, OpenAI

implements mechanisms to avoid training on the same training data excerpts until all of the data within an entire training set have been depleted. SUF ¶ 74. Moreover, OpenAI implements *regularization* techniques like drop-out and weight decay, which encourage the model to focus on broad, generalizable patterns and to ignore or deemphasize idiosyncratic features of specific training examples. SUF ¶ 75.

Post-training. OpenAI also conducts copyright-specific post-training to train models to avoid regurgitating any memorized training data. SUF ¶ 76. These anti-regurgitation campaigns were some of the earliest implementations of the innovative post-training techniques that OpenAI had invented and described in academic papers in the years leading up to the release of ChatGPT in 2022. *See* SUF ¶¶ 70–76. OpenAI uses these techniques to train its models to recognize and refuse requests for copyrighted content. SUF ¶ 76. According to OpenAI's internal evaluations, these techniques were enormously successful at training the model to refuse such requests. SUF ¶ 76. OpenAI continued to improve these techniques throughout 2023 and 2024. SUF ¶ 78.

Bloom Filter. As a fail-safe, OpenAI also designed and implemented a filter to assess and screen model outputs as they are generated. SUF ¶ 79. But because it is infeasible to cross-reference every generated model output against the hundreds of billions or trillions of words in OpenAI's training datasets, OpenAI used a probabilistic data structure called a *bloom filter* to evaluate each model output against a subset of training data sequences. SUF ¶ 80. OpenAI then ran a series of internal tests to identify text sequences that were more susceptible to regurgitation and used those sequences to populate the filter's index. SUF ¶ 81.

These mitigations, implemented separately and taken together, are highly effective. OpenAI's expert searched for regurgitation of the books at issue in the case ("the Asserted

Works")[6] in a sample of 95 million outputs from the 20 million ChatGPT conversation logs that OpenAI produced in this case and found only 14 instances containing verbatim or near-verbatim regurgitation—*i.e.*, an alleged regurgitation rate of 0.00007%. Of these 14 instances, all but two were 32 words or fewer (the two outliers were 67 and 122 words), and none represented more than 0.017% of an Asserted Work. SUF ¶¶ 82–84. Plaintiffs reviewed a larger, overlapping sample and identified an even smaller rate of regurgitation: 31 instances out of 108 million ChatGPT conversation logs (0.00003%). SUF ¶ 91.

Separate from review of these actual user chats, Plaintiffs' expert spent significant effort trying to brute force OpenAI's LLMs to reproduce text similar to text from the Asserted Works by relying almost exclusively on prompts that fed the models passages from Plaintiffs' own books and directed them to continue the text verbatim. *See* SUF ¶ 85. After more than 5.3 million prompts, the longest span of contiguous text he could generate was a 1,899-word excerpt from "A Game of Thrones," which is only 0.62% of the book. SUF ¶¶ 85–88.

E. The Plaintiffs, the Asserted Works, and the Allegations in this Case

Plaintiffs are thirteen well-known authors, the Authors Guild, and related entities that own copyrights in books written by the author plaintiffs.[7] They allege that OpenAI infringed their copyrights in the Asserted Works, most of which are works of fiction. SUF ¶ 98. For the purposes of this motion only, OpenAI does not dispute that it used one or more of the Asserted Works (in addition to billions of other works) in pretraining some of its LLMs.

Plaintiffs include international bestselling authors like David Baldacci, John Grisham,

[6] The Asserted Works are 212 published books authored or purportedly owned by Plaintiffs.

[7] Plaintiffs propose to represent a class of authors whose books were "downloaded or otherwise reproduced by OpenAI or Microsoft[]." ECF 183.

George R.R. Martin, and Jodi Picoult, and several other prominent authors. Collectively, they have won dozens of awards and sold hundreds of millions of books. SUF ¶¶ 92–93. Several of them have made [redacted] over the course of their careers. SUF ¶ 94. They have continued to write and sell new books after ChatGPT was released, including several books that have been highly successful. SUF ¶ 95.

The Asserted Works are "trade books," meaning they are sold via bookstores and other standard channels for general readership. SUF ¶ 97. They span a range of genres, from mystery, to fantasy, to romance. Several of them are bestsellers. SUF ¶¶ 98–99.

Plaintiffs allege that OpenAI downloaded some of the Asserted Works from Library Genesis in 2018-2019 and used those downloaded copies to pretrain GPT-3 and GPT-3.5. ECF 183 ¶ 111. They also allege that OpenAI used the Asserted Works to train later models. *Id*. at ¶ 117. Although OpenAI did not use content from Library Genesis for those later models, it did use material from other publicly available websites, including material from a publicly available index of billions of webpages created by the non-profit organization Common Crawl. SUF ¶ 139. Some, if not all, of the Asserted Works (or portions of them) exist on those websites. SUF ¶ 103.

Plaintiffs admit that ChatGPT does not provide complete copies or even excerpts of their books to users. SUF ¶ 105. They instead claim that ChatGPT provides short "summaries" or "outlines" of the Asserted Works. SUF ¶ 103. Plaintiffs also allege that OpenAI's LLMs can be used to create new books at "the press of a button," and that those new books—although they do not reproduce any part of the Asserted Works—displace the Asserted Works in the marketplace. ECF 183 ¶ 3. But Plaintiffs have admitted they are unaware of any books generated by OpenAI's LLMs or any particular lost sales. SUF ¶¶ 110-119, 130-131. Mr. Grisham, for example, testified he is [redacted]

[redacted]," and "[redacted]." SUF ¶ 116. Mr. Martin testified that he [redacted]." SUF ¶ 117. Mr. Franzen, when asked if he "[redacted]," responded, "[redacted]" SUF ¶ 118. And multiple plaintiffs admitted they are [redacted]" or "[redacted]." SUF ¶¶ 110–12.

Plaintiffs also allege that they have been harmed because OpenAI did not pay to acquire or use their books to train OpenAI's LLMs. It is undisputed that none of the Plaintiffs have ever [redacted]. SUF ¶ 121. Nor did they ever [redacted]. SUF ¶ 122. Most of the Asserted Works were written years before modern LLM technology even existed. SUF ¶ 100.

Plaintiffs do not dispute that "[redacted]," and that [redacted] SUF ¶ 132. They further admit there is [redacted] and that "[redacted] SUF ¶ 132. Plaintiffs have also used ChatGPT to help them research, write, and brainstorm new ideas. SUF ¶ 135. Plaintiffs' expert in this case opined that ChatGPT is "[redacted]." SUF ¶ 135.

III. LEGAL STANDARD

Summary judgment is warranted where "the movant shows that there is no genuine dispute as to any material fact and the movant is entitled to judgment as a matter of law." Fed. R. Civ. P. 56(a). A fact is "material" only if it could affect the outcome under the governing law. *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 248 (1986). A dispute is "genuine" only "if the evidence is such that a reasonable jury could return a verdict for the nonmoving party." *Id.*

Fair use is a mixed question of fact and law. *Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 24 (2021). The "ultimate question" of fair use "primarily involves legal work[,]" and is therefore a "legal question" for the Court to decide. *Id.* Because fair use is an affirmative defense to copyright infringement, the proponent of the defense bears the ultimate burden of proof. *Romanova v. Amilus Inc.*, 138 F.4th 104, 109-110 (2d Cir. 2025). But the copyright owner bears at least an initial burden to identify a legally cognizable market that has been harmed by the alleged use. *Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163, 194 (2d Cir. 2024).

"[T]he difficulty of the fair use question is not a valid reason to shy away from [resolving it at] summary judgment." *Maxtone-Graham v. Burtchaell*, 803 F.2d 1253, 1259 (2d Cir. 1986). Courts regularly "resolve issues of fair use at the summary judgment stage where there are no genuine issues of material fact as to such issues." *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 608 (2d Cir. 2006); *Authors Guild v. Google, Inc.* (*Google Books*), 804 F.3d 202, 225 (2d Cir. 2015) (affirming summary judgment finding fair use); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 101 (2d Cir. 2014) (same); *Swatch Grp. Mgmt. Servs. Ltd. v. Bloomberg L.P.*, 756 F.3d 73 (2d Cir. 2014) (same); *Blanch v. Koons*, 467 F.3d 244, 250 (2d Cir. 2006) (same); *Wright v. Warner Books, Inc.*, 953 F.2d 731, 740 (2d Cir. 1991) (same).

IV. OVERVIEW OF KEY FAIR USE PRINCIPLES

No property right is unlimited. Copyright law "has never accorded the copyright owner complete control over all possible uses of his work." *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 432 (1984). The Copyright Act contains built-in flexibility for new technologies, providing that "the fair use of a copyrighted work . . . is not an infringement of copyright." 17 U.S.C. § 107. Fair use "avoid[s] rigid application of the copyright statute when . . . it would stifle the very creativity which that law is designed to foster." *Stewart v. Abend*, 495 U.S. 207, 236 (1990); *see also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579 (1994) (stating that fair use "guarantee[s] [] breathing space within the confines of copyright").

For decades, courts have applied the fair use doctrine to make space for new, utility-expanding technologies like the VCR, Internet search engines, services for the print-disabled, book research tools, and mobile computing platforms—notwithstanding the objections of copyright owners who insisted that such rulings would "strangle[]" their industries,[8] "disempower[] Authors[,]"[9] pose "enormous harm to copyright holders[,]"[10] "decimate existing and developing markets for literary works[,]"[11] or condone "egregious act[s]" of "piracy."[12] At the same time, courts have declined to extend the doctrine to technologies that simply re-publish the expressive content of the original work without sufficient justification. *See, e.g.*, *Fox News Network, LLC v. Tveyes, Inc.*, 883 F.3d 169, 175 (2d Cir. 2018) (users could watch an "unlimited number of clips" from copyrighted TV broadcasts); *Hachette*, 115 F.4th at 175 (users could read

[8] *Home Recording of Copyrighted Works: Hearing on H.R. 4783, H.R. 4794, H.R. 4808, H.R. 5250, H.R. 5488, and H.R. 5705 Before the H. Comm. on the Judiciary*, 97th Cong. 8 (1982).
[9] Appellants' Brief at 55 *in Authors Guild v. Google, Inc.*, No. 13-4829, 2014 WL 1509692 (2d Cir., Apr. 11, 2014).
[10] Appellant's Brief at 1 *in Perfect 10, Inc. v. Google Inc.*, No. 06-55406, 2006 WL 3096732 (9th Cir. May 31, 2006).
[11] Appellants' Brief at 3 *in Authors Guild, Inc. v. HathiTrust*, No. 12-4547, 2013 WL 3771981 (2d Cir. July 11, 2013).
[12] Respondent's Brief *in Google LLC v. Oracle Am., Inc.*, No. 18-956, 2020 WL 832871 (U.S. Feb. 12, 2020).

copyrighted books "in full, for free"). All the while, courts have been "reluctan[t] to expand the protections afforded by the copyright without explicit legislative guidance[,]" particularly when doing so would threaten "major technological innovations" that advance the "Progress of Science" and "society's [] interest in the free flow of ideas, information, and commerce[.]" *Sony*, 464 U.S. at 429–30.

Two clear principles emerge from this body of case law. First, it is fair use for a new technology to expand the utility of copyrighted work when it does not deliver the entirety of that content to the public. *Romanova*, 138 F.4th at 115 (copying that "enable[s] the furnishing of valuable information . . . without allowing public access to the copy" is fair). That is why it is fair use to copy thousands of essays to create a plagiarism detector, *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 644–45 (4th Cir. 2009), or scan millions of books to make a book search tool, *HathiTrust*, 755 F.3d at 97–98. This is true even where transformative uses display portions of the original copyrighted content, so long as the new use does not serve as a "competing substitute for the original." *Google Books*, 804 F.3d at 223–24; *see also Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 821 (9th Cir. 2003) (holding that thumbnail images were not substitutes for the original full-size images); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1168 (9th Cir. 2007) (same). Where there is no meaningful likelihood of "significant" market harm from substitution, the use is fair. *Google Books*, 804 F.3d at 223–24.

Second, a use that enables the creation of new works is a copyright good—even if it may cause some (non-cognizable) "harm" to the rightsholder. *See*, *e.g.*, *Campbell*, 510 U.S. at 592 (parody is fair use even where it "kills demand for the original"). The Supreme Court reaffirmed this principle recently in *Oracle* by assigning great weight to the fact that the challenged use enabled "new applications," even though it may deprive the plaintiff of significant profit and

market opportunities. 593 U.S. at 39. The Second Circuit held the same in *Google Books*, 804 F.3d at 224 (no cognizable harm despite the fact that "the snippet function can cause *some* loss of sales" by conveying "historical fact[s]" "not protected by the copyright").

The Copyright Act identifies four factors the Court must consider in the fair use analysis:

> (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. These factors are "not exhaustive" and "some factors may prove more important in some contexts than in others." *Google*, 593 U.S. at 19. Application of these factors "requires judicial balancing, depending upon relevant circumstances, including 'significant changes in technology.'" *Id.* (quoting *Sony*, 464 U.S. at 430).

V. ARGUMENT

Any use by OpenAI of the Asserted Works to pretrain its LLMs was a fair use as a matter of law. OpenAI's use was highly transformative, did not result in the display of any meaningful portion of the Asserted Works to ChatGPT users, and caused no cognizable harm.

A. Factor 1: The purpose of OpenAI's use strongly favors fair use.

The first fair use factor considers "the purpose and character of the use." 17 U.S.C. § 107(1). "The central question it asks is whether the use 'merely supersedes the objects of the original creation (supplanting the original), or instead adds something new, with a further purpose or different character." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 525 (2023)*. The "larger the difference" between the purpose of the original and the

purpose of the copy, "the more likely the first factor weighs in favor of fair use." *Id.* A use with a further purpose is "transformative." *Id.*

The *Warhol* Court explained that transformative uses are "justified" because, "in a broad sense," they "further[] the goal of copyright, namely, to promote progress of science and the arts, without diminishing the incentive to create." *Id.* And, "[i]n a narrower sense, a use may be justified because copying is reasonably necessary to achieve the user's new purpose." *Id.*; *accord Romanova*, 138 F.4th at 114–116 (discussing justifications that have been held to be fair uses).

The first factor also considers whether the use is "of a commercial nature." 17 U.S.C. § 107(1). "The commercial nature of a use is relevant, but not dispositive." *Warhol*, 598 U.S. at 510. Indeed, "many common fair uses are indisputably commercial." *Google*, 593 U.S. at 32.

1. OpenAI's use was new, different, and highly transformative.

OpenAI's alleged use of the Asserted Works is indisputably different from the purpose for which those works were created. Plaintiffs wrote their books to be read and enjoyed by a human audience. SUF ¶ 95. OpenAI's alleged use of those books was for a fundamentally different purpose: to distill broad statistical patterns about human language, representing unprotectable linguistic rules, general concepts, and facts about the world, with the ultimate goal of creating general purpose software that is able to interpret and output natural human language. Indeed, OpenAI's research papers (published years before this litigation began) confirm that its purpose in building the GPT architecture was to "capture longer-range linguistic structure" (Gratz Decl. Ex. 6 at 2) and to improve "generalization performance" (Gratz Decl. Ex. 7 § 5). Unrebutted testimony of OpenAI witnesses in this case confirms that OpenAI's goal during the pretraining process was to model the underlying patterns of language itself, not to repackage Plaintiffs' works and display them to a human audience. *See* SUF ¶¶ 38–50. This is further

demonstrated by how OpenAI evaluates its models: by testing them on never-before-seen data. SUF ¶¶ 46–48. It is difficult to imagine a purpose more "distinct" from the original. *Warhol*, 598 U.S. at 510.

The only two courts to address the use of copyrighted books to train an LLM agreed that the use was highly transformative. In *Kadrey,* the court found that "[t]here 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." 788 F. Supp. 3d at 1044 (citation modified); *id.* at 1035. The *Bartz* court went further, finding that the use of copyrighted books to train LLMs was "spectacularly" and "quintessentially" transformative, and stating that "[t]he technology . . . was among the most transformative many of us will see in our lifetimes." 787 F. Supp. 3d at 1021–22, 1033.

OpenAI's use was justified for additional reasons. As Judge Leval recently explained in *Romanova*, courts have recognized a broad array of justifications that support a finding of fair use—including that the secondary use "rendered a valuable service to the public . . . without allowing public access to the copy." 138 F.4th at 115–18. Here, it is undisputed that OpenAI's LLMs are broadly useful to the public and specifically useful to authors. SUF ¶¶ 132–135.

Nor does OpenAI "allow[] public access" to the copies it makes during the training process. *Romanova*, 138 F.4th at 115–18. OpenAI does not make its training data accessible to the public. SUF ¶ 34. Moreover, regurgitation of copyrighted works is exceedingly rare. OpenAI's expert found that only 0.00007% of a sample of 20 million chats contained verbatim or near-verbatim text from any of the Asserted Works. SUF ¶ 82. Plaintiffs' own expert conducted a study (using a different methodology and a different, but partially overlapping, sample) that arrived at a similarly infinitesimally small number of outputs (0.00003%) that

allegedly contained text from Asserted Works. SUF ¶ 90. Even when Plaintiffs' expert attempted to force OpenAI's LLMs to regurgitate text from the Asserted Works by prompting the model via the API more than 5.3 million times—using sophisticated and targeted techniques that do not work in ChatGPT and that no user would ever employ—he could only force regurgitation in 0.53% of these attempts, and the longest contiguous block of allegedly regurgitated text he was able to coax out of the model in those attempts was approximately 0.6% of "A Game of Thrones." SUF ¶¶ 85–89.

Isolated and unintended instances of a tiny amount of "regurgitated" content are not sufficient to create a genuine issue of material fact regarding the purpose and character of OpenAI's use. In *Google Books*, Google intentionally displayed three different three-line "snippets" of verbatim text from millions of books. 804 F.3d at 209–10. Those snippets could be combined to reveal up to 16% of the book's text. *Id.* at 223. The Court found that even this ***intentional*** display of copyrighted text was transformative because Google put safeguards in place to ensure that the technology served its purpose but did not reveal sufficient text to "threaten the author's copyright interests." *Id.* at 218. Similarly, the court in *Kadrey* found that Meta's LLMs were transformative where, given the "mitigations" Meta put in place, plaintiffs' experts could get the models to regurgitate no more than 50 words from plaintiffs' books. 788 F. Supp. 3d at 1041–42.

Here, it is undisputed that OpenAI put into place extensive measures to prevent any meaningful amount of regurgitation from copyrighted works to ChatGPT. Those measures are highly effective. SUF ¶¶ 58–81. Regardless of the precise amount of regurgitation for any particular work, it is indisputable that OpenAI's LLMs are not designed to repackage protected expression from the Asserted Works and OpenAI's LLMs cannot be used to "read or otherwise

meaningfully access plaintiffs' books." *Kadrey*, 788 F. Supp. 3d at 1042; *Google Books*, 804 F.3d at 222 (noting that "only small and randomly scattered portions of a book will be accessible").[13]

Moreover, Plaintiffs cannot dispute that OpenAI's LLMs are useful for an enormous range of tasks that have nothing to do with books, which further weighs in favor of fair use. *See Google*, 593 U.S. at 30 (factor one favored fair use where Android platform could be used for a wide variety of purposes); *Kadrey*, 788 F. Supp. 3d at 1045 (factor one favored fair use where Meta's LLM was a "tool that can generate a wide range of text" and that "[a]ny person can use that tool to help them create further expression"); *see also Google Books*, 804 F.3d at 209 (factor one favored fair use where Google's tool "ma[de] possible new forms of research").

2. OpenAI's downloading was part of a single transformative use.

Plaintiffs wrongly argue that the Court should analyze OpenAI's downloading of books separate from its use of those books to pretrain its LLMs. Section 107 of the Copyright Act, however, requires courts to analyze "the purpose and character of the use." 17 U.S.C. § 107(1). Where the ultimate purpose is transformative, the entire course of conduct undertaken in service of that transformative purpose—including intermediate copying—is fair use. It is undisputed that OpenAI downloaded books for the sole purpose of acquiring content to pretrain the LLMs that would become GPT-3 and GPT-3.5. SUF ¶ 104. To the extent OpenAI made digital copies of the books when it acquired them or at any other stage of the process through which it pretrained its LLMs, those copies are justified by the same transformative purpose discussed above.

[13] To the extent Plaintiffs assert that any particular output from OpenAI's LLMs displays enough protected expression from the Asserted Works to be considered a substantially similar infringing work, that assertion would require a separate and individualized analysis of copyright infringement in an output. The Court need not resolve any such claim to find that OpenAI's alleged use of the Asserted Works for pretraining its LLMs was fair use as a matter of law.

Google Books compels this conclusion. There, the plaintiffs argued that "Google has scanned over four million in-copyright print books into e-books and has permanently stored multiple copies of those e-books on its servers. This use—mechanical conversion—is not transformative." Brief for Plaintiffs-Appellants in *Google Books*, 2014 WL 1509692, at *32. The Second Circuit rejected plaintiffs' approach. It did not analyze whether the act of scanning books was transformative in isolation. Rather, the court analyzed the overall purpose for which Google scanned those books: to build a transformative new technology. Because ***that*** purpose was different from the purpose for which plaintiffs wrote the books, the court held that Google's use—including the initial act of scanning—was fair use as a matter of law. 804 F.3d at 214–19.[14] So too here. OpenAI's downloading of books from the Internet served the same transformative purpose discussed above, and fair use applies for all the same reasons.[15]

The *Kadrey* court reached a similar conclusion. Like Plaintiffs here, plaintiffs there were book authors who alleged that Meta downloaded their books from Library Genesis. 788 F. Supp. 3d at 1047–48. The court held that plaintiffs were "wrong" to argue that Meta's downloading of the books "must be considered wholly separate" from training. Citing *Google Books* and *Warhol*, the court held that Meta's "downloading must still be considered in light of its ultimate, highly

[14] Plaintiffs' argument that the act of acquiring a work must be analyzed separately from the subsequent transformative use is irreconcilable with other precedent. *See* Plaintiffs-Appellees' Brief in *Sony Comput. Ent. Inc.*, *v. Connectix Corp.*, No. 99-15852, 1999 WL 33623858, at *8 (9th Cir. June 24, 1999) (arguing that defendant "found and downloaded an illicit and unauthorized copy" of the work); *Connectix*, 203 F.3d at 606–07 (conducting holistic analysis of defendant's purpose and finding that the use was "transformative" because defendant created "a wholly new product"); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1164 (9th Cir. 2007) (factor one favored fair use even though the defendant acquired and used "thumbnail images derived from infringing websites" for search engine service).

[15] Plaintiffs suggest that they also assert a copyright claim based on "Defendants' distribution of pirated books between themselves and to third parties (including via transfers of pirated training data, scraped and crawled data, and sending books to contractors)." ECF 1658 (Joint Letter) at 2. Plaintiffs failed to plead this claim in their Consolidated Class Action Complaint. *See* ECF 183. Setting this failure aside, the alleged "distribution" plaintiffs complain about, to the extent it occurred at all, served the same transformative purpose discussed above. OpenAI reserves the right to further address this claim if and when Plaintiffs attempt to raise it.

transformative purpose: training Llama." *Id.* "Because Meta's ultimate use of the plaintiffs' books was transformative, so too was Meta's downloading of those books." *Id.* at 1048.

The *Bartz* court reached a different conclusion, but on materially different facts. There, the court cited internal Anthropic documents stating that Anthropic downloaded books to build a "general purpose library." 787 F. Supp. 3d at 1016, 1034. Unlike Anthropic, however, there is no evidence that OpenAI downloaded books to build a "general purpose library." SUF ¶ 104.[16]

3. Commerciality Does Not Tilt This Factor Against Fair Use.

The first factor also requires consideration of "whether [the defendant's] use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1). But the fact that a business is "conducted for profit" does not "carr[y] presumptive force against a finding of fairness." *Campbell*, 510 U.S. at 584; *Google*, 593 U.S. at 32 ("[M]any common fair uses are indisputably commercial."). Indeed, both the Supreme Court and the Second Circuit have affirmed findings of fair use in favor of some of the largest and most profitable companies in history. *See, e.g.*, *Google*, 593 U.S. at 38 (affirming fair use even though Google's use was commercial and "helped Google make a vast amount of money"); *Swatch*, 756 F.3d at 83 (affirming fair use even though "Bloomberg is a commercial enterprise" that distributed plaintiffs' work via a "subscription service available to paying users").

OpenAI did not directly "exploit[]" the Asserted Works for "commercial advantage" by republishing them for a profit. *Am. Soc'y for Testing & Materials v. UpCodes, Inc.*, 172 F.4th 253, 266 (3d Cir. 2026). Rather, any use of those works was as an infinitesimal portion of a massive training data corpus that is an input to the very first stage of the incredibly complex

[16] To the extent *Bartz* suggests, in *dicta*, that downloading books from Library Genesis can never be part of a fair use regardless of the purpose of the download, that suggestion is wrong for all the reasons explained above.

model-development process. SUF ¶¶ 23-29, 136; *Swatch*, 756 F.3d at 83 (assigning "little weight" to commerciality because "it would strain credulity to suggest" that the copying "more than trivially affected the value of [defendant's] service"). In any event, "in light of the inherently transformative" nature of OpenAI's use, commerciality does not tilt this factor against fair use. *Google*, 593 U.S. at 32.

B. Factor 2: OpenAI used non-expressive aspects of previously published works.

The second factor considers "the nature of the copyrighted work[.]" 17 U.S.C. § 107(2). Works that are highly creative are closer to the "core" of copyright protection than works that are mostly factual. *Campbell*, 510 U.S. at 586. Additionally, fair use is more likely for works that are published than works that are unpublished. *Blanch*, 467 F.3d at 256.

The Asserted Works are published (some many years ago), which favors fair use. SUF ¶ 100. And, while OpenAI does not dispute here that they contain creative expression, OpenAI did not copy the works for that expression. Rather, as explained above, any copying of the works as part of OpenAI's pretraining process was done to access unprotectable learnings from and elements in those and millions of others works, including broad linguistic patterns and rules, facts, and common idioms. SUF ¶¶ 60–61. *See* 17 U.S.C. § 102(b); *Google Books*, 804 F.3d at 209 (copyright interests not implicated by service that extracts "word frequencies, syntactic patterns, and thematic markers" and "information on [] nomenclature, linguistic usage, and literary style"). Copying to access and make use of these elements does not weigh against fair use. *Google*, 593 U.S. at 28 (finding fair use where copyrightable elements were "bound together with uncopyrightable ideas"); *see also Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992), *as amended* (Jan. 6, 1993) (finding fair use where defendant copied "expressive

elements of the work that must necessarily be used as incident to expression of the underlying ideas, functional concepts, or facts"). In any event, "[t]he second factor has rarely played a significant role in the determination of a fair use dispute." *Google Books*, 804 F.3d at 220; *accord Lennon v. Premise Media Corp.*, 556 F. Supp. 2d 310, 325 (S.D.N.Y. 2008).

C. Factor 3: Any use by OpenAI was reasonable in amount.

The third factor considers "the amount and substantiality of the portion used in relation to the copyrighted work as a whole." 17 U.S.C. § 107(3). "[T]he extent of permissible copying varies with the purpose and character of the use," and favors fair use if the amount used is "reasonable in relation to the purpose of the copying." *Campbell*, 510 U.S. at 586–87; *accord Bill Graham*, 448 F.3d at 613. The key inquiry under factor three 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 for which it may serve as a competing substitute." *Google Books*, 804 F.3d at 222 (emphasis in original).

OpenAI's alleged use of entire books for pretraining was reasonable in relation to its transformative purpose. Courts have repeatedly recognized that "copying the entirety of a work is sometimes necessary to make a fair use" where it is "tailored to further [a] transformative purpose." *Bill Graham*, 448 F.3d at 613; *HathiTrust*, 755 F.3d at 98. It is undisputed that pretraining the LLMs at issue required use of enormous volumes of text to distill broad patterns that emerge across hundreds of billions or trillions of words. Given the need for such an enormous volume of text, use of entire books is reasonable. *See Kadrey*, 788 F. Supp. 3d at 1050 ("[I]t was 'reasonably necessary' for Meta to 'make use of the entirety of the works'" to train its LLMs.); *Bartz*, 787 F. Supp. 3d at 1030 ("Because using so many works was reasonably necessary, using any one work for *actually training* LLMs was about as reasonable as the next."

(emphasis in original)).

Further demonstrating reasonableness, OpenAI implements extensive technical measures to prevent its LLMs from displaying copyrighted works to the public. *See Google Books*, 804 F.3d at 222 (discussing the "limitations" and "restrictions" that Google "built into the program"). In the rare instances where OpenAI's LLMs regurgitate any amount of a copyrighted work, the amount of text shown to users is minimal. For example, Plaintiffs' expert purports to have prompted ChatGPT to produce "verbatim" excerpts of the Asserted Works, but it took him over 5.3 million attempts to do so, and he was only able to generate three excerpts that exceeded 300 contiguous words,[17] all of which came from a massively popular George R.R. Martin book. SUF ¶¶ 85–86. The longest excerpt was a 1,899-word excerpt from "A Game of Thrones," which has 305,507 words—*i.e.*, approximately 0.6% of the book—and the identified excerpt is readily available elsewhere on the internet, for example, in the Amazon Kindle sample for the book. SUF ¶¶ 87–89. Given the time, effort, and expert knowledge needed to obtain such a tiny amount of text from a book used in pretraining, no ordinary ChatGPT user would ever do this, and it would be much cheaper and easier for the user to simply buy the book or read it elsewhere on the Internet. *See Google Books*, 804 F.3d at 224 ("Especially in view of the fact that the normal purchase price of a book is relatively low in relation to the cost of manpower needed to secure an arbitrary assortment of randomly scattered snippets, we conclude that the snippet function does not give searchers access to effectively competing substitutes.").

D. Factor 4: OpenAI's use did not harm the market for the Asserted Works.

Factor four considers the effect the copying has on the "market for or value of the

[17] This is the equivalent of about one page. *See* E-book Experts, "How Many Words Per Page in a Book," Books Publishing Help, May 17, 2025, available at <https://bookspublishinghelp.com/blog/how-many-words-per-page>.

copyrighted work." 17 U.S.C. § 107(4). The "only harm" that matters for purposes of factor four is "the harm of market substitution." *Campbell*, 510 U.S. at 593. The first and fourth factors are closely linked. The more transformative the purpose, "the less likely it is that the copy will serve as a satisfactory substitute for the original." *Google Books*, 804 F.3d at 223. Thus, courts ask whether the defendant's copying "results in widespread revelation of sufficiently significant portions of the original as to make available a significantly competing substitute." *Id.*

1. OpenAI's use is not substitutive as a matter of law.

OpenAI's alleged use does not "substitute" for the Asserted Works. As discussed above, pretraining an LLM is fundamentally different from reading a book. And despite Plaintiffs' complaints, outputs from OpenAI's LLMs do not reveal a meaningful amount of protected expression from the Asserted Works to ChatGPT users.

Factor four weighs even more strongly in favor of fair use here than it did in *Google Books*. Unlike Google's snippet view, which was designed to display up to three different three-line excerpts of verbatim text from copyrighted books, 804 F.3d at 209–10, OpenAI's LLMs are not designed to display ***any*** verbatim text from copyrighted works. *See* SUF ¶¶ 65-66, 80, 105. And whereas plaintiffs' experts in Google Books were able to aggregate individual snippets to display 16% of a book, *id.* at 223, Plaintiffs' experts here were unable to generate anything close to that amount for any book. SUF ¶¶ 85-88. Given these undisputed facts, OpenAI's LLMs could not possibly be considered "an effectively competing substitute for Plaintiffs' books." *Google Books*, 804 F.3d at 222.

Plaintiffs claim that ChatGPT can produce "summaries" of their books or "outlines" for a sequel or new chapter. ECF 183 at ¶¶ 104, 177–78, 191–92, 202–03, 213–14, 235–36, 257–58, 268–69. Even if some summaries and outlines could be considered infringing works (which

Plaintiffs have not shown), they are not "effectively competing substitutes" for purchasing the original books or derivatives of them. The only summaries Plaintiffs have identified to date are short, high-level factual recitations of the plot and main characters of a book. They are no different from countless other book reviews or summaries available for free online from sources like Wikipedia (https://www.wikipedia.org/) and Goodreads (https://www.goodreads.com/), with no apparent objection from Plaintiffs. SUF ¶ 102. It is implausible that someone who wanted to read, for example, Jonathan Franzen's 576-page novel *The Corrections* would choose to read a short summary instead. Indeed, Mr. Franzen testified in his deposition that he [redacted]. SUF ¶¶ 108, 119. And Mr. Martin's personal website links to other websites that contain chapter-by-chapter summaries of some of his books. SUF ¶ 109. The same is true for "outlines" of a new book or new chapters of an existing book. There is no evidence whatsoever that these user-prompted outlines have caused any harm to the market for or value of the Asserted Works.

Regardless, the *Google Books* court expressly recognized that "the snippet function can cause *some* loss of sales," explaining that "[t]here are surely instances in which a searcher's need for access to a text will be satisfied by the snippet view[]" thereby causing a lost sale. 804 F.3d at 224. "But the possibility, or even the probability or certainty, of some loss of sales does not suffice to make the copy an effectively competing substitute that would tilt the weighty fourth factor in favor of the rights holder in the original." *Id.* "There must be a ***meaningful*** or ***significant*** effect 'upon the potential market for or value of the copyrighted work.'" *Id.* (quoting 17 U.S.C. § 107(4)). Here, there is no evidence of any effect, let alone a meaningful or significant one.

2. Plaintiffs' "dilution" theory of harm is not legally cognizable.

Plaintiffs claim they have been or will be harmed because prospective authors could use OpenAI's LLMs either to write or help write new books, in similar or even unrelated genres, that compete with Plaintiffs' books. According to Plaintiffs, "AI-generated" books will harm them by "diluting" the market or "indirectly substituting" for their own books. There is no evidence to support these claims, given that multiple Plaintiffs have admitted that they are unaware of any ChatGPT output that [redacted] for their Asserted Works. SUF ¶¶ 110-112. Even if there were evidence of "indirect" substitution, it would not matter because alleged harm from new works that do not reproduce substantial protected expression from the original is not cognizable as a matter of law.

The "only harm" cognizable under copyright law is harm from substitution ***for the expressive content*** of the original. *Campbell*, 510 U.S. at 593; Patry on Copyright § 10.150 ("[E]ven though the taking of unprotectable material such as important ideas, laboriously researched facts, or the copying of a work's overall style may cause lower sales, such losses should not be considered. The harm must be caused by the use of expression."). Alleged "harm" from other sources—whether called "dilution," "indirect substitution," or any other name—is not the type of "harm" that copyright law protects against. *HathiTrust*, 755 F.3d at 99 ("[U]nder Factor Four, any economic 'harm' caused by transformative uses does not count because such uses, by definition, do not serve as substitutes for the original work."); *Sega*, 977 F.2d at 1524 (Any "indirect" effect from a new work that does not reproduce protected expression from the original is not cognizable harm given the "statutory purpose of promoting creative expression").

Google Books is instructive. There, the court acknowledged that snippet view could cause authors to lose sales by providing "a historical fact" that might "satisfy a searcher's need for

access to a copyrighted book . . . eliminating any need to purchase it or acquire it from a library." 804 F.3d at 224. But the court held that this harm was legally irrelevant because it "will generally occur in relation to interests that are not protected by . . . copyright." *Id.* Facts are not copyrightable. *Id.* So "harm" that occurs when the defendant reproduces facts is not cognizable under factor four. *Google Books*, 804 F.3d at 224; *accord Wright*, 953 F.2d at 739.

The same rule applies to alleged harm from new books ("AI-generated" or otherwise) that do not reproduce protected expression from the original. Plaintiffs' copyrights in the Asserted Works protect only their "manner of expression." *Google Books*, 804 F.3d at 224. They do not prohibit others from writing books that express ideas in new and different ways. *See id.* Thus, even if it were true that consumers were choosing to purchase "AI-generated" books instead of the Asserted Works (which they are not), Plaintiffs' claims still would fail because those books do not reproduce Plaintiffs' expression. Plaintiffs have no more right to claim harm from non-infringing "AI-generated" books than they do to claim harm from new non-infringing books written by humans who have read Plaintiffs' books and been inspired by them. *See id.*; *Bartz*, 787 F. Supp. 3d at 1031–32; *Sony Comput. Ent. Am., Inc. v. Bleem, LLC*, 214 F.3d 1022, 1029 (9th Cir. 2000), *amended on denial of reh'g* (July 10, 2000) (holding that even the "los[s] [of] a significant share of [the plaintiff's] present market" is not cognizable if it resulted "not from the display of plaintiff's [expressive material] but from commercial competition with a work that does not in any way make use of plaintiff's copyrighted material"); *NXIVM Corp. v. Ross Institute*, 364 F.3d 471, 482 (2d Cir. 2004) ("[T]he relevant market effect with which we are concerned is the market for plaintiffs' 'expression,' and thus it is the effect of defendants' use of that expression on plaintiffs' market that matters, not the effect of defendants' work as a whole."). Indeed, Plaintiffs admit that their own books were inspired by other books they have

read. As Mr. Baldacci testified, "[redacted]" SUF ¶ 120; *see also Campbell*, 510 U.S. at 575 ("Every book in literature, science and art, borrows, and must necessarily borrow, and use much which was well known and used before.").

Plaintiffs' "dilution" or "indirect substitution" theory of harm would turn copyright law on its head by transforming it from a law that *incentivizes the creation* of new expression into a law that *prohibits competition from* new expression. Numerous appellate courts, including the Second Circuit, have rejected attempts to misuse copyright law to prevent general competition in the marketplace. *See Bleem*, 214 F.3d at 1029; *Sega*, 977 F.2d at 1523 ("[A]n attempt to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression and cannot constitute a strong equitable basis for resisting the invocation of the fair use doctrine."); *Connectix*, 203 F.3d at 607 (loss of sales from "a legitimate competitor in the market" does not weigh against the defendant); *Triangle Publ'ns, Inc. v. Knight-Ridder Newspapers, Inc.*, 626 F.2d 1171, 1177 (5th Cir. 1980) (holding that loss of market share from "commercial competition with a work that does not in any way make use of plaintiff's copyrighted material" does not weigh against fair use).

The *Bartz* and *Kadrey* decisions address a "dilution" theory of harm under Ninth Circuit law. In *Bartz*, the court flatly rejected it. 787 F. Supp. 3d at 1031–32. The court "assume[d]" for the purposes of argument that "training LLMs will result in an explosion of works" that competed with the author plaintiffs by "creating alternative summaries of factual events, alternative examples of compelling writing about fictional events, and so on." *Id.* It then held: "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." *Id.*

In *Kadrey*, the court granted summary judgment on the defendant's fair use defense, but suggested, in *dicta*, that the "dilution" theory could be "relevant" in a case where the plaintiff presented sufficient evidence of this type of "harm." 788 F. Supp. 3d at 1052–54. As explained below, Plaintiffs here have presented no such evidence, but even if they had, it would not matter because this aspect of *Kadrey* is wrong under Second Circuit law. Alleged "harm" from new, non-infringing works that do not reproduce protected expression from the original is non-cognizable. *See*, *e.g.*, *Google Books*, 804 F.3d at 224–25. A non-cognizable harm does not become cognizable simply because there is the potential for more of it. *See id.; Campbell*, 510 U.S. at 591 ("We do not, of course, suggest that a parody may not harm the market at all, but when a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act."); *see also* Patry on Copyright § 10:155.40 (criticizing dilution theory as "nuts" and a "fringe piece of advocacy").

3. Plaintiffs have suffered no meaningful harm.

Because (1) no reasonable factfinder could find that OpenAI's use created a meaningful substitute for the Asserted Works, and (2) Plaintiffs' "dilution" theory is not legally cognizable, the Court need go no further to resolve this factor in favor of fair use. If, however, the Court considers Plaintiffs' legally incorrect theories of harm, it should reject them because the undisputed evidence shows that Plaintiffs have not been harmed by OpenAI's alleged use.

To establish lack of harm, OpenAI's economic expert, Dr. Gregory Leonard, conducted a detailed analysis of sales data for the Asserted Works. SUF ¶ 125. He found no systematic decrease in sales trend after the launch of OpenAI's products. SUF ¶ 126. Dr. Leonard also analyzed pricing data for certain Asserted Works that met his criteria and found that OpenAI's

products had no statistically significant impact. SUF ¶ 127. Lastly, Dr. Leonard constructed a control-group of books that were not used to train OpenAI's LLMs and compared sales trends for those books to the sales trends for over 100 of the Asserted Works.[18] SUF ¶ 128. He found no statistically significant difference. SUF ¶129. Plaintiffs complain about Dr. Leonard's analysis, but offer no empirical analysis of their own. SUF ¶ 130. Plaintiffs' own factor-four expert admitted that he did not have the "[redacted]." SUF ¶ 131.

Instead, Plaintiffs merely point to the existence of supposed "AI-generated" books on Amazon. Plaintiffs provide no evidence that anyone has bought any of these books instead of the Asserted Works. SUF ¶¶ 110-119; 130-131. This is no surprise. Plaintiffs are well-known authors, many of whom have written best-selling books and made millions of dollars selling those books. Demand for their books and other "trade" books—especially those published by major publishers—is driven by factors like author reputation, marketing, and word-of-mouth. It defies common sense to claim that Plaintiffs face a meaningful risk of significant "harm" from little-known AI-generated books (none of which copy Plaintiffs' expression) that exist (along with other self-published books) on Amazon. *See Kadrey*, 788 F. Supp. 3d at 1052 ("It seems unlikely . . . 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.").

Plaintiffs also claim that they were harmed because OpenAI downloaded books from Library Genesis instead of buying those books from Plaintiffs. This claim fails for at least three reasons. **First**, the mere fact that OpenAI did not buy the books is insufficient and demonstrates nothing. "[B]y definition every fair use involves some loss of royalty revenue because the

[18] The remainder of the Asserted Works did not meet Dr. Leonard's criteria because, among other things, some of the books did not have sufficiently high net sales.

secondary user has not paid royalties." *Bill Graham*, 448 F.3d at 614 (quoting Pierre N. Leval, Toward a Fair Use Standard, 103 Harv. L.Rev. 1105, 1111 (1990)). Factor four asks whether there is a cognizable market harm beyond the mere act of the copying. **Second**, there is no evidence that OpenAI would have purchased new books from Plaintiffs had it not downloaded copies from Library Genesis. OpenAI could have acquired the same books from libraries (as Google did in *Google Books*) or used booksellers or not downloaded them from Library Genesis. **Third**, even if a "lost sale" to the defendant for the use at issue was a valid form of "harm," it is not significant enough to tilt the scales in the factor four analysis in this case. Factor four "focuses on whether the copy brings to the marketplace a competing substitute for the original, or its derivative, so as to deprive the rights holder of ***significant*** revenues." *Google Books*, 804 F.3d at 223. A single "lost" book sale "does not suffice to make the copy an effectively competing substitute that would tilt the weighty fourth factor in favor of the rights holder in the original. There must be a meaningful or significant effect[.]" *Id.* at 224. There is no such effect here. The factor-four "widespread conduct" inquiry does not change the result. *See Campbell*, 510 U.S. at 590. For the same reason Plaintiffs have suffered no significant harm from OpenAI's alleged use, Plaintiffs would suffer no significant harm from other, similar (transformative) uses.

4. Plaintiffs' licensing theory of harm is not legally cognizable.

Plaintiffs attempt to salvage their claim by arguing that OpenAI should have paid to license their books for pretraining its LLMs. This circular argument assumes the conclusion. "If the use is otherwise fair, then no permission need be sought or granted." *Campbell*, 510 U.S. at 585 n.18; *Google*, 593 U.S. at 38 (quoting Nimmer on Copyright § 13.05[A][4] and "cautioning against the 'danger of circularity posed' by considering unrealized licensing opportunities because 'it is a given in every fair use case that plaintiff suffers a loss of a potential market if that

potential is defined as the theoretical market for licensing the very use at bar").

Indeed, both *Kadrey* and *Bartz* rejected this argument. *Kadrey*, 788 F. Supp. 3d at 1052 ("[W]hether 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."); *Bartz*, 787 F. Supp. 3d at 1032 ("[S]uch a market for that use is not one the Copyright Act entitles Authors to exploit").

Plaintiffs are not entitled to a licensing fee for a use that is "transformatively different from their original expressive purpose." *Bill Graham*, 448 F.3d at 614; *see also* Leval, Fair Use Rescued at 1465 ("Market harm must be assessed in relation to the purpose and character of the use"). Alleged harm from a "lost" license fee in a "transformative market" is not cognizable as a matter of law. *Bill Graham*, 448 F.3d at 614–15.

Even if the Court were to consider alleged "harm" of this type, it would not change the result because the purported "market" in which plaintiffs claim to be harmed is not a "traditional, reasonable, or likely to be developed" market for their books. *Id.* at 614. Plaintiffs have [redacted]. SUF ¶ 121; *Blanch*, 467 F.3d at 250, 258 (finding that factor four "greatly favor[ed]" defendant where plaintiff had "never licensed any of her photographs for use in works of graphic or other visual art"). They assert that they [redacted]. SUF ¶ 124; *White v. West Pub. Corp.*, No. 12 Civ. 1340 (JSR), 2014 WL 3385480, at *3 (S.D.N.Y. July 11, 2014) ("[N]o secondary market exists in which White could license or sell the briefs to other attorneys, as no one has offered to license any of White's motions, nor has White sought to license or sell them."). And it is highly unlikely that anyone ever will, because the purported "market" that Plaintiffs posit makes little economic sense in

reality. OpenAI's LLMs are pretrained on hundreds of billions to tens of trillions of tokens.[19] SUF ¶ 23. Any given work is an infinitesimally small part of the whole, and the transaction costs for licensing any individual works (or even groups of works) would vastly exceed the marginal benefit of those works. SUF ¶¶ 23-29, 136; *see White*, 2014 WL 3385480, at *3 ("Although White argues that Lexis and West impede a market for licensing briefs, the Court finds that no potential market exists because the transaction costs in licensing attorney works would be prohibitively high."). Regardless, Plaintiffs "cannot prevent others from entering fair use markets merely by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work." *Bill Graham*, 448 F.3d at 614–15.[20]

5. OpenAI's use unlocks substantial public benefits.

Factor four also considers the "public benefits the copying will likely produce." *Google*, 593 U.S. at 35. "This analysis requires a balancing of the benefit the public will derive if the use is permitted and the personal gain the copyright owner will receive if the use is denied." *Bill Graham*, 448 F.3d at 613. This strongly favors fair use. ChatGPT provides innumerable public benefits, but most relevant here are benefits that Plaintiffs and other authors get from their use of it.

Plaintiffs do not dispute that "[redacted]." SUF ¶ 132. They admit that [redacted]. SUF ¶ 133. And Plaintiffs' own expert concedes that ChatGPT "[redacted]

[19] Each token can represent one or part of a word.

[20] OpenAI expects that Plaintiffs will invoke certain commercial partnerships that OpenAI and/or other entities have entered to argue that there is a market for LLM training licenses that has been harmed by OpenAI's failure to license their works. Nothing about that argument can or does change the analysis presented here, and OpenAI will address it in due course.

[redacted] SUF ¶ 135.

Promoting the progress of science and the arts through the dissemination of new works is the purpose of copyright law. *See*, *e.g.*, *Google*, 593 U.S. at 39. OpenAI's LLMs fulfill that very purpose, in addition to providing many other benefits. Given the significance of the public benefits and the lack of any cognizable harm to the Plaintiffs, factor four favors fair use. *See id.*

VI. CONCLUSION

For the foregoing reasons, OpenAI respectfully requests that the Court grant this motion.

Dated: September 4, 2026

Respectfully submitted,

LATHAM & WATKINS LLP
Andrew M. Gass (*pro hac vice*)
andrew.gass@lw.com
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
Telephone: 415.391.0600

Sarang V. Damle
sy.damle@lw.com
Luke A. Budiardjo
luke.budiardjo@lw.com
1271 Avenue of the Americas
New York, NY 10020
Telephone: 212.906.1200

Elana Nightingale Dawson (*pro hac vice*)
elana.nightingaledawson@lw.com
555 Eleventh Street, NW, Suite 1000
Washington, D.C. 20004
Telephone: 202.637.2200

MORRISON & FOERSTER LLP
Joseph C. Gratz (*pro hac vice*)
jgratz@mofo.com
Tiffany Cheung (*pro hac vice*)
tcheung@mofo.com
Caitlin Sinclaire Blythe (*pro hac vice*)
cblythe@mofo.com
425 Market Street
San Francisco, CA 94105
Telephone: 415.268.7000

Lily Li (*pro hac vice*)
YLi@mofo.com
755 Page Mill Road
Palo Alto, CA 94304
Telephone: 650.813.5600

KEKER, VAN NEST & PETERS LLP
Robert A. Van Nest (*pro hac vice*)
rvannest@keker.com
R. James Slaughter (*pro hac vice*)
rslaughter@keker.com
Paven Malhotra
pmalhotra@keker.com
Michelle S. Ybarra (*pro hac vice*)
mybarra@keker.com
Reid P. Mullen (*pro hac vice*)
rmullen@keker.com
Nicholas S. Goldberg (*pro hac vice*)
ngoldberg@keker.com
633 Battery St.
San Francisco, CA 94111
Telephone: 415.391.5400

Attorneys for OpenAI

CERTIFICATE OF COMPLIANCE

In accordance with Local Civil Rule 7.1(c), I hereby certify that the foregoing OPENAI'S MEMORANDUM OF POINTS AND AUTHORITIES IN SUPPORT OF MOTION FOR SUMMARY JUDGMENT is 12,394 words, exclusive of the caption page, table of contents, table of authorities, and signature block. The basis of my knowledge is the word count feature of the word-processing system used to prepare this memorandum.

Dated: September 4, 2026 By: