2024

A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence

Art Neill
California Western School of Law, aneill@cwsl.edu

James Thomas

Erika Lee
California Western School of Law, ELee@cwsl.edu

Recommended Citation

Art Neill, James Thomas & Erika Lee, *A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence*, 32 Tex. Intell. Prop. L.J. 225 (2024).
Available at: https://scholarlycommons.law.cwsl.edu/fs/451

A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence

Art Neill,[*] James Thomas[†] & Erika Lee[‡]

Abstract

The rise in the popularity of consumer-facing generative artificial intelligence (GenAI) has created considerable confusion and consternation among some copyright owners. The ability to automate the generation of original works based on user input is considered by some copyright holders to have been made possible by large-scale direct infringement by OpenAI, Microsoft, and other major GenAI developers. This article explores the application of copyright law to the training of OpenAI's ChatGPT, specifically focusing on the legal issues surrounding the unauthorized use of copyrighted textual works in the GenAI training process.

The large language models (LLMs) that drive ChatGPT and similar GenAI can summarize written works, generate movie scripts, write poetry, and compose stories nearly instantaneously. LLMs can only function in this way due to the use of vast, diverse training datasets comprised of billions of websites and expansive repositories

[*] Art Neill is an Associate Clinical Professor of Law and the Director of the New Media Rights Program at California Western School of Law, teaching Internet & Social Media Law, as well the Internet & Media Law Clinic. Art practices, educates, and advocates in the areas of intellectual property, privacy, and media law. He has testified and provided regulatory comments before the United States Copyright Office and the Federal Communications Commission. Art previously served on the FCC's Consumer Advisory Board. Art is the co-author of the book *Don't Panic :) A Legal Guide (in plain English) for Small Businesses and Creative Professionals*. He has written about legal issues in copyright and internet law through numerous scholarly articles, op-eds, and as a Contributor for creators and small businesses for Forbes. He is also led creation of the Fair Use App, which teaches concepts surrounding fair use and content reuse. Learn more about Art at https://www.newmediarights.org/about_us/nmr_staff/art_neill.

[†] James Thomas is a *magna cum laude* graduate of California Western School of Law, where he served as the Executive Director of Notes & Comments for the *Law Review*. James is deeply grateful to his co-authors for their invaluable collaboration and support. He also extends his heartfelt thanks to his wife, whose unwavering love and encouragement have been instrumental in his success. This bio was written by ChatGPT.

[‡] Erika Lee is the Assistant Director of the New Media Rights program and an Adjunct Professor at California Western School of Law, teaching the Internet & Media Law Clinic. Erika provides transactional preventative legal services, develops educational resources, engages in policy advocacy in regulatory proceedings at the Copyright Office for New Media Rights, and contributes to copyright legal scholarship, including an article previously published in the Texas Intellectual Property Law Journal (Art Neill & Erika Lee, *Fixing Copyright Registration for Online Video Creators*, 28 TEXAS INTELL. PROP. L. J. 87 (2019)).

of books. These datasets are analyzed to study the functionality and syntax of the language, allowing the LLMs to generate new works.

This article discusses the recent lawsuits launched by high-profile authors and copyright owners against OpenAI and Microsoft, claiming direct, vicarious, and derivative infringement. Authors such as George RR Martin, Sarah Silverman, Christopher Golden, and professional organizations such as the Authors Guild contended their works were infringed upon to turn OpenAI into an $80 billion company.

In considering the merits of these lawsuits, we discuss the curation and content of training datasets used in the known iterations of ChatGPT and characterize the protectability of the different works the datasets included. We then explore whether the transitory nature of OpenAI's training process uses acceptable, non-infringing copies and how that would affect the outcome of an action for direct infringement.

The article then looks at the applicability of current fair use precedent to textual GenAI and the various types of works used in training datasets. To do so, we apply settled caselaw and leading decisions to discuss OpenAI's use of copyrighted works regarding purpose and character, nature of the original work, the amount and substantiality of the works used, and the impact on the market value of the works by ChatGPT. We pay special attention to other innovative technologies that rely on a fair use defense to draw analogies and comparisons to GenAI.

Finally, this article considers the policy and legislation of other countries and their approach to ChatGPT and copyright. In doing so, policy considerations are taken into account to argue the necessity of a finding of fair use to maintain international competitiveness and to prevent an erosion of fair use in other sectors outside of GenAI. The article concludes that there is substantial support for arguments that GenAI training involves only transitory, non-actionable copying and that it is also permissible under fair use.

Table of Contents

Introduction

The legality of companies like Open AI, Microsoft, Meta, and Google's use of existing copyrighted works to train their generative artificial intelligence (GenAI) without permission is under scrutiny in the legislature, courts, and public opinion. The rapid increase in the effectiveness and utility of generative artificial intelligence[1] and the corresponding increase in valuations[2] related to artificial intelligence (AI) is built using large datasets of copyrighted works.[3] Numerous plaintiffs, including the Authors Guild (with George RR Martin, John Grisham, and 15 other authors), Sarah Silverman, The New York Times, and others have filed lawsuits challenging artificial intelligence companies' ability to use textual works without permission to train their technology.[4] This Article focuses on copyright law issues surrounding the use of textual copyrightable works in training datasets for artificial intelligence. Specifically, we will use OpenAI and its ChatGPT technology to frame many of these issues. Part I discusses the claims in some of the leading lawsuits against OpenAI. Part II discusses what kinds of copyrighted works are used in training artificial intelligence and how those works are collected and curated. Part III establishes which training inputs for generative artificial intelligence, if any, are protected. Part IV considers how training methods for ChatGPT may only use transitory copies of protected works and, therefore, may be noninfringing. Finally, Part V provides a fair use analysis, discussing: (1) the purpose and character of the use; (2) the nature of the original work; (3) the quantitative and qualitative scope of the works used; and (4) the impact of the use on the copyrightable material's market value. This analysis demonstrates that under current United States copyright law, (1) the use of

[1] For purposes of this paper, artificial intelligence will be considered under the Oxford Dictionary definition: "the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages." *Artificial intelligence*, OXFORD DICTIONARY OF PHRASE AND FABLE (2d ed. 2005).

[2] Molly Schuetz, *OpenAI Seeks $90 Billion Valuation in Possible Share Sale, WSJ Says*, BLOOMBERG (Sep. 26, 2023, 12:36 PM), https://www.bloomberg.com/news/articles/2023-09-26/openai-seeks-90-billion-valuation-in-possible-share-sale-wsj.

[3] *See* Complaint at 3, Chabon v. OpenAI, Inc., No. 23-CV-04625 (N.D. Cal. filed Sep. 8, 2023); *see also* First Consolidated Complaint at 3–4, Authors Guild v. OpenAI Inc., No. 23-CV-08292 (S.D.N.Y. filed Sep. 19, 2023); Complaint at 2, Silverman v. OpenAI, Inc., No. 23-CV-03416 (N.D. Cal. filed Jul. 7, 2023); Complaint at 2–3, New York Times Co. v. Microsoft Corp., No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023).

[4] *See* First Consolidated Complaint at 3–4, *Authors Guild*, No. 23-CV-08292; *see also* Complaint at 2, *Silverman*, No. 23-CV-03416; *see also* Complaint at 2–3, *Times*, No. 23-CV-11195.

unprotected works will be permitted, (2) certain fleeting uses of copyrighted works may be permissible and not infringement, and (3) for protected textual copyrighted works used without permission, there is likely a significant fair use defense. Part VI briefly discusses selected international approaches to GenAI and copyright and their implications for policy choices in the United States.

I. Current Lawsuits Against OpenAI and Text Generative Artificial Intelligence

There are numerous lawsuits against Open AI and other AI companies relating to their use of copyrighted materials as training inputs.[5] Companies are being sued for the use of text, software, images, and music as training inputs. Since this article is focused on the use of text works and uses OpenAI as our example technology, we will focus on the lawsuits involving OpenAI's use of textual works as training inputs. There are a few leading cases, including *Chabon v. OpenAI*, *Tremblay/Silverman v. OpenAI*, *Authors Guild v. OpenAI*, and *Sancton v. OpenAI*. In essence, the plaintiffs claim that OpenAI committed copyright infringement by using their works to train the ChatGPT language model,[6] violating the plaintiff's exclusive rights to reproduce, distribute, create derivatives, and publicly display or perform their works.[7] Although there are many commonalities in the complaints filed by various plaintiffs, there are some variations and nuances.

Cases such as *Tremblay* and *Authors Guild* suggest that large language models (LLMs) may collect and store copyrighted works. Indeed, *Tremblay*'s plaintiffs claim "the reason ChatGPT can accurately summarize a certain copyrighted book is because that book was copied by OpenAI into the underlying OpenAI Language Model (either GPT-3.5 or GPT-4) as part of its training data."[8] Essential to this claim is that "ChatGPT retains knowledge of particular works" to output similar summaries.[9] *The New York Times Company v. Microsoft Corporation* follows a similar pattern as *Tremblay* and *Authors Guild*, claiming that defendant's LLMs "were built by copying and using *millions* of The Times' copyrighted" works.[10] The Times then takes a further step, arguing that "Defendants' GenAI tools can generate output that recites *The New York Times* content verbatim, closely summarizes it, and mimics its expressive style, as demonstrated by scores of examples."[11] To support its arguments, Times introduced an exhibit allegedly showing 100 examples[12] where ChatGPT

[5] *See infra* Part II (explaining the various types of datasets used to train artificial intelligence for LLMs); *infra* Part III (discussing the level of protectability of the various inputs used in training datasets).

[6] *See* Complaint at 3, *Chabon*, 3:23-CV-04625; Tremblay v. OpenAI, Inc., No. 23-CV-03223, 2024 WL 557720, at *1 (N.D. Cal. Feb. 12, 2024); Complaint at 3–4, *Authors Guild*, 1:23-CV-08292; First Amended Complaint at 3, Sancton v. OpenAI Inc., No. 23-CV-10211 (S.D.N.Y. filed Nov. 21, 2023).

[7] *See* 17 U.S.C § 106.

[8] Complaint at 8, *Tremblay*, 2024 WL 557720 (No. 23-CV-03223).

[9] *Id.*

[10] Complaint at 2, New York Times Co. v. Microsoft Corp., No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023).

[11] *Id.*

[12] *Id.* ex. J, at 2.

provided near identical replications of copyrighted works and summaries "significantly longer and more detailed" than what is accessible through search engines.[13]

Silverman v. OpenAI may act as a signpost for how other ChatGPT and LLM proceedings will be handled by the courts. OpenAI is currently attempting to dismiss the claims brought in *Silverman*, including direct copyright infringement, vicarious copyright infringement, and 17 U.S.C. § 1202(b) Digital Millennium Copyright Act claims.[14] These same claims are common in *Chabon*, *Tremblay*, *Times*, and *Authors Guild*,[15] and, if they are resolved in *Silverman*, they could lead to a dismissal of some of the claims in similar cases against OpenAI. Silverman's request for dismissal relied on a recent decision in *Kadrey et al. v. Meta Platforms, Inc.* and seems to favor OpenAI's position.[16]

Kadrey consists of several authors (including Silverman and Chabon) claiming direct copyright infringement based on the belief that Meta copied their protectable works to train the LLaMA 1 and 2 language models.[17] The alleged direct infringement occurred in a similar fashion to the OpenAI cases, wherein Meta used a training dataset called Books3 which was allegedly "derived from a copy of the contents of . . . Bibliotik."[18] Plaintiffs contend not only that Books3 contains their copyrighted works and that the use of that dataset was infringing[19] but also that the entire language model is an infringing derivative work due to their reliance on the expressive content of works in the dataset.[20] The court in *Kadrey* granted a Motion to Dismiss, calling the claim that the language models were derivative works "nonsensical" and that there "is no way to understand the LLaMA models . . . as a recasting or adaptation" of the works included in the dataset.[21] The court addressed infringing output claims by rejecting the plaintiffs' contention that "every output . . . is an infringing derivative work."[22] The court also rejected the plaintiff's claim that they did not need to allege similarities between the outputs and original works to show derivative infringement.[23] The court noted that, to ultimately prevail on a claim of infringement, the outputs themselves must include "'in some form a portion of' the plaintiffs' books."[24] *Kadrey*

[13] *Id.* at 3.

[14] Complaint at 10–12, Silverman v. OpenAI, Inc., No. 23-CV-03416 (N.D. Cal. filed Jul. 7, 2023).

[15] *See* Complaint at 15–18, Chabon v. OpenAI, Inc., 23-CV-04625 (N.D. Cal filed Sep. 8, 2023); Complaint at 10–12, Tremblay v. OpenAI, Inc., 3:23-CV-03223, 2024 WL 557720 (N.D. Cal. Feb. 12, 2024); Complaint at 44–46, Authors Guild v. OpenAI Inc., 1:23-CV-08292 (S.D.N.Y. filed Sep. 19, 2023); Complaint at 60–64, *Times*, No. 23-CV-11195.

[16] *See* Statement of Recent Decision at 2, *Silverman*, No. 23-CV-03416.

[17] Complaint at 7, Kadrey v. Meta Platforms, Inc., No. 23-CV-03417, 2024 WL 235199 (N.D. Cal. filed Nov. 20, 2023).

[18] *Id.* at 6.

[19] *Id.*

[20] *Id.* at 7.

[21] Order Granting Motion to Dismiss, Kadrey v. Meta Platforms, Inc., No. 23-CV-03417, 2023 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023).

[22] *Id.* at 1–2.

[23] *Id.* at 2.

[24] *Id.*

has seemingly provided a resolution to the argument that entire AI technologies and their outputs are infringing derivatives simply due to the use of copyrightable training inputs. As such, contention over vicarious copyright liability and derivative works may be soon resolved, leaving the question of direct infringement of works during the training process as the key inquiry pending in further proceedings.

If the use of training datasets containing copyrightable material is found to be infringing, there will be a significant impact on current and future development of GenAI. With no way to remove specific copyrightable material from language models, many current generative AI models would have to be shut down. Language models could only be trained using works for which they had obtained permission or which are public domain. It could further consolidate the power over future artificial intelligence development, solidifying a few artificial intelligence companies' dominance due to the expense of paying for the access to and use of copyrighted material for training purposes. In addition to lawsuits, OpenAI, other AI companies, and copyright holders are monitoring regulatory proceedings such as the study at the United States Copyright Office aimed at determining key questions at the intersection of AI and copyright.[25] That study is reviewing the use of copyrightable works as training inputs and also the protectability of outputs, among other questions. Meanwhile, news organizations have "urged Congress . . . to clarify that use of copyrighted content to train large language models is not fair use."[26] The battle over use of copyrighted works is currently being fought in the court room, the legislature, and through government regulatory agencies. With courts willing to throw out claims on vicarious infringement and derivative works,[27] litigation will likely focus on direct infringement and fair use related to the training process.

II. Curation, Collection, and Categories of Copyrighted Works Used in ChatGPT Training Models

Any understanding of the application of the Fair Use Doctrine to inputs of LLMs begins with an understanding of (a) what large language models are and (b) the datasets used for training artificial intelligence. Although many different LLMs are currently in use, with many more to follow, this section will focus on OpenAI's ChatGPT, specifically its data collection and training. The following information about data collection and the functionality of ChatGPT is based on currently available literature and has been used to the best of our understanding.

LLMs are deep learning algorithms that can "recognize, translate, predict, or generate text or other content," and are usually trained on massive datasets.[28] The

[25] *See Artificial Intelligence Study*, U.S. COPYRIGHT OFFICE, https://copyright.gov/policy/artificial-intelligence/ (last visited Jan. 31, 2024, 11:22 AM).

[26] Ivan Moreno, *News Orgs Ask Congress For Copyright Clarity On AI Training*, LAW360 (Jan. 10, 2024, 8:36 PM), https://www.law360.com/media/articles/1782747?nl_pk=a9a8c0e3-f059-4c96-a0eb-dc5e21b882de&utm_source=newsletter&utm_medium=email&utm_campaign=media&utm_content=2024-01-11&read_main=1&nlsidx=0&nlaidx=1.

[27] *See* Order Granting Motion to Dismiss, Kadrey v. Meta Platforms, Inc., No. 23-CV-03416, 2024 WL 8039640, at *1 (N.D. Cal. Nov. 20, 2023).

[28] *What is a Large Language Model (LLM)?*, ELASTIC, https://www.elastic.co/what-is/large-language-

consumer-facing form of ChatGPT 4.0, the most popular LLM, began with GPT-1.[29] The dataset used to train this iteration was relatively small compared to the subsequent forms.[30] GPT-1 seems to have utilized BooksCorpus, a collection of 7,000 unpublished and self-published books[31] collected from Smashwords,[32] a repository of unpublished and self-published books.[33] GPT-1 utilized this dataset to study "large stretches of contiguous text" to train the AI on word dependencies in a large range.[34] The creation of BooksCorpus itself was done by researchers from the University of Toronto and the Massachusetts Institute of Technology.[35] While the published paper lists seven authors, it does not explicitly mention who collected the data.[36] However, reports indicate that the data was collected from scraping software, which likely generated a list of links to free versions of the ebooks and converted them from epub files to plain text files for inclusion in the corpus.[37]

GPT-2 used a much larger dataset than the previous iteration by scraping Reddit to find upvoted articles and pulling data from all outbound links in the targeted Reddit posts.[38] To build this dataset, OpenAI focused on using human curation to use only higher-quality text.[39] Other web-scraped datasets were considered too broad with "unintelligible" content, so OpenAI apparently focused on using Reddit posts with at least three karma[40] to focus on the text that was "interesting, educational, or just funny."[41] After removing all Wikipedia documents, which were already included in test sets used in training, the WebText dataset was created, containing information from over 45 million links, which were then paired down to over 8 million documents.[42]

OpenAI's third iteration, GPT-3, seems to have once again expanded the amount of information in its dataset by utilizing five corpora: Common Crawl, WebText2,

models (last visited Dec. 14, 2023).

[29] Bernard Marr, *A Short History of ChatGPT: How We Got to Where We Are Today*, FORBES (May 19, 2023), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/?sh=4ec27165674f.

[30] Priya Shree, *The Journey of Open AI GPT Models*, MEDIUM (Nov. 9, 2020), https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2.

[31] *Id.*

[32] Jack Bandy & Nicholas Vincent, *Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus*, at 1, 2 (May 11, 2021), https://arxiv.org/pdf/2105.05241.pdf.

[33] Shree, *supra* note 30.

[34] *Id.*

[35] Bandy & Vincent, *supra* note 32, at 1, 2.

[36] *Id.* at 1, 2.

[37] *Id.* at 1, 6.

[38] Shree, *supra* note 30.

[39] Alec Radford et al., *Language Models are Unsupervised Multitask Learners*, 1, 3, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[40] Karma is the Reddit system used for placing preferred content higher on search pages, with users "liking" a post to give it "positive" karma. *What is Karma?*, REDDIT, https://support.reddithelp.com/hc/en-us/articles/204511829-What-is-karma.

[41] Radford et al., *supra* note 39, at 1, 3.

[42] *Id.*

Books1, Books2, and Wikipedia.[43] Common Crawl contains information from over 250 billion web pages since 2007 using web crawlers and provides the data for free to researchers.[44] The vast amount of data available in Common Crawl was filtered before being incorporated into GPT-3 to maintain a higher level of quality for training purposes.[45] Websites are not notified when Common Crawl scrapes their data, but they can opt out by configuring a specific site to block the crawler.[46] The second corpora used, WebText2, is an expanded version of WebText and utilizes a similar parameter for choosing targeted websites based on posts with at least three upvotes from Reddit users.[47] The contents of the Books1 and Books2 datasets are far less precise, although they appear to be comprised of books in the public domain.[48] The final dataset, Wikipedia, seems to have contained all the data and text available on the platform.[49]

OpenAI's current form, GPT-4, and its consumer-facing component, ChatGPT, seems to have the most expansive training dataset, including web texts, books, news articles, social media posts, code snippets, and other unspecified sources.[50] The datasets used for this training are currently unknown, as the company has closed off much of the information previously shared in different GPT iterations.[51] OpenAI lacks transparency about its training data for GPT-4, unlike previous iterations, which complicates the analysis of the copyrightable nature of inputs and applicability of fair use.[52] Importantly, GPT-4 utilized self-supervised learning where the model used information from its various datasets to learn from its own generated texts without human interference or guidance.[53]

[43] Shree, *supra* note 30.

[44] COMMON CRAWL, https://commoncrawl.org/ (last visited Mar. 21, 2024).

[45] Tom B. Brown et al., *Language Models are Few-Shot Learners*, 33 ADVANCES IN NEURAL INFO. PROCESSING, 1877, 1893 (2020).

[46] *Frequently Asked Questions*, COMMON CRAWL, https://commoncrawl.org/faq#:~:text=How%20can%20I%20block%20the,%2DAgent%20string%20is%3A%20CCBot (last visited Oct. 30, 2023).

[47] Roger Montti, *How to Block OpenAI ChatGPT from Using Your Website Content*, SEARCH ENGINE JOURNAL (Feb. 2, 2023), https://www.searchenginejournal.com/how-to-block-chatgpt-from-using-your-website-content/478384/#close.

[48] Gregory Roberts, *AI Training Datasets: The Books1+Books2 that Big AI Eats for Breakfast*, GREGOREITE (Dec. 14, 2022), https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/#:~:text=Books1%20%26%20Books2%20are%20two%20internet,fact%20check%20ASAP!%5D; *see also* Kyle Barr, *GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery*, GIZMODO (Mar. 16, 2023), https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989.

[49] Shree, *supra* note 30.

[50] *Id.*

[51] Barr, *supra* note 48.

[52] *See* discussion *supra* Part II (discussing the obscure nature of books used in Books2 and the use of pre-existing infringing shadow libraries to easily procure large datasets). The existence of shadow libraries, and their potential use as dataset training material, may not affect the overall fair use analysis but is acknowledged as part of the various plaintiffs' arguments. *See also supra* Part I (discussing the existence of shadow libraries as vast copyright infringing repositories in the training datasets of GenAI language models).

[53] E2Analyst, *GPT-4: Everything you want to know about OpenAI's new AI model*, MEDIUM (Mar. 14, 2023), https://medium.com/predict/gpt-4-everything-you-want-to-know-about-openais-new-ai-

Knowing what the datasets are, it must next be determined how data is collected and stored. At least some of the data collected into various datasets may be copied and stored somewhere. That said, our current understanding of the training functionality is that text from various datasets is not always processed wholesale.[54] In fact, reports indicate that ChatGPT was not trained by "reading" an entire novel at once, but rather through the analysis of small portions of a text at a time.[55] The program then jumps to another section of a different text in the attempt to create a prediction of what text will follow a given word.[56] This process is repeated through the entire dataset to assign values to create predictions and simulate human creation.[57] As such, entire books are not "read" by the machine sentence by sentence, but rather small sections are compared to sections in other books to compare the relatability of words. This comparison of small sections trains the LLM on how to place words together. Tremblay's complaint alleging copying acknowledges that outputs summarizing books contain inaccuracies but assumes retention in the form of copying to later summarize.[58] This seems to assume that ChatGPT is able to "pull" information directly from stored data that includes copyrighted works. In light of current knowledge on how training occurs, retention of entire copyrighted materials is not yet substantiated, and training as described is likely to be considered transitory, nonactionable copying as we will discuss in Part IV below. The specifics of what is and is not copied in the training process will have significant ramifications on the copyright analysis. To the extent entire works are copied, courts will be more inclined to find at least a prima facie case of copyright infringement and will need to undertake a fair use analysis. However, if the training data is not copied and stored in a separate dataset and if only transitory copies are made, courts may simply decide that the underlying works are not being copied at all. Before we analyze infringement and fair use, we must first look at the copyrightability of the training inputs.

III. Protectability of Training Inputs

The copyrightability of information used in the datasets used to train ChatGPT runs the gamut of traditional protection, including public domain works, openly licensed works, and protected works that are not openly licensed. There is likely a considerable amount of works from the public domain used for training LLMs such as ChatGPT. Although the exact content of the datasets referred to as Books1 and Books2 for ChatGPT's training are unknown, they are believed by some to be books that have entered the public domain.[59] Public domain works are not protected and can be used for training LLMs without violating copyright law.

model-a5977b42e495.

[54] Ross Anderson, *Does Sam Altman Know What He's Creating*, THE ATLANTIC (Jul. 24, 2023), https://www.theatlantic.com/magazine/archive/2023/09/sam-altman-openai-chatgpt-gpt-4/674764/.

[55] *Id.*

[56] *Id.*

[57] *Id.*; *see generally* Bandy & Vincent, *supra* note 32, at 1, 2.

[58] Complaint at 8, Tremblay v. OpenAI, Inc., 3:23-CV-03223-AMO (N.D. Cal. filed June 28, 2023).

[59] Barr, *supra* note 48.

Other inputs used, such as Wikipedia, are published under open licenses. Wikipedia itself, other than quoted portions, is openly licensed under the Creative Commons Attribution-Sharealike 4.0 International License and the GNU Free Documentation License.[60] Generally, Wikipedia content can be used without infringement so long as there is attribution.[61] Considering OpenAI's use of Wikipedia content is entirely in the training process and not consumer-facing, the form of this attribution or the need for attribution is unclear. So long as this attribution requirement is fulfilled, however, there is an argument that use of openly licensed works could be permitted under the relevant license language.

The other category of inputs is fully protected works that are not openly licensed. It seems that some work protected by copyright is present in the datasets used to train ChatGPT without prior permission. One example is the use of Reddit's Application Programming Interface (API) to create curated content in GPT-2.[62] The apparent use of Reddit content caused some controversy, with co-founder Steve Huffman stating it was "unacceptable" that other companies were scraping data from the social media site to train their systems without compensation.[63] In response, Reddit announced it would be charging for its API, the tool that allowed OpenAI to access the website's text.[64] Notably, many of the books collected on Smashwords for use in GPT-1 contain a license that limits reproduction and distribution and states "for [the reader's] personal enjoyment only."[65] There is also some evidence that the authors whose books were used in this dataset were not able to opt out of the inclusion of their works.[66] To the extent LLMs are bound to terms of services restrictions for individual users, there may be contractual questions around the use of platform content without permission of the platform. Regardless, the takeaway is that at least some of the datasets used in training LLMs like ChatGPT appear to include works protected by copyright. Knowing that, we must now look at whether those works are copied and infringed and if any defenses are available to the companies creating the LLMs.

IV. Training Methods for ChatGPT May Use Transitory Copies of Protected Works and Are Likely Non-Infringing

There is little information about the retention of copyrightable material in the datasets used by ChatGPT. Current literature indicates that works are copied to datasets to be used for training without any substantial modification[67] other than a

[60] *Wikipedia:FAQ/Copyright*, WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:FAQ/Copyright#:~:text=Most%20text%20in%20Wikipedia%2C%20excluding,be%20reused%20only%20if%20you (last visited Oct. 30, 2023).

[61] *Id.*

[62] APIs are a software intermediary, allowing different applications to "talk" to each other.

[63] Gintaras Fadauskas, *Redditors on Strike but Company Wants OpenAI to Pay Up for Scraping*, CYBERNEWS (June 12, 2023), https://cybernews.com/news/reddit-strike-api-openai-scraping/.

[64] Mike Isaac, *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems*, THE NEW YORK TIMES (Apr. 18, 2023), https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html.

[65] Bandy & Vincent, *supra* note 32, at 5.

[66] *Id.* at 5–6.

[67] Carmit Yulis et al., *Opinion: Uses of Copyrighted Materials for Machine Learning*, MINISTRY OF JUSTICE 1, 18 (Dec. 18, 2023), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf (Isr.).

potential conversion of file type.[68] However, because of the functionality of the AI model, it is not clear how much is copied or how long the copy is retained. For the most part, and as far back as GPT-1, the amount copied during training appears to be small portions at a time as the GPT model compares one section of text to another.[69] *The Atlantic* recently described this function as "a group of students who share a collective mind running wild through a library, each ripping a volume down from the shelf, speed-reading a random short passage, putting it back, and running to get another."[70] A paper published by the Israeli Ministry of Defense analogizes this function to an autonomous driving system which "'watch[es]' movies in order to teach the system . . . to anticipate a pedestrian."[71] Essentially, the LLM "understands" only a small portion of the larger work in order to compare to the overall dataset without understanding the entirety of the original work.

If the model is indeed moving sporadically between texts for comparison of the functionality of syntax and sentence structure, then there is a likelihood that the copying itself would be considered fleeting.[72] Such copying for training purposes might even be so transitory as to be noninfringing.[73] Here, *Cartoon Network LP, LLLP v. CSC Holdings, Inc.* provides some guidance on the legality of transitory copies. CSC Holdings' remote DVR system allowed for recording live television through the storage of data on a server.[74] The system operated by creating a "buffer" stream, which had all the data required for the recording but only recorded data that was selected by customers to record onto the remote device.[75] The data in the "buffer" contained everything needed for the recording, but the information only remained in this system for "a fleeting 1.2 seconds."[76] This 1.2 second time frame was considered a transitory period and did not meet the duration requirement to have been considered a copy.[77] An analogy may be drawn as ChatGPT's training is very similar to CSC's use of a buffer to hold the data required for a recording. If the dataset is considered the entire stream of data used for copying and the training functions of the LLM are comparable to the "buffer" used by CSC, the LLM's use of the dataset is likely to be considered fleeting. The LLM briefly scans a portion of the data from the dataset in order to process the information necessary to train the model. Due to the massive amounts of data in the training set, it would appear that this "scan" of the data must be transitory in nature.

Considering this, copying done to train the language model itself can be distinguished from the copying of works into the dataset that trains the language model. The copying of copyrighted works into datasets themselves would likely also

[68] Bandy & Vincent, *supra* note 32, at 9.

[69] *See* Anderson, *supra* note 54.

[70] *Id.*

[71] Yulis et al., *supra* note 67, at 18.

[72] *See* Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 127 (2d Cir. 2008).

[73] *See id.* at 130.

[74] *Id.* at 124.

[75] *Id.* at 125.

[76] *Id.* at 130.

[77] *See generally id.*

need to be successfully defended to avoid liability for copyright infringement. While some of the use of copyrighted works in LLM training might be considered transitory, if portions of copyrighted works are copied for longer periods of time by the LLM, fair use becomes the most viable defense to a claim of direct infringement.

V. Application of Current Fair Use Precedent to OpenAI's ChatGPT

The fair use analysis is highly fact-dependent, and the use of copyrightable material is considered on a case-by-case basis. As such, the analysis below will still focus on OpenAI alone and its use of copyrighted materials for the generative artificial intelligence ChatGPT. While it is not possible to make conclusive, broad statements that any and all activity around training existing LLMs is summarily protected by a fair use defense, the arguments below paint a picture of how the training of generative artificial intelligence fits into current fair use jurisprudence.

In order to apply this jurisprudence, the fair use defense must be applied as it is written in 17 U.S.C. § 107 by considering:

> (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 on potential market value of the copyrighted work.[78]

A. Purpose and Character of the Use in LLMs

In this section, we must consider both the commerciality of the use by an LLM like ChatGPT as well as whether that use was transformative. First, the commerciality of the use must be discussed. It appears that OpenAI is more likely to be considered a commercial enterprise than a non-profit, though it does have a nonprofit component. Current reports suggest that 99% of OpenAI's staff are engaged in commercial affairs, with 1% of the staff operating in the company's nonprofit endeavors.[79] The amended complaint filed in *Authors Guild v. OpenAI* claims that OpenAI would have no commercial product without the authors' copyrighted works "to power their lucrative commercial endeavor, taking whatever datasets of relatively recent books they could get their hands on."[80] While the commercial character of ChatGPT as a subscription-based service will weigh against the fair use assessment, commerciality is not determinative, as cases finding fair use related to Google Books and Google's image search engine indicate.

Notably, the Supreme Court's recent decision in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith* focused heavily on the similarity of the commercial use.[81] *Warhol* was chiefly concerned with the use of the allegedly infringing work being for the same purpose as the original.[82] The Court noted the commercial uses of

[78] 17 U.S.C. § 107.

[79] Anderson, *supra* note 54.

[80] Amended Complaint at 2, Authors Guild v. OpenAI Inc., No. 1:23-CV-08292 (S.D.N.Y. filed Dec. 5, 2023).

[81] Patrick K. Lin, *Retrofitting Fair Use: Art & Generative AI After* Warhol, 64 SANTA CLARA L. REV. (forthcoming 2024) (manuscript at 15).

[82] *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 508–09 (2023).

the original image and the Warhol images were substantially similar: to be licensed for use in a magazine.[83] However, even viewing OpenAI's ChatGPT under *Warhol*'s interpretation of commerciality, OpenAI's fair use argument remains strong. The original purposes of the underlying books and articles vary, but they were largely created to inform and entertain individual readers who would actually read the works for their educational and entertainment value. In contrast, ChatGPT's commercial use of copyrighted works consists of simply gathering datapoints to derive functional language rules and syntax so as to train a commercial language generation tool for use by individuals. Therefore, despite *Warhol*'s apparent elevation of commerciality in its fair use analysis, its specific interpretation favors OpenAI's fair use defense. Courts will then analyze the transformation of the use, which again provides an argument strongly in OpenAI's favor.

OpenAI's use of underlying original work to train its LLM is significantly transformative. Generally, courts consider the purpose of the original work, then how significantly the downstream work aesthetically changes the original work, as well as if it brings a significant new purpose, meaning, or message.[84] First, it must be understood that when OpenAI's natural language processing functions "read" text, they do not analyze the meaning but instead the functionality of sentence structure and syntax.[85] So, while the LLM is built by digesting creative material,[86] that material is not processed for its copyrightable expression but rather for its non-copyrightable aspects: the functions of language itself. Consider Chabon's claim that ChatGPT functions require "OpenAI to capture, download, and copy copyrighted written works, plays and articles."[87] Chabon states that OpenAI uses such written works to "unfairly profit from and take credit for developing a commercial product based on unattributed reproductions."[88] While it may be true that plaintiff's works were used in pursuit of a commercial endeavor, ChatGPT's function and intent is quite different than the purpose and character of the initial copyrighted works.[89] This claim is similar to many claims against technology companies in the past, who succeeded in arguing fair use as a defense in commercial endeavors.

Similar technology used for data collection on the internet provides a strong comparison and allows for a clear application of precedent, including actions against Google's use of web crawlers. Generally, web crawlers archive data by storing inputs as snapshots so that they can be used as a contained version of the live internet.[90] Notably, web crawling has been found to be fair use and is a common practice for search engine functions. *Field v. Google* found caching websites through the use of a

[83] *Id.* at 509.

[84] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

[85] Yulis et al., *supra* note 67, at 18.

[86] *Id.*

[87] Complaint at 8, Chabon v. OpenAI, Inc., 3:23-CV-04625, (N.D. Cal filed Sep. 8, 2023).

[88] *Id.* at 18.

[89] *See id.* at 6 (discussing OpenAI's use of written works as "valuable training material").

[90] Cory James, *Crawlers: What Do They Do, and How Do They Work?*, MEDIUM (Mar. 25, 2022), https://medium.com/@coryjames.proxycrawl/crawlers-what-do-they-do-and-how-do-they-work-d036eb38c0a9.

web crawler to be transformative.[91] The crawler in question, known as "Googlebot," cached the websites as a complete copy to be used in an index for Google's search engine.[92] Although the complete copying of a website could not be disputed, the court found the use transformative because Google's use was intended to create an archive that allowed users to track changes in websites and determine why a website resulted from a particular search, making its purpose distinct from Field's.[93] ChatGPT's use of datasets can be directly compared to this use of web crawling. The data is transformed because OpenAI's LLM parses the data for functional, rather than creative, aspects. While Chabon's argument that data was scraped and used without permission may not be effectively disputed,[94] OpenAI's new purpose, meaning, and message support a fair use defense.

Similarly, *Perfect 10, Inc. v. Amazon.com, Inc.* focused on the use of web crawlers that cached images and displayed them as thumbnails for use in Google's image search functionality.[95] The court ruled that the use of thumbnails weighed in Google's favor because the thumbnails were used to help internet users simply find content on the internet, rather than letting users experience the photos for their original aesthetic purposes.[96] Both the aforementioned cases featured significant retention of copyrighted materials, yet courts found these uses to be highly transformative.

While it is true that some LLMs can be designed to produce an output that resembles copyrighted inputs, that does not appear to be the case for OpenAI's current functionality. The *Times* plaintiffs make the argument that ChatGPT, rather than providing new purpose, meaning or message, acts as a substitute for The Times' works.[97] In support of this, The Times includes several prompts and responses, the methods of which seem to follow two main themes. Many of the included examples include prompts that consist of actual portions of *The New York Times* articles to try to cause the model to recreate the entire article.[98] This method shows uneven results both in the amount of the article ChatGPT outputs and its accuracy.[99] The next method used in support of The Times' argument includes directly asking what a specific author said about a subject in a *New York Times* article and then asking for the subsequent paragraphs in those articles.[100] Again, this method showed various degrees of success in recreating the actual articles.[101] Instead of providing the next paragraph in an article, ChatGPT most often provided articles several paragraphs after

[91] Field v. Google, Inc., 412 F. Supp. 2d 1106, 1119 (2006).

[92] *Id.* at 1115.

[93] *Id.*

[94] Complaint at 8, *Chabon*, 3:23-CV-04625.

[95] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1155 (9th Cir. 2007).

[96] *Id.* at 1165.

[97] Complaint at 2, New York Times Co. v. Microsoft Co., No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023).

[98] *See, e.g.*, *id.* ex. J, at 3.

[99] *See OpenAI and Journalism*, OPENAI (Jan. 8, 2024), https://openai.com/blog/openai-and-journalism; Complaint ex. J, *Times*, No. 23-CV-11195.

[100] Complaint at 18–19, *Times*, No. 23-CV-11195.

[101] *See* OPENAI, *supra* note 99; *see also* Complaint ex. J, *Times*, No. 23-CV-11195.

the first recreated paragraph.[102] In response to these exhibits, and the *Times* complaint in general, OpenAI's blog characterized the "regurgitation" of copyrighted materials as a "bug" and argued that The Times "intentionally manipulated prompts, often including lengthy excerpts of the articles" to achieve these results.[103] This *Times* claim suggests an ability for users to access copyrighted material through the use of the correct prompts.[104] This ability, considered by OpenAI to be a bug and exploitation of ChatGPT functionality,[105] appears to take concerted effort and has varied and inconclusive results.

Lending credence to OpenAI's response is the presence of considerable guardrails against outputs of copyrighted material. Consider, for example, these prompts fed into ChatGPT on October 19, 2023. First, the interface was asked for a transcript of a public domain work, Robert Frost's *The Road Not Taken*, and produced a nearly identical transcript of that work, which is allowable under copyright law.[106] However, when asked to recite the lyrics of The Beatles' *Penny Lane*, ChatGPT gave the first twelve words of the song and then provided a notice stating the content may violate their content policy or terms of use, showing some intent on OpenAI's part to prevent infringing outputs. Finally, ChatGPT was asked for a transcript of the President's speech from the film *Independence Day*. Instead of providing this transcript, the LLM offered to summarize the themes and plot of the film, protecting the expression of the speech itself. The LLM was then asked to paraphrase the speech, which it was able to do with some degree of success. The central theme remained intact but did not include the expressive language used in the movie's actual script. This mirrors admissions made in *Authors Guild*'s amended complaint wherein the plaintiffs admit that ChatGPT no longer provides quotations from copyrighted works.[107] *Authors Guild*'s amended complaint was filed just twenty-two days prior to *Times*'s complaint, indicating the inability of the plaintiffs in a similar suit to produce the same results.[108]

An effective system of guardrails preventing access to copyrighted material could prove critical to a fair use defense. Guardrails undermine the argument that LLMs simply reproduce copyrighted works and shift the focus back to the highly transformative uses by LLMs. Relative to the technologies considered in *Field* and *Perfect 10*, ChatGPT users seem to have access to far less retained copyrighted content. The purpose of the use—to derive functional language relationships and syntax—is also far from the original aesthetic and expressive purposes of the works that ChatGPT utilizes. *Google LLC v. Oracle Am., Inc.* focused on a similar use in

[102] *See* OPENAI, *supra* note 99; *see also* Complaint ex. J, *Times*, No. 23-CV-11195.

[103] OPENAI, *supra* note 99.

[104] Complaint ex. J, *Times*, No. 23-CV-11195.

[105] OPENAI, *supra* note 99.

[106] There was one errant comma in this output.

[107] Amended Complaint at 14, Authors Guild v. OpenAI Inc., No. 23-CV-08292 (S.D.N.Y. filed Dec. 5, 2023).

[108] *See* Complaint, *Times*, No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023); Amended Complaint, *Authors Guild*, No. 23-CV-08292 (S.D.N.Y. filed Dec. 5, 2023).

the innovation of a new or improved technology.[109] The Supreme Court noted that Google's use of a copyrighted program to create a new platform "was consistent with that creative 'progress' that is the basic" intent of copyright protection itself under the Constitution.[110] Similar to *Google*, OpenAI is not copying the lines of text "because of their creativity, their beauty, or even . . . their purpose" but to support the functionality of a new technological platform.[111]

We previously discussed the possibility that the fleeting copies used solely for functional language training may not be actionable under copyright law. Even if those copies, however, could be the basis of a copyright infringement claim, OpenAI has a response. If Silverman is correct in claiming that the reason ChatGPT "can accurately summarize a certain copyrighted book is because that book was copied by OpenAI," the method of copying must also be taken into consideration.[112] The sporadic nature of the ingestion of the copyrighted works clearly focuses on function, rather than expressive content.[113] OpenAI can argue that it is using only small portions of each copyrighted work at a time to train its language models and that it is using those portions solely for functional purposes.

Generally, the viability of a fair use defense decreases as the amount of the copyrighted work used increases.[114] However, even a complete copy of the entire work does not prevent a finding of fair use in instances where the use is highly transformative, such as OpenAI's creation of ChatGPT. OpenAI's "sole purpose and intent" does not lie in reproducing the expressive content contained in the copyrighted works but rather in its functionality, similar to iParadigm's Turnitin.com.[115] In *iParadigm,* the plaintiff's works were students' original written assignments,[116] which would traditionally receive copyright protection. The defendant clearly copied and saved the entire works in their database.[117] The database created by iParadigms was used to perform automated comparisons of student works in search of plagiarism, and the court found that this purpose was "unrelated" to the works' expressive components.[118] Similarly, ChatGPT uses copyrighted material to perform automated comparisons of language to derive functional language relationships and syntax. In fact, ChatGPT's use is arguably more transformative as it trains for a limited time (as opposed to constantly referring to its dataset), and it does not process the complete works to compare to other material but rather sporadically jumps between them to compare language solely to glean functional relationships of words and syntax.[119]

[109] *See* Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021).

[110] *Id.* at 30.

[111] *See id.* at 34.

[112] Complaint at 8, Silverman v. OpenAI, Inc., 23-CV-03416 (N.D. Cal. filed Jul. 7, 2023).

[113] *See supra* Part II (discussing the sporadic nature of "reading" collected works to derive their functionality).

[114] A.V. v. iParadigms, LLC, 562 F.3d 630, 642 (4th Cir. 2009).

[115] *Id.*

[116] *Id.* at 641–42.

[117] *Id.*

[118] *Id.* at 640.

[119] *See supra* p. 238 (discussing the parsing of inputs for functionality and not expressive content).

Further clarification on fair use and complete copies can be found in *Authors Guild v. Google*. This precedent-setting case established that even wholesale copying of digital works may be acceptable so long as the use is sufficiently transformative.[120] Here, Google made digital copies of "tens of millions" of copyrighted books and then scanned those digital copies for use in a search function.[121] Users were able to search using a term that would result in the relevant book coming up along with a "snippet" of text from that book.[122] Plaintiff authors argued that such use was not transformative and that allowing users to see a snippet of the copyrighted material should be considered infringement.[123] However, when the court considered the database of books as a whole, they found that the work was transformative as it allowed Google to "examine 'word frequencies, syntactic patterns, and thematic markers.'"[124]

This functionality is similar to that of ChatGPT, which also analyzes the structure of the works—in that case to predict which words are likely to follow others when the program generates original content. The court's finding in *Authors Guild* was dependent on the ability for users to only see a small quantity of the works used.[125] The search function only allowed users to see small snippets of every page and disabled the ability to view snippets of works where a single snippet might "satisfy the searcher's present need for the book."[126] Following this reasoning, ChatGPT might be even more transformative than Google's use in *Authors Guild*, as users are unable to search for even a snippet of the original works on some occasions and on others are only given the very beginning of searched works. The *Times* complaint against OpenAI argues that ChatGPT's user interface may be exploited to produce larger amounts of copyrighted works, but this reproduction appears to be not only difficult but also inaccurate and incomplete.[127] Further, the alleged ability to access copyrighted material does not negate the fact that complete copying is found justified when it was "reasonably appropriate" to achieve the transformative purpose of the copying party.[128] The technical limitation requiring full copying into the dataset before use in training should not preclude fair use where only small portions of that copy are referenced briefly and used solely to derive their functional relationships and syntax.[129]

In sum, OpenAI does not appear to retain entire copyrighted works, nor does it offer those works to users. In contrast, defendants in the various cases discussed engaged in more substantial copying, keeping extensive cached archives of works and sometimes even reproducing portions of those works in response to user requests.

[120] *See* Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

[121] *Id.* at 207.

[122] *Id.*

[123] *Id.*

[124] *Id.* at 209 (quoting Authors Guild, Inc. v. Google, Inc., 954 F. Supp. 2d 282, 287 (S.D.N.Y. 2013)).

[125] *Id.* at 210.

[126] *Id.*

[127] *See supra* pp. 238–39 (discussing the *Times* complaint and exhibit and the inconclusive nature of attempts at reproducing copyrighted material).

[128] *Authors Guild*, 804 F.3d at 221.

[129] Yulis et al., *supra* note 68, at 20.

Yet these defendants' uses were found to be transformative and ultimately protected by fair use. Unlike extensive cached archives in the various cases discussed, users of ChatGPT cannot access this archived data outside of rare occurrences and with great effort. Rather than a library of copyrighted content, users of ChatGPT are provided generative artificial intelligence capability that creates new content using functional language and syntax derived from millions of individual works.

The highly transformative nature of ChatGPT's GenAI training weighs heavily in favor of a finding of fair use under current case law.

B. The Nature of the Original Works Used by LLMs

Next, the nature of the works used in both datasets and training must be considered both in the type of the work and the degree of protection the work receives under copyright law. Generally, reuse of factual and nonfiction works supports a finding of fair use. Reuse of highly creative works like fictional literature or music weighs against a finding of fair use. Current litigation focuses on fictional literature, all of which appear to be published works.[130] Reuse of published works supports fair use, while reuse of unpublished works weighs against fair use. Here, it appears that ChatGPT is only using published works (which modestly supports a fair use argument). As far as factual versus highly creative works, the diverse nature of the datasets used to train ChatGPT means the nature of the work used will span the entire gamut of types of protectible works. Outside of the use of musical or unpublished works, this factor weighs significantly less than the others, especially when the use is highly transformative, so discussion of this area will receive far less scrutiny in this comment.

The likely result is that the use of fact-based research and news reporting, further broken down by the LLM for only its functionality, will likely support a finding of fair use. Use of more highly creative works will likely weigh against a fair use finding. However, as discussed, this factor is typically not determinative, and in this case OpenAI is likely to overcome any obstacles this factor presents due to the transformative use of underlying works.

C. Amount and Substantiality of the Use

This factor considers the amount and substantiality of the portion of the work used, both quantitatively and qualitatively.

1. Quantitatively

Generally, copyrighted materials are copied in full for use in the datasets.[131] This copying is necessary for the computer to access the unprotected functional components of the work.[132] Humans can study copyrighted material to understand

[130] *See, e.g.*, Chabon v. OpenAI, Inc., No. 23-CV-04625 (N.D. Cal. filed Sep. 8, 2023); Authors Guild v. OpenAI Inc., No. 23-CV-08292 (S.D.N.Y. filed Sep. 19, 2023); Silverman v. OpenAI, Inc., No. 23-CV-03416 (N.D. Cal. filed Jul. 7, 2023); New York Times Co. v. Microsoft Corp., No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023).

[131] Yulis et al., *supra* note 67, at 20.

[132] *Id.*

syntax, sentence structure, and the relationship between words. But to train a LLM to develop a similar "understanding" of language, computers need to have works copied into a dataset for processing.[133] As previously discussed in relation to *Field*, *Perfect 10*, and *iParadigms*, the full copying of works does not preclude fair use in cases where the use is highly transformative. Even if the complete copying of works reduces the fair use argument for OpenAI somewhat, ChatGPT does not share the heart of the work of its datasets, and the necessity of that copying for a highly transformative purpose, along with the fact that each work is only an extremely small fraction of the training dataset, means that the overall fair use argument will likely still weigh in OpenAI's favor.

2. *Qualitatively*

The language model used to power ChatGPT is only effective if there is a diverse, expansive dataset used to analyze the functionality of language. Without this dataset, ChatGPT cannot exist. Once again considering the nature of the training, which scans small portions of each work before moving on to a different work included in the dataset, the qualitative substance of the work is greatly diminished.[134] It does not focus on the entire creative expression of any single work and instead focuses on the use of language in one small area of that work.[135] Each work included in the datasets and parsed by the training model are miniscule compared to the entirety of the data used in the training process.[136] In addition, the portions used from each work are also very small compared to the individual work from which it derives, considering the process ChatGPT uses to "learn" the functionality of language.[137] What ChatGPT takes is not the expressive heart of the copyrighted works but simply functional language relationships and syntax. Again, OpenAI can distinguish itself from precedent set by *Harper & Row v. Nation Enterprises*, which considered the publication of a portion of Gerald Ford's memoir by *The Nation*.[138] *The Nation* published between 300 and 400 words of the 500-page book, comprising only a very short portion of the overall work.[139] The court, however, found against *The Nation* because the selected text included Ford's reasoning for pardoning former President Nixon, which was found to be the "heart" of the work.[140] The considerable guardrails in place on ChatGPT's consumer interface act as a way to prevent access to the "heart of the work" contained in any of the works used in the training datasets.

[133] *Id.*

[134] *See supra* Part IV (positing the possibility of transient copying); *supra* Part V.A (discussing the use of functionality from training material rather than copyright protected expressive content).

[135] Anderson, *supra* note 54.

[136] *See supra* Part II (analyzing the different and varied sources and types of works included in training datasets).

[137] *See supra* Part IV (discussing the nature of OpenAI's training techniques, its sporadic nature, and how that affects "copying" analysis in fair use).

[138] *See* Harper & Row Publishers. v. Nation Enterprises, 471 U.S. 539, 542 (1985).

[139] *Id.* at 545.

[140] *Id.* at 600.

Consider the complaint filed in *Silverman*, which claims the language model is "extracting expressive information" from the works that are used in the dataset.[141] Compared to current knowledge on how ChatGPT training functions, this argument that the LLM is extracting expressive information appears inaccurate. Even if courts consider the functionality of LLMs to go to the heart of the expressive content, its output is designed to not share that expressive content with the user. The model is simply using the functional language rules derived from comparing short passages of millions of individual works to create new expressive content. In this way, LLMs use complex mathematical algorithms and training to mimic human operations much in the way previous works inspire future works under human authorship. Copyright law's historical aversion to protecting the functionality of works suggests courts will be unlikely to see the functional elements of copyrighted works as the "heart" of those works.

D. Impact on Market Value

The final fair use factor to consider is the impact on the value of the work and its potential market value. Here, the value of the works themselves must be considered both individually as well as by the value of the license of a work to be used in a large-scale dataset used by LLMs such as OpenAI's ChatGPT. The use is unlikely to have significant impact on the existing, actual markets for the expressive works in the dataset. The dataset is intended to remain hidden from the public, and the output system contains guardrails. Therefore, ChatGPT is designed to provide no additional ways to read or access the original expressive work, and individuals must continue to access books, articles, and other written work in through copyright owner authorized distribution.

A market effect argument would have to rely on a court finding a viable market for the functional, non-expressive elements of copyrighted works. The potential market value for large datasets of licensed works to be used for their non-expressive functional elements is difficult to assess at this time because there are not extensive established markets for licensing works to be used in the training of LLMs. While some licensing has occurred for datasets, it is still in its infancy. Recent literature suggests that purchasing a license for all the works required for a complete dataset is likely impossible.[142] While it may be possible to purchase a license for a set of works from a single author, publisher, or content holder, acquiring these piecemeal from their copyright holders would be prohibitively expensive for many LLM creators. Mandating a market for licensing all data would likely restrict the ability for many LLM developers to compete with large established corporations, such as the Microsoft-backed OpenAI.[143] Indeed, some of these early LLM competitors appear to have used works without authorization and only later, once they had established market position, added the ability for copyright owners to opt out or provide a license.

[141] Complaint at 1, Silverman v. OpenAI, Inc., No. 23-CV-03416 (N.D. Cal. filed Jul. 7, 2023).

[142] Yulis et al., *supra* note 67, at 20.

[143] *Id.* at 21.

Also important to consider is the market for each individual work in the large-scale license of dataset material. The inherent value of datasets is the sheer volume of diverse content. Any single work has relatively little value to the dataset as a whole. For example, our understanding of the functionality of ChatGPT is that the LLM does not place greater value on certain works over the others in assessing the language's functionality based on that work's commercial appeal or success. The *Times* plaintiffs make an argument about the value of its works in comparison to other works included in training datasets but base this argument on The Times' Google PageRank and not a more concrete example concerning ChatGPT functionality.[144] This value-based argument is undermined in *Times* as the plaintiffs note WebText2 is weighted at 22% of GPT-3's training parameters, and that *The New York Times* content makes up 1.23% WebText2 corpus.[145] Therefore, the total contribution of *New York Times* articles to GPT-3's training is likely at 0.276%.[146] While *The New York Times* may represent one of the most significant private sources used in ChatGPT's training (though still less than 1%), the individual books at the heart of book authors' claims will represent a much smaller portion of training data.

In summary, the four factors as a whole weigh in favor of fair use in the case of ChatGPT. While entire copyrighted works are used, the use is highly transformative, using underlying works only for their functional language elements rather than their expressive qualities. The second factor is likely indeterminative in this case, and the third factor's quantity element is diminished by the highly transformative use. Finally, under the fourth factor, there is little effect on the actual and potential market for the expressive works themselves. Further, if there is a mandated market for the non-expressive functional and syntax elements of works, there is a public interest argument that this could create monopolies in training information and quickly close the development of AI down to just a few powerful companies with the means to license copyrighted works.

VI. International Approaches to Copyright and AI

The above argument for fair use of copyrighted material in LLM training datasets through the application of case precedent should be considered along with the practical and socioeconomic effects of not adopting such fair use arguments. The approach must take into account other nations' applications of copyright law to GenAI. The following discussion will focus on the approaches of Israel, the European Union (EU), and the United Kingdom, which represent diverse policy approaches to GenAI.

[144] Complaint at 56, New York Times Co. v. Microsoft Corp., No. 23-CV-11195 (S.D.N.Y. filed Dec. 27, 2023).

[145] *Id.* at 26.

[146] The overall contribution of *The New York Times* articles was calculated as follows. WebText2 made up 22% of all training data. This 22% was divided by 100 to equal 0.22. 0.22 was multiplied by 1.23 (the percentage *The New York Times* contributed to the WebText2). 0.22 multiplied by 1.23 is 0.276, or the overall amount of *The New York Times* content included in GPT-3.

Israel has taken an approach focusing on what it calls "responsible innovation," designed to clarify GenAI use and keep the nation at the forefront of technological innovation.[147] Israel's corresponding draft document is meant to act as both "a moral and business oriented compass" for companies to innovate and grow within a defined scope of regulation.[148] The draft attempts to avoid a "lateral framework legislation" and instead create "'soft' regulatory tools" that consider widespread ethical principles.[149] The long-term goal seems to be allowing widespread development for economic gain while balancing the public's privacy and security interests.[150] The nation's Ministry of Justice has released guidance stressing the importance of finding input datasets to be fair use in order to further promote GenAI expansion.[151] Supporting this assessment included a strong argument for not analyzing fair use on an ad hoc basis but instead carving out a large exception for LLM datasets to reduce litigation, promote efficiency, and "enhance certainty for market players on both sides."[152] This finding stressed the need to consider more than just traditional fair use factors, also including the need for consideration of how other countries will allow the technology to grow.[153] By finding for fair use in the training of LLMs, the Israeli authorities have made a choice to foster the economic and social benefits afforded by its development. The takeaway is that some countries will adopt policies aimed at establishing their nations as early adopters and homes for GenAI development.

The EU AI Act as it passed in 2023 appears to require a "sufficiently detailed summary" of the works included in training.[154] There is some ambiguity about what constitutes "sufficiently detailed" as well as how often that information needs to be updated.[155] This transparency requirement appears to be a compromise between some legislators who favored a general ban on allowing copyrightable work to be used in training for GenAI and others who wished to allow for the promulgation of the technology.[156] Questions remain about the purpose and effect of such a disclosure requirement.

While the existence of a regulated copyright registry in the U.S. could allow for disclosure of training inputs, it is unclear whether the U.S. should desire to mirror such an approach. LLM developers will have concerns about the effect of reporting

[147] *For the First Time in Israel: The Principles of the Policy for the Responsible Development of the Field of Artificial Intelligence Were Published for Public Comment*, MINISTRY OF INNOVATION, SCIENCE AND TECHNOLOGY (Nov. 17, 2022), https://www.gov.il/en/departments/news/most-news20221117 (Isr.).

[148] *Id.*

[149] *Id.*

[150] *Id.*

[151] *See* Yulis et al., *supra* note 67.

[152] *Id.* at 25.

[153] *Id.* at 28–33.

[154] Lutz Reide et al., *Has Copyright Caught Up with the AI Act?*, LEXOLOGY (May 16, 2023), https://www.lexology.com/library/detail.aspx?g=d9820844-8983-4aec-88d7-66e385627b4a.

[155] *See id.*

[156] Supantha Mukherjee et al., *EU Proposes New Copyright Rules for Generative AI*, REUTERS (Apr. 27, 2023 11:51PM), https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/.

on intellectual property protection. In addition, this new requirement would represent a new regulatory and liability hurdle, a sort of attribution that has never previously been required under fair use. It is also unclear what positive effect such a registry would have for creators or LLM developers. Finally, the United Kingdom provides an additional data point, as they appear to be working towards fostering licensing markets for copyrightable material, including potentially mandatory licensing requirements for copyright owners.[157] Conflicting approaches to GenAI in different countries could lead to a complicated, disparate legal approach around the world and could also put countries that do not provide an exception for GenAI training at an economic disadvantage.

Conclusion

GenAI development has relied on the availability of copyrightable materials for the use in training datasets. The current state of AI development has left copyright owners seeking clarification, through the courts and the legislature, over their rights when works are included in training datasets.[158] Application of current fair use case precedent creates a strong argument for continued GenAI development and innovation, allowing a narrow use for particularly transformative purposes that does not affect any traditional markets for copyrightable works. Some copyright owners have argued that if the courts and legislature agree with GenAI companies on their interpretation of fair use and inputs, the "exception [will swallow] the rule."[159] However, new technology does not necessarily require a new set of rules. The framework for handling questions related to whether the inputs of GenAI are infringing exists in our case precedents, and we can continue to apply that precedent. If U.S. courts and regulators take an alternative approach that indicates the use of copyrightable works to train without permission is infringement, it will upend the development of GenAI. Such an approach would mandate a licensing market, creating potentially significant oligopolies in AI development, where the only companies that can afford to research and develop effective artificial intelligence solidify their market position through their ability to pay for training inputs. This approach also effectively protects the functional aspects of copyrighted works, an aspect of works never before protected by copyright law. New protections for the functional aspects of works would put established copyright doctrines and traditional fair use at risk. Protection for functional elements would have negative effects on fair use beyond commercial GenAI, including potentially chilling journalistic efforts which "ingest data . . . [and] information . . . [and] put it out in a different way."[160] Finally, treating the use of copyrightable works in LLM training datasets as infringing may cause major GenAI developers simply move overseas. Rather, treating the use of copyrightable works in LLM training datasets as fair use is likely critical to maintaining U.S. competitiveness in artificial intelligence development. This

[157] Reide et al., *supra* note 154.

[158] *See generally* Moreno, *supra* note 26.

[159] *Id.*

[160] *Id.*

approach is also consistent with longstanding copyright law doctrines that are well articulated in our statutes and case law.