BEFORE THE COPYRIGHT OFFICE

LIBRARY OF CONGRESS, Washington, D.C.

Artificial Intelligence and Copyright	**Docket No.** **2023-6**

Comments of New Media Rights

by Art Neill, James Thomas, and Erika Lee

I. Commenting Party

New Media Rights is a non-profit program of California Western School of Law, a 501(c)(3) non-profit, that provides preventative, one-to-one legal services to creators, entrepreneurs, and internet users whose projects require specialized internet, intellectual property, privacy, and media law expertise. These legal services include advising on issues surrounding artificial intelligence and copyright law. Further information regarding New Media Right's mission and activities can be obtained at http://www.newmediarights.org.

Comments

This comment focuses on legal issues surrounding the use of copyrightable inputs in training datasets for artificial intelligence. Our focus is the training of large language models, specifically ChatGPT. We primarily focus on whether or not such training uses are fair use. To evaluate fair use, we will apply the basic approach to a copyright question. First, we will establish which inputs, if any, are protected. Then we will discuss whether or not there is an actual infringement of those underlying works in the training process. Finally, we will undertake a Fair Use analysis by discussing: (1) the purpose and character of the use; (2) the character of the work used; (3) the quantitative and qualitative scope of the works used; and (4) the impact of the use on the market for the copyrightable materials.

I. Data Collection – Response to Question Category 6 – What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

Application of the Fair Use Doctrine large language model ("LLM") training inputs must inherently begin with an understanding of the datasets used for training artificial intelligence ("AI"). Although many different LLMs are currently in use and many more to follow, this section will focus on OpenAI's ChatGPT training, 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.

The consumer-facing form of ChatGPT 4.0, the most popular LLM, began with GPT-1. The dataset used to train this iteration was relatively small compared to the subsequent forms. GPT-1 seems to have utilized BooksCorpus, a collection of 7,000 unpublished and self-published books[1] collected from Smashwords,[2] a repository of unpublished and self-published books. [3]

[1] 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.

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

[3] Shree *supra* note 1.

GPT-1 utilized this dataset to study "large stretches of contiguous text" to train the AI on word dependencies in a large range.[4] The creation of BooksCorpus itself was done by researchers from the University of Toronto and the Massachusetts Institute of Technology.[5] While the paper published lists seven authors, it does not explicitly mention who collected the data.[6] 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.[7]

GPT -2 used a much larger dataset than the previous iteration by using select upvoted Reddit posts and pulling data from outbound links in the targeted Reddit posts.[8] To build this dataset, OpenAI appears to have focused on using human curation to use only higher-quality text.[9] Other web-scraped datasets were considered too broad with "unintelligible" content, so OpenAI apparently focused on using Reddit posts with at least three karma[10] to focus on the test that was "interesting, educational, or just funny."[11] 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 apparently paired down to over 8 million documents.[12]

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, Books1, Books2,

[4] *Id.*
[5] Bandy, *supra* note 2, at 1.
[6] *Id* at 5.
[7] *Id.* at 9.
[8] Shree *supra* note 1.
[9] 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.
[10] Karma is the Reddit system used for placing preferred content higher on search pages, with users "liking" a post to give it "positive" karma.
[11] Radford *supra* note 9, at 3.
[12] *Id.*

and Wikipedia.[13] Common Crawl apparently contains information from over 250 billion web pages since 2007 using WebCrawler's and provides the data for free to researchers.[14] The vast amount of data available in Common Crawl was apparently filtered before being incorporated into GPT -3 to maintain a higher level of quality for training purposes.[15] Websites are not notified when Common Crawl scrapes their data, but they can opt out by configuring a specific site to block the crawler.[16] 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.[17] The contents of the Books1 and Books2 datasets are far less precise, although they appear to be comprised of books in the public domain.[18] The final dataset, Wikipedia, seems to have contained all the data and text available on the platform.[19]

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.[20] The datasets used for this training are currently unknown, as this information is much less available than in previous GPT iterations.[21] GPT-4 seems to have utilized self-supervised learning where the model used information from

[13] Shree *Supra* note 1.
[14] *Overview*, COMMON CRAWL (2023), https://commoncrawl.org/.
[15] Tom B. Brown et al., *Language Models are Few-Shot Learners*, ARXIV 1, 43 (July 22, 2020), https://arxiv.org/pdf/2005.14165.pdf.
[16]*Frequently Asked Questions*, COMMON CRAWL, https://commoncrawl.org/faq#:~:text=How%20can%20I%20block%20the,%2DAgent%20string%20is%3A%20CCBot (last visited October 30, 2023).
[17] 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.
[18] *See* AI Training Datasets: the Books1+Books2 that Big AI eats for breakfast, GREGOREITE, https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/#:~:text=Books1%20%26%20Books2%20are%20two%20internet,fact%20check%20ASAP!%5D (last visited Oct. 30, 2023); *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.
[19] Shree *supra* note 1.
[20] *Id.*
[21] Barr *supra* note 18.

its various datasets to learn from its own generated texts without human interference or guidance.[22]

Knowing what the datasets are, the next question is how is all this data 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.[23] 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.[24] 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.[25] This process is repeated through the entire dataset to assign values to create predictions and simulate human creation.[26] As such, entire books are not "read" by the machine in one 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 model how to place words together. This strategy of training has significant ramifications on the fair use of copyrightable material, and to the extent any of the data is not collected and stored in a separate dataset, could affect the question of whether the underlying work is even being copied at all.

<u>II. Are the works used to train protected? – Response to Question Category 8 – Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute Fair Use?</u>

[22] 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-model-a5977b42e495.
[23] Ross Anderson, *Does Sam Altman Know What He's Creating?*, THE ATLANTIC (July 24, 2023).
[24] *Id.*
[25] *Id.*
[26] Anderson *supra* note 23; *see generally* Bandy *supra* note 2.

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 are likely a considerable amount of works from the public domain used for training LLM models 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.[27] Public domain works are not protected, and therefore can be used for training purposes in LLM models without violating copyright law.

Other inputs used, such as Wikipedia, are 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.[28] Generally, Wikipedia content can be used without infringement so long as there is attribution.[29] 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 does appear that some work protected by copyright is present in the datasets used to train ChatGPT without prior permission. One example seems to be the use of Reddit's API to create curated

[27] Barr *supra* note 18.

[28] *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).

[29] *Id.*

content in GPT-2. 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.[30] In response, Reddit announced it would be charging for its application programming interface ("API"), the tool that allowed OpenAI to access the website's text.[31] 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."[32] 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.[33] 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 of LLMs like ChatGPT appear to include works protected by copyright. We must now look at whether those works are actually copied and infringed, and also if any defenses may be available to the companies creating the LLMs.

A. Is there copying?

There is little information about the retention of copyrightable material used in the datasets used by ChatGPT. Current literature indicates that works are copied to datasets to be used for training without any substantial modification[34] other than a potential conversion of file

[30] 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/.

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

[32] Bandy *supra* note 2, at 3.

[33] *Id.* at 6.

[34] 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.

type.[35] 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.[36] *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."[37] 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 . . . rather than enjoy the aesthetic quality or content of the film." [38] Essentially, the LLM processes only a small portion of the larger work in order to compare to the overall dataset without processing the entirety of the original work at one time.

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.[39] Such copying for training purposes might even be so transitory as to be non-infringing.[40] Here, *Cartoon Network LP, LLLP v. CSC Holdings, Inc.* provides some applicable precedent on the nature of transitory copies. CSC allowed for recording of live television through the storage of data on a server.[41] The system operated by creating a "buffer" stream which had all of the data required for the recording, but only recorded data that was selected by customers to record onto the remote device.[42] The data in the "buffer" contained

[35] Bandy *supra* note 2, at 6.
[36] *See* Ross Anderson, *Does Sam Altman Know What He's Creating?*, THE ATLANTIC (July 24, 2023).
[37] *Id.*
[38] Yulis note 34, at 18.
[39] Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 127.
[40] *Id.* at 130
[41] *Id.* 139
[42] *Id.* at 129

everything needed for the recording, but the information only remained in this system for "a fleeting 1.2 seconds."[43] This 1.2 seconds was considered a transitory period and did not meet the duration requirement to have been considered a copy.[44] ChatGPT's training is very similar to CSC's use of a buffer to hold the data required for a recording. If the training functions of the LLM are compared to the "buffer" used by CSC, the LLM's use of the dataset might 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 might be transitory in nature.

Considering this, copying done to train the language model itself can be distinguished somewhat from the copying of works into a dataset that is then used to train the language model. The copying of copyrighted works into datasets themselves would likely also need to be successfully defended to avoid liability for copyright infringement. While the LLM's use of that data in training might be considered transitory, portions of copyright material are copied for a period of time by the LLM, and so Fair Use must be discussed in relation to the training process.

B. Fair Use

Fair Use analysis is 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

[43] *Id.* at 130

[44] *See generally Id.*

the various steps to training of generative artificial intelligence fit into current fair use jurisprudence.

In order to apply this jurisprudence, the fair use defense must be considered as it is written in 17 U.S.C.S. § 107 by considering: (1) the purpose and character of the use, both its commercial nature and whether it is transformative; (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.

1. 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, we must consider whether or not the use is commercial. It appears that OpenAI is a more commercial enterprise than a non-profit, though it does have a nonprofit component. Current reports suggest that 99% of the OpenAI's staff are engaged in commercial affairs, with 1% of the staff operating in the company's non-profit endeavors.[45] While the commercial nature 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.[46] The Court noted the commercial use of the image was substantially similar; to be licensed for use in a magazine cover.[47] However, even viewing OpenAI's ChatGPT under Warhol's interpretation of commerciality, OpenAI's fair

[45] Ross Anderson, *Does Sam Altman Know What He's Creating*, The Atlantic (July 24, 2023).

[46] Patrick K. Lin, *Retrofitting Fair Use: Art & Generative AI After* Warhol 64 SANTA CLARA L. REV. Working Paper 1, 15 (forthcoming 2024), https://ssrn.com/abstract=4566945.

[47] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1264 (2022).

use argument remains strong. The commercial nature of the use of copyrighted works to train ChatGPT is to create functional language capability to support language generation by the user. The copyrighted works within its dataset have a substantially different commercial use that relates to their original expressive purpose. Courts will then analyze the transformative nature of the use, which here provides an argument strongly in OpenAI's favor.

The transformative nature of OpenAI's LLM is significant. Generally, courts consider the purpose of the original work, and then how significantly the downstream work aesthetically changes the work, as well as if it brings significant new purpose, meaning or message. 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.[48] So, while the LLM is built by digesting creative material,[49] that material is not processed for its copyrightable expression, but rather for its non-copyrightable aspects: the function of language itself.

Similar technology used for data collection on the internet must be used for comparison. Generally, web crawlers archive data by storing inputs as a snapshot so that they can be used as a contained version of the live internet.[50] Notably, web crawling has been found to be fair use and is a common practice for search engine functions. *Field v. Google* found the transformative nature of caching websites through the use of a web crawler.[51] This crawler, known as "Googlebot" cached all of the websites as a complete copy to be used in an index for Google's search engine.[52] Although the complete copying of a website could not be disputed, the court

[48] Yulis note 34, at 18.
[49] *Id.*
[50] 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.
[51] *See* Field v. Google, 412 F.Supp.2d 1106 (2006).
[52] *Id.* at 1115.

found the use transformative because Google's use was distinct from Field's through the creation of an archive that allowed users to track changes in websites and determine why a website resulted from a particular search.[53]

ChatGPT's use of datasets can be directly compared to this use of web crawling. The data is transformed by the use of OpenAI's LLM, which parses the data for functional, rather than its creative, aspects. Unlike Google's cached archive, it seems that users of ChatGPT cannot access this archived data. Rather than a library of copyrighted content, users of ChatGPT are provided generative AI capability that creates new content using functional language and syntax derived from millions of individual works. Similarly, *Perfect 10 Inc. v. Amazon.com, Inc.*, focused on the use of web crawlers which cached images and displayed them as thumbnails for use in Google's image search functionality.[54] 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. Both the aforementioned cases featured significant retention of copyrighted materials, yet courts found these uses to be highly transformative. Even when prompted to produce archived copyrighted content, ChatGPT has guardrails that do not allow the reproduction of most copyrighted works.

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. 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.[55] However, when asked to recite the lyrics of

[53] *Id.*

[54] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1155 (2007).

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

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 was intact but did not include the expressive language used in the movie's actual script.

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 purposes of the works that ChatGPT utilizes. *Google LLC v. Oracle Am., Inc.* focused on a similar use in the innovation of a new or improved technology.[56] The 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. As in *Google*, OpenAI is not copying the lines of text "because of their creativity, their beauty, or even . . . their purpose" but to support the interoperability of computer systems.[57] We previously discussed the possibility that the fleeting copies used solely for functional language training may not be actionable under copyright law. However, even if those copies could be the basis of a copyright infringement claim, OpenAI has a response. Open AI can argue that it is using only small portions of each copyrighted work at a time to train its language models and is using those portions solely for functional purposes.

[56] *See* Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021).
[57] *Id.* at1188.

Generally, the viability of a Fair Use defense decreases as the amount of the copyrighted work used increases.[58] 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.[59] In *iParadigm,* the plaintiff's works were students' original works that traditionally receive copyright protection. The defendant clearly copied and saved the entire works in their database.[60] The database created by iParadigm's 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.[61] 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 works as a whole to compare to other material, but rather sporadically jumps between them to compare language solely to glean functional relationships of words and syntax.

In addition, *Authors Guild v. Google* further establishes that even wholesale copying of digital works may be acceptable so long as the use is sufficiently transformative.[62] Here, Google made digital copies of "tens of millions" of copyrighted books and then scanned those digital copies for use in a search function.[63] Users were able to search using a term which would result in the relevant book coming up along with a "snippet" of text from that book.[64] Plaintiff authors

[58] A.V. v. iParadigms, LLC, 562 F.3d 630, 642.
[59] *Id.*
[60] *Id.* at 641.
[61] *Id.*
[62] *See* Authors Guild v. Google, Inc., 804 F.3d 202 (2015).
[63] *Id.* at 207.
[64] *Id.*

argued such use was not transformative and allowing users to see a snippet of the copyrighted material should be considered infringement.[65] 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.'"[66]

This functionality is similar to that of ChatGPT, which also analyzes the structure of the works, in this case to predict which words are likely to follow 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.[67] 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."[68] Following this reasoning, ChatGPT might be even more transformative than Google's use in *Author's Guild,* as users are unable to search for even a snippet of the original works. Complete copying is found justified when it was "reasonably appropriate" to achieve the transformative purpose of the copying party.[69] 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.[70]

The highly transformative nature of ChatGPT's AI training weighs heavily in favor of a finding of Fair Use under current case law.

2. <u>The Character of the Work Used by LLMs</u>

[65] *Id.*
[66] *Id.* at 209.
[67] *Id.* at 210.
[68] *Id.*
[69] *Id.* at 221.
[70] Yulis *supra* note 34, at 20.

Next, the character 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 non-fiction works supports a finding of fair use. Reuse of highly creative works like fictional literature or music weigh against a finding of fair use. In addition, 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 character of the work used will span the entire gamut of types of 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 nature of its use of underlying works.

3. **Amount and Substantiality of the Use**

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

a. **Quantitative**

Generally, copyrighted materials are copied in full for use in the datasets.[71] This copying is necessary for the computer to access the unprotected functional components of the work.[72] Humans are able to study copyrighted material to understand syntax, sentence structure, and the relationship between words. To develop a similar "understanding" of language, computers need to have works copied into a dataset.[73] As previously discussed in *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 means that the overall fair use argument will still weigh in OpenAI's favor.

b. **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. 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. 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. 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

[71] Yulis *supra* note 34, at 19.
[72] *Id.*, at 20.
[73] *Id.*

the publication of a portion of Gerald Ford's memoir by *The Nation* magazine.[74] *The Nation* published between 300 and 400 words of the 500-page book, comprising only a small portion of the entirety of the work.[75] However, the court found against *The Nation* because the text included Ford's reasoning for pardoning Former President Nixon, which was found to be the "heart" of the entire book.[76] OpenAI's user facing application, ChatGPT, does not provide the user with information in this way. The guardrails discussed above[77] act as a way to prevent access to the "heart of the work" contained in any of the works used in the training datasets.

4. **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 remains hidden from the public, and the output system contains guardrails. Therefore, ChatGPT provides 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

[74] *See* Harper & Row, Publrs. v. Nation Enters., 471 U.S. 539 (1985).
[75] *Id. at* 545.
[76] *Id.* at 600.
[77] *Supra* Section II(B)(1). (Discussing guardrails in place on ChatGPT to preclude users from accessing copyrightable inputs).

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.[78] 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.[79] 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.

Also important is to consider 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, as such, 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.

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 undeterminative in this case, and the third factor's quantity element is diminished by the highly transformative use. Finally, under the fourth factor, there is

[78] Yulis *supra* note 34, at 20.
[79] *Id.* at 21.

little effect on the actual and potential market for the expressive works themselves. If we mandate a 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.

III. Practical/Socioeconomic – Response to Question 4 – Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States?

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. Our approach must take into account other nations' approach to copyright and generative AI. We will focus on the approach of Israel, the European Union ("EU"), and the UK.

Israel has taken an approach focusing on what it calls "responsible innovation," designed to clarify AI use and keep the nation at the forefront of technological innovation.[80] The document itself is meant to act as both "a moral and business oriented compass" for companies to innovate and grow within a defined scope of regulation.[81] The draft attempts to avoid a "lateral framework legislation" and instead create "'soft' regulatory tools" that consider widespread

[80] *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.
[81] *Id.*

ethical principles.[82] The long-term goal seems to be allowing widespread development for economic gain while balancing the public's privacy and security interests.[83] The nation's Ministry of Justice has released guidance stressing the importance of finding input datasets fair use in order to further promote AI expansion.[84] 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."[85] This finding stressed the need to consider more than just traditional fair use factors, but also the need for consideration of how other countries will allow the technology to grow.[86] By finding for fair use in the training of LLMs, Israel appears poised 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 AI development.

The EU AI Act as it passed in 2023 appears to require a "sufficiently detailed summary" of the works included in training.[87] There is some ambiguity about what constitutes "sufficiently detailed" as well as how often that information needs to be updated.[88] 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 generative AI and those wishing to allow

[82] *Id.*
[83] *Id.*
[84] *See* Yulis *supra* note 34.
[85] *Id.* at 25.
[86] Yulis *supra* note 34, at 28-33.
[87] Lutz Riede 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.
[88] Reide supra note 87.

promulgation of the technology.[89] 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 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.[90] Conflicting approaches to AI in different countries could lead to a complicated, disparate legal approach around the world, and also put countries that don't provide an exception for AI training at an economic disadvantage.

Conclusion

At this point, AI development relies on the availability of copyrightable materials for the use in training datasets. Application of current fair use case precedent creates a strong argument for continued development and innovation. The framework for handling questions related to whether the inputs of generative AI are infringing exists in our case precedents and we can continue to apply that precedent. If U.S. courts and regulators consider an alternative approach that mandates a licensing market, they should be wary of creating potentially significant

[89] 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/.
[90] Reide supra note 87.

oligopolies in AI development. Such an approach would effectively protects the functional aspects of copyrighted works. Finally, U.S. regulators should be aware that such a choice could put the U.S. significantly out of step with other governments around the world, diminishing U.S. competitiveness in artificial intelligence development.