Reuters Legal News

Summarization and ChatGPT: a look at *Silverman v. OpenAI*

By Sasha Rosenthal-Larrea, Esq., Daniel M. Barabander, Esq., and Lucille Dai-He, Esq., Cravath

FEBRUARY 2, 2024

Although ChatGPT was not designed solely for the task of summarization, its generative AI capabilities have incidentally made it a useful tool for prompting summaries of text and concepts. This summarization capability is a focal point in the recent suit *Silverman v. OpenAI, Inc.*, 3:23-cv-03416, (N.D. Cal.), in which plaintiffs, all authors of books, are suing OpenAI, the creator of ChatGPT, claiming that OpenAI had infringed these authors' copyrights by using their works to train GPT, the model underlying ChatGPT.

This case is among several recent cases that have drawn intense interest because the claims test the bounds of copyright law as it relates to the use of copyrighted works to train a machine learning model.

The case raises several factual issues, most critically, how GPT was trained. To make their claim, the plaintiffs must prove that their copyrighted works were actually ingested by GPT during the training process, and that this ingestion constitutes infringing activity.

Regarding whether the copyrighted works were actually ingested by GPT during training, what data exactly was used in training is not publicly documented, so we are unable to know for certain from where GPT learned this information.

It is worth discussing what summarization means in the context of a user's experience of the service. Generally speaking, there are two different ways users can prompt ChatGPT to generate summaries. First, the user can type in a summarization command and then paste in the text she wants summarized (*e.g.*, "summarize this: [pasted content]"). Second, the user can type in a summarization command without providing the thing to be summarized (*e.g.*, "summarize historical scholarship on the war of 1812"). We will refer to the first type as an "in-context summary" and the second type as an "out-of-context summary."

Both types of summaries are generated by the model using the same function. GPT is an "autoregressive language model," which means it generates outputs one "token" (generally, a word or part of a word) at a time, repeatedly using the "context" of the previous tokens, including the prompt, to produce the best next token. It uses a mechanism called "attention" to determine which tokens in the context to focus on.

As a loose analogy for how this attention mechanism works, imagine a human author writing a paper, but before each new word is typed, the author rereads the previous words typed and focuses on the words that are important for the meaning of the word that is about to be typed. The GPT model's objective is always to generate the next best token to follow the sequence of words in the context, regardless of whether the user's prompt includes the text to be summarized, or merely a reference to some outside concept.

> *This case is among several recent cases that have drawn intense interest because the claims test the bounds of copyright law as it relates to the use of copyrighted works to train a machine learning model.*

Yet, because the GPT model attends to context that the user provides in her prompt (along with other background context that we will not discuss in this article), the two types of summary prompts can lead to different inferences about the data GPT uses to reach its next token prediction.

When ChatGPT is asked to produce an in-context summary, the model has as its context both the command (*e.g.*, "summarize this text:") *and* the excerpted text to be summarized. This means the model applies its attention mechanism to the actual text it is asked to summarize. As a consequence, the model identifies information from the text, which it can use to produce the summary output. Thus, if ChatGPT produces an accurate summary in the in-context scenario, it is difficult to know what information in the summary is informed by training data versus the text in the user's prompt.

This differs from when ChatGPT is prompted to produce an out-of-context summary, which is the type of summary the plaintiffs are relying on to make their claims, which include an assertion that GPT ingested their books as part of training (and not that the book texts were inserted as part of a user's prompt). In an exhibit to the Complaint, the plaintiffs provide copies of prompts requesting summaries of plaintiffs' books and summaries, which they allege prove GPT was trained on the plaintiffs' works.

For example, one prompt is: "Summarize in detail the beginning of 'Sandman Slim' by [plaintiff] Richard Kadrey" (and no other context

Thomson Reuters is a commercial publisher of content that is general and educational in nature, may not reflect all recent legal developments and may not apply to the specific facts and circumstances of individual transactions and cases. Users should consult with qualified legal counsel before acting on any information published by Thomson Reuters online or in print. Thomson Reuters, its affiliates and their editorial staff are not a law firm, do not represent or advise clients in any matter and are not bound by the professional responsibilities and duties of a legal practitioner. Nothing in this publication should be construed as legal advice or creating an attorney-client relationship. The views expressed in this publication by any contributor are not necessarily those of the publisher.

is provided). ChatGPT responds, in part, as follows: "The novel begins with the protagonist, James Stark, returning to Earth after 11 years of enforced residency in Hell, during which he was the only living human". If this summary is accurate, as the plaintiffs claim (Compl. ¶ 42), where is that information coming from?

Silverman v. OpenAI is one of the first major copyright cases to claim that the generation of text based on a particular prompt request (e.g., summarization) can dispositively prove that certain data must have been ingested as part of the training process.

A reasonable inference is that this information is captured somehow in the model's weights (parameters adjusted during training that affect the different levels of importance the model assigns to input features). Assuming the version of ChatGPT the plaintiffs used to produce these summaries did not retrieve content related to the books at runtime and consider it in its context, the only words the model is attending to at the outset are the words within the prompt — the request, the level of detail, the part of the book to summarize, the title of the book and the author — but this prompt does not provide any further information about the book itself. Put simply, the model is not directly attending to the text inside the book "Sandman Slim" because it was not part of the user's prompt.

If the information in the summary is captured somehow in the model's weights, there is a further question of where this information came from. One possibility, cited by the plaintiffs, is that GPT was trained on the plaintiffs' copyrighted works. (Compl. ¶ 5.) As another possibility, GPT could have been trained on a large number of reviews or plot summaries found on the internet (not written by Kadrey) of Kadrey's book, including summaries that break down the book chapter by chapter (just as humans could provide a summary of a book without reading the actual book, simply through reading others' summaries of the book on the internet).

In fact, 85% of GPT-3's training data originates from online content (https://bit.ly/4bdFkgS) (Common Crawl, WebText2 and English- language Wikipedia datasets), so it is certainly possible these datasets contained accurate summaries of the books (not written by the plaintiffs) and that these were ingested during training.

Silverman v. OpenAI is one of the first major copyright cases to claim that the generation of text based on a particular prompt request (*e.g.*, summarization) can dispositively prove that certain data must have been ingested as part of the training process. Defendants have not sought dismissal at the pleading stage of the plaintiffs' allegations of direct copyright infringement, so these questions will be resolved at a later stage when Defendants can raise defenses such as fair use, as well as address the conclusions plaintiffs are drawing from the evidence presented.

In navigating these questions of fact and law, knowledge of the function of the underlying technology is essential in order to understand what conclusions can — and cannot — be drawn with full confidence.

About the authors

Sasha Rosenthal-Larrea (L) is a partner in **Cravath**'s corporate department in New York. She focuses her practice on supporting the firm's clients with their most complex intellectual property issues, including complex licensing, collaboration agreements and other commercial arrangements, as well as handling the technology, intellectual property, data privacy and data security aspects of major corporate transactions. She can be reached at srosenthal-larrea@cravath.com. **Daniel M. Barabander** (C) is an associate in the firm's corporate department in New York. He can be reached at dbarabander@cravath.com. **Lucille Dai-He** (R) is a member of the firm's corporate department in New York. She can be reached at ldaihe@cravath.com.

This article was first published on Reuters Legal News and Westlaw Today on February 2, 2024.