AI Discovery Battles Heat up as AI Developer Ordered to Produce Training Data

February 5, 2025

After many rounds of motions to dismiss, intellectual property cases against AI developers are moving into the discovery phase. As we previewed in our 2024 AI year in review, one of the big areas to watch in 2025 will be how much discovery courts are prepared to order into the inner workings of AI companies, especially in the face of arguments that discovery would reveal trade secrets or would be overbroad in cases based on specific claimed works. Just weeks into 2025, we got our first answer, with a court ordering OpenAI to produce a complete training dataset to plaintiffs.

What Happened? On January 27, 2025, a federal judge in the Northern District of California ordered OpenAI to produce a dataset to plaintiffs' counsel that was used by the company to train its generative AI model, GPT-4. The order came in the closely watched *Tremblay v. OpenAI, Inc.* case (which has been consolidated with the similarly situated *Silverman v. OpenAI, Inc.* and *Chabon v. OpenAI, Inc.* cases since last February).[1]

The *Tremblay* plaintiffs, a proposed class of artists and authors, alleged that GPT-4 was trained on datasets containing the plaintiffs' copyrighted works, and that the training of GPT-4 directly infringed plaintiffs' copyrights and amounted to unfair competition under California law. Defendants denied the allegations and stated that to the extent any copying of copyrighted works occurred, that copying constituted fair use.[2]

In a joint discovery letter filed earlier this month, plaintiffs argued that the court should compel the defendants to produce the complete "English Colang" dataset that OpenAI used for training GPT-4, both because it was directly relevant to their copyright infringement claim and also because the production of the dataset would increase the efficiency of the discovery process. Plaintiffs claimed that OpenAI had already produced two key training datasets which allegedly contained pirated copies of plaintiffs' works—

---

1 *Tremblay v. OpenAI, Inc.*, *Silverman v. OpenAI, Inc.*, *Chabon v. OpenAI, Inc.*, Stipulation and Order Consolidating Cases (N.D. Cal., 2/16/24).

2 *Tremblay v. OpenAI, Inc.*, Defendants' Answer to First Consolidated Amended Complaint, (N.D. Cal., 8/27/24) (Defendants also raised other affirmative defenses, including valid express or implied licenses, waiver, estoppel, innocent infringement, statute of limitations, failure to mitigate damages, the doctrine of merger).

LibGen 1 and 2—and defendant's proposed compromise to produce only a subset of the English Colang training dataset consisting of documents over 20,000 words or more was insufficient to identify all instances of infringement.[3] Defendants responded that their 20,000-word solution under the existing protective order was a reasonable compromise and that production of the entire dataset would be burdensome and expose the dataset on a less secure system.

Ultimately, the court agreed with plaintiffs, ordering OpenAI to produce the complete, native English Colang dataset. While that means the *Tremblay* plaintiffs will gain access to the complete training data, the court was clearly sympathetic to the data security concerns raised by OpenAI during the briefing and argument. The court granted OpenAI's request to seal certain information regarding the English Colang dataset which OpenAI considers to be proprietary, such as the size and technical specifications, and ordered the parties to file an updated security protocol under seal to ensure that the dataset would be adequately protected from cyber breaches. The court's order has echoes of last year's order in the high-profile litigation between *The New York Times* and OpenAI, in which the court permitted discovery into the source code for ChatGPT but only in a secure room entirely disconnected from the Internet.

What to Expect. Many of the earliest-filed copyright infringement cases against AI developers are well into discovery, and the parties and the courts are grappling with the practical implications of a myriad of questions: what datasets are relevant, what should be sealed, how datasets should be protected and secured, and in what format the requested productions should be produced.

The battle over the production of training datasets, which AI developers consider to be commercially sensitive and proprietary, will surely continue to be hard-fought in other cases. Recent victories for plaintiffs, including in *Tremblay*, suggest that courts view training data to be central to the question of direct copyright infringement, a claim which has survived every motion to dismiss thus far.

In the *Kadrey v. Meta* case, for example, a court recently ruled that the plaintiffs' request for additional datasets used to fine-tune AI models was firmly within the scope of the plaintiffs' copyright infringement claims. In the words of the *Kadrey* court, "Plaintiffs have made a sufficient factual showing that the use of datasets that contain copyrighted works to create datasets that were used in fine-tuning the Llama models concerning intellectual property violations may have or could have resulted in portions of the copyrighted works ending up in the fine-tuning datasets, such that Plaintiffs are entitled to learn if that in fact happened."[4] The *Kadrey* court accordingly granted plaintiffs'

---

[3] *Tremblay v. OpenAI, Inc.*, Joint Discovery Letter Brief, (N.D. Cal 1/17/2025).

[4] *Kadrey v. Meta Platforms, Inc.*, Public Version of Discovery Order at ECF No. 374 at 3, (N.D. Cal. 1/17/25).

motion to compel—though the judge did find that requiring the production of raw or original data would be unduly burdensome due to the size of the datasets.[5]

As these cases proceed through discovery, AI developers are likely to continue to resist discovery into their proprietary training data, especially in more burdensome and costly formats, while plaintiffs continue to expand the scope of their discovery requests in order to get greater insight into the training process. Courts have repeatedly noted the broad scope of U.S. civil discovery, however, and we may see other courts follow the recent *Tremblay* decision's balance between disclosure to plaintiffs while ensuring data security and restricted access to the broader public through protective orders.

Megan K. Bannigan
Partner, New York
+ 1 212 909 6127
mkbannigan@debevoise.com

Christopher S. Ford
Counsel, New York
+ 1 212 909 6881
csford@debevoise.com

Samuel J. Allaman
Associate, New York
+ 1 212 909 6026
sjallaman@debevoise.com

Abigail Liles
Associate, New York
+ 1 212 909 6387
aeliles@debevoise.com

[5] *Kadrey v. Meta Platforms, Inc.*, Public Version of Discovery Order at ECF No. 374, (N.D. Cal. 1/17/25).