ALM | LAW.COM

CORPORATE COUNSEL

Discovery of Training Data in AI Litigation

By Sasha S. Rao and Richard A. Crudo

April 30, 2025

Although AI litigation is still in its infancy, discovery disputes are now emerging. In AI copyright cases, for example, parties have disputed the discoverability of data used to train defendants' AI models, as well as the protocols that should govern the review of such data. Recognizing the data's relevance to infringement, courts have compelled its production. Thus, defendants facing likely disclosure of training data must consider how best to protect it.

In approaching this issue, litigants have relied on their experience with source-code-inspection protocols. But the unprecedented scale and complexity of AI data require adaptation of such protocols. This article discusses those adaptations as well as opportunities to leverage the unique properties of training data to protect it during litigation.

The commercial value of training data

Training data refers to the text, images, audio, and other information fed into algorithms to develop and train AI models. Sources of such data can include public datasets, proprietary or internal datasets, and content generated from partnerships with content holders. Better-quality and larger training data leads to higher-performing models.

Credit: Spot Decor/Adobe Stock

Training data is a highly valuable commercial asset. AI companies invest significant resources and expenses (up to tens of millions of dollars) to collect, curate, and structure this data to create their models. In addition to these high costs, gathering the data can be challenging given the time required and potential inconsistency or bias in available data. Thus, AI companies often treat the data as a valuable asset whose disclosure could cause competitive harm.

The value of training data in discovery

Training data can also be valuable to litigants in copyright-infringement actions. For example, copyrighted works can be swept up in training

data when a company gathers the data. And unauthorized inclusion of copyrighted works in the training data—and the AI's use of such works to generate substantially-similar works—can be an act of copyright infringement.

Accordingly, training data is highly relevant in AI-based copyright litigation. In *Thomson Reuters Enter. Ctr. GMBH v. Ross Intel. Inc*., for example, access to the defendant's training data led to summary judgment of infringement because it allowed the court to see instances of the copied work in the data. No. 1:20-cv-613-SB, 2025 WL 458520, at *5 (D. Del. Feb. 11, 2025). Indeed, the potential for training data to uncover evidence of infringement has led courts to compel its production. Thus, defendants should be prepared to disclose their training data in litigation, particularly in copyright matters.

Training-data inspection protocols: challenges and opportunities

With the disclosure of training data likely, parties are stipulating to inspection protocols that govern a plaintiff's ability to request and review a defendant's data. OpenAI introduced one of the first standalone training-data inspection protocols in *Tremblay v. OpenAI, Inc*. No. 3:23-03223, D.I. 182 (N.D. Cal. Sept. 24, 2024), which has been borrowed in other cases as well.

Most of these protocols mirror source-code-review protocols that litigants are very familiar with. For example, training-data protocols typically restrict review to individuals with access to the highest level of confidentiality under a protective order and may additionally require an NDA. *Id.* at ¶ 7.i. Training-data protocols also often require inspection on a standalone computer in a secure room. Reviewers are typically permitted to take notes and request printouts of the data but may not copy the data directly. *Id.* at ¶¶ 7.f, 7.h.

But the unprecedented scale and complexity of AI data will likely require adaptation of these protocols, as described below.

Scale and Volume. A major difference between source code and AI training data is the scale and volume of the latter compared to the former. For example, Meta's Llama-3 was trained on 15 trillion tokens (i.e., fundamental units of data that AI reads and learns from). "Introducing Meta Llama 3: The most capable openly available LLM to date," AI at Meta Blog (April 18, 2024) (available at https://ai.meta.com/blog/meta-llama-3/). Meanwhile, Meta's Facebook app had about 100 million lines of code. Lily Newman, "How Facebook Catches Bugs in its 100 Million Lines of Code," WIRED (Aug. 15, 2019). This scale makes it difficult to simply "hand over" raw data in any human-reviewable form. The plaintiffs in *Trembley v. OpenAI*, for example, were forced to cancel their search for just one of their copyrighted works while reviewing one of OpenAI's training datasets during discovery because it was going to take more than six hours to complete. No. 3:23-cv-03223, D.I. 254, at 1 (N.D. Cal. Jan. 17, 2025).

Defendants can use their training data's size to their advantage. Specifically, courts may be sympathetic to proportionality arguments when the data is burdensome to produce. For example, in *Kadrey v. Meta Platforms, Inc.*, the court limited discovery to the refined post-training data, recognizing that the raw/original data was "massive compared to the datasets actually used" and not proportional to the needs of the case. No. 23-cv-03417, D.I. 399, 4–5 (N.D. Cal. Sept. 9, 2024).

Further, parties can negotiate shortcuts for reviewing raw training data. For example, rather

than producing the raw data itself, a defendant can generate hash values or an index for all works in the dataset and allow the plaintiff to compare its copyrighted work against that index. AI companies, anticipating litigation, could also document their training efforts (e.g., what data was used, from where, how it was filtered, etc.). This type of information was sufficient for the court in *Kadrey* to determine that the plaintiff did not need certain raw data itself. *Id.* These types of shortcuts thus benefit both parties—defendants avoid opening up their entire training datasets to review, while plaintiffs can review the data more efficiently.

Heterogeneity and Accessibility. Also, unlike source code, which is typically written in a single programming language, training data can use multiple languages and might require different types of indices and tools for facilitating review of different data types and formats.

For example, the organization of training data can vary by data type (e.g., text, image, audio), granularity (e.g., full documents, token sequences), labeling structure (e.g., supervised, unsupervised), data source, and training purpose. Given these numerous ways to organize training data, most protocols require technical guidance for efficient review. For example, the producing party may agree to supply a README file that provides a directory of the data and describes the layout, format, and means of searching the data. *Tremblay*, D.I. 182, at ¶ 7.a.

Additionally, many protocols require defendants to provide software that plaintiffs can use to effectively view and search the training data. For example, in one recent case, an audio AI company allowed plaintiffs to use Audible Magic—content identification software—to analyze the audio files in their training data. *UMG Recordings, Inc. v. Uncharted Labs, Inc. d/b/a/ Udio.com*, No. 1:24-cv-04777, D.I. 82 (S.D.N.Y. March 17, 2025).

Conclusion

Training data's importance, both as a commercial asset and in litigation, demands an approach that balances protection with access during discovery. Achieving this balance requires accounting for the unique size, formats, and sources of the training data. To date, parties have borrowed from source-code-review protocols, but the needs of AI litigation go beyond such protocols. Thus, creative solutions are needed by litigants to ensure efficient review while minimizing overproduction and misuse.

Sasha S. Rao *is counsel in Sterne Kessler's Trial & Appellate Practice Group, where he has been involved in various matters before federal district courts, the International Trade Commission and the U.S. Court of Appeals for the Federal Circuit. His experience spans from initial fact investigation through trial in a wide range of technologies.* ***Richard A. Crudo*** *is a director in Sterne Kessler's Electronics and Trial & Appellate Practice Groups and has more than a decade's worth of experience litigating intellectual property cases. He has represented clients from a broad range of industries in high-stakes cases before the Supreme Court, the Federal Circuit and the district courts.*