IP Alert | Britannica v. Perplexity: Newly-Filed Lawsuit Questions Legality of Scraping, Grounding of Machine Learning Models

By Kirk Sigmon and Jacob Hinish

Encyclopedia Britannica, Inc. and Merriam-Webster, Inc. have sued Perplexity AI, Inc., alleging copyright infringement related to Perplexity's alleged use of their online content to "ground" its AI-powered "answer engine." This case has significant potential to refine how copyright and fair use law addresses web scraping and generative AI grounding. Perhaps more importantly, a ruling in favor of Encyclopedia Britannica and Merriam-Webster could have a significant adverse impact on Large Language Model ("LLM") developers that scrape public websites to collect model training data.

Background

Perplexity provides an "AI-powered answer engine" that retrieves online content in real-time and uses it to strengthen machine learning-generated responses to user queries. This augmentation process, referred to as "retrieval-augmented generation" ("RAG") or "grounding," allegedly enables Perplexity's model to provide more accurate answers to research questions (in contrast to conventional LLMs, which can often provide incorrect or "hallucinated" responses).

Encyclopedia Britannica and Merriam-Webster allege that Perplexity's model infringed their copyrighted works (such as Encyclopedia Britannica's online encyclopedia) in three ways:

1. When Perplexity scraped and crawled their websites;
2. When Perplexity's model uses that scraped information as input to generate responses to user queries; and
3. When Perplexity's model provides output allegedly substantially similar to Encyclopedia Britannica and Merriam-Webster's copyrighted articles.

Case Raises Questions regarding Legality of Scraping and Use of Scraped Data

This case may be the first case to decide whether the use of copyrighted material to ground LLMs infringes copyright and/or whether that grounding may comprise fair use. Cases such as

Associated Press v. Meltwater U.S. Holdings, Inc., 931 F.Supp.2d 537 (S.D.N.Y. 2013) are often cited for the proposition that mere scraping and re-delivery of online content is not sufficiently transformative to warrant fair use protection; however, cases such a Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir., 2007) ("Perfect 10") have often been cited as permitting framing and hyperlinking of original content (there, copyrighted images) to effectuate an image search engine (there, Google's image search). It is unclear how such cases might be interpreted in view of LLM grounding: on one hand, grounding arguably involves scraping and re-delivery; on the other hand, Perplexity's grounding is arguably used to effectuate searches.

The case might also hinge on the legality and intent of Perplexity's scraping. As discussed in our previous IP Alerts, at least two cases (Bartz et al. v. Anthropic PBC, No. 24-cv-05417, Doc. No. 231 (N.D. Cal. June 23, 2025) and Kadrey et al. v. Meta Platforms, Inc., No. 23-cv-03417, Doc. No. 598 (N.D. Cal. June 25, 2025)) have held that the use of legally obtained copyrighted material to train a machine learning model may constitute fair use. That said, it is not immediately clear whether Perplexity's scraping would constitute legally obtaining such content, given that Encyclopedia Britannica and Merriam-Webster allege that Perplexity "ignored or evaded" technological features to prevent crawling/scraping on their websites. Moreover, at least one court (in Thomson Reuters Enterprise Center GMBH v. Ross Intelligence Inc., 765 F.Supp.3d 382 (D. Del. 2025)) has held that fair use might not protect use of non-licensed material to generate a directly competitive market substitute. Along those lines, Encyclopedia Britannica and Merriam-Webster's complaint suggests that they view Perplexity's use of their content as an attempt to provide a competitive market substitute.

It is equally unclear whether the Court will find that Perplexity's outputs are copyright-infringing. A similar battle is being fought in California in Disney Enterprises v. Midjourney Inc., No. 2:25-CV-05275 (C.D. Cal. 2025) ("Midjourney"), where complainants such as Universal, Disney, DreamWorks, Marvel, Lucasfilm, and Twentieth Century Fox allege that Midjourney (and, for instance, Midjourney's video generation functionality) can be used to create copyright-infringing content (e.g., videos that use popular Disney or Marvel characters). This question is made even more complicated given that the scraped content (such as encyclopedia entries and dictionary entries) arguably contains basic facts, potentially affording that content less protection under U.S. copyright law.

Looking Ahead

Developers of machine learning models (including LLMs) should monitor this case closely: the Court's ruling could have significant implications regarding the legality of both LLM "grounding" as well as the use of scraped online data to train machine learning models. This case could also provide valuable guidance regarding the limits of fair use to develop alleged market substitutes. As we have recommended before, organizations training machine learning models should carefully audit the training data they have used (or plan to use), recognizing that rulings in this and similar cases could drastically affect the legality of their operations.

Posted: November 13, 2025