Canadian News Outlets Seek What Could Amount to Billions From OpenAI in New Copyright Infringement Case | ArentFox Schiff

Dan Jasnow

Allegations and Legal Claims

Copyright Infringement

The core allegation made by the Canadian news and media companies is that OpenAI's method of training its GPT LLMs infringes the publishers' copyrights. To train an LLM, developers like OpenAI rely on large data sets. In the case of the GPT LLMs, this included datasets such as Webtext, Webtext2, and CommonCrawl, all of which included copies of text data scraped from the internet, including news articles. The models analyze these data sets to learn to generate natural-sounding text.

Like their US counterparts, the Canadian publishers allege that the act of scraping content from the internet and storing copies of that content in data sets infringes the publishers' copyrights. Notably, however, in their US court filings, OpenAI alleges that these copyright infringement claims are time-barred because the underlying conduct, presumably the scraping and creation of the initial data sets, occurred more than three years ago.

In a possible attempt to preempt this defense, the Canadian publishers' go a step further, alleging that it's not just the initial training data that infringes their rights, but also the process OpenAI uses to augment its models after they have been developed and released to the public. Specifically, the publishers allege that through a process known as "Retrieval-Augmented Generation" (RAG), OpenAI provides its models with "continuous access" to an additional data set which is "continually updated" in response to user prompts. Unlike the initial data sets, the publishers allege that OpenAI continuously updates the RAG data by scraping and/or copying information from the internet, including by repeatedly scraping or copying the publishers' websites and copyrighted works.

Even if not time-barred, it remains an open question whether using copies of "scraped" content to train an LLM is actionable infringement under copyright law. Many artificial intelligence (AI) developers argue that it is not because the scraped content is used only for internal purposes and does not result in unauthorized dissemination of the copyrighted works. In other words, the view taken by many AI developers is that the copying serves merely as an intermediate technical step in an analytical process that results in a transformative new technology, not in the dissemination of an author's original expression to a new audience.

Circumventing Technological Protections

The publishers separately allege that in order to access and scrape the publishers' copyrighted content, OpenAI circumvented safeguards like paywalls, subscriptions, or copyright disclaimers intended to prevent "scraping" or other unauthorized copying of protected content. Similar allegations have been made by US rights holders under the Digital Millennium Copyright Act, which expressly prohibits circumvention of such technological protection measures.

Breach of Terms of Use

Finally, the plaintiffs allege that OpenAI breached their respective terms of use by scraping and reproducing the publishers' content for unauthorized commercial purposes. Specifically, the publishers allege that OpenAI was on notice of each site's terms of use and agreed to comply with those terms when it accessed and used the plaintiffs' websites, and that since 2015, the relevant terms have clearly stated that the content is for the "personal, non-commercial use of individual users only."

Relief

The plaintiffs are seeking what could amount to a substantial amount in damages — $20,000 per article they claim was illegally "scraped" and used to train OpenAI's GPT LLMs. They are also seeking a share of the profits OpenAI has made through the alleged infringement and are asking the court to prevent OpenAI from using their content in the future.

Industry Impact and Future Considerations

The outcome of this lawsuit may have dramatic effects on not only OpenAI but also the larger relationship between the AI industry and the news and media industries. The immediate implications for OpenAI are the damages it could face, potentially totaling billions of dollars, and possibly a need to reassess the ways its LLMs are trained.

Meanwhile, as these lawsuits play out in the United States and Canada, OpenAI has entered into licensing agreements with news and media organizations to license their content for training purposes. This move creates a potentially lucrative market for rights holders while providing an alternative and potentially less risky source of content for OpenAI. But a finding in this case or others that OpenAI's use of third-party content to train its models is fair dealing or fair use under Canadian or US law, respectively, may stunt the emerging licensing market.

ArentFox Schiff will continue to monitor these issues and is available to answer any questions you may have. For additional information, please contact the authors or the attorney who usually handles your matters.