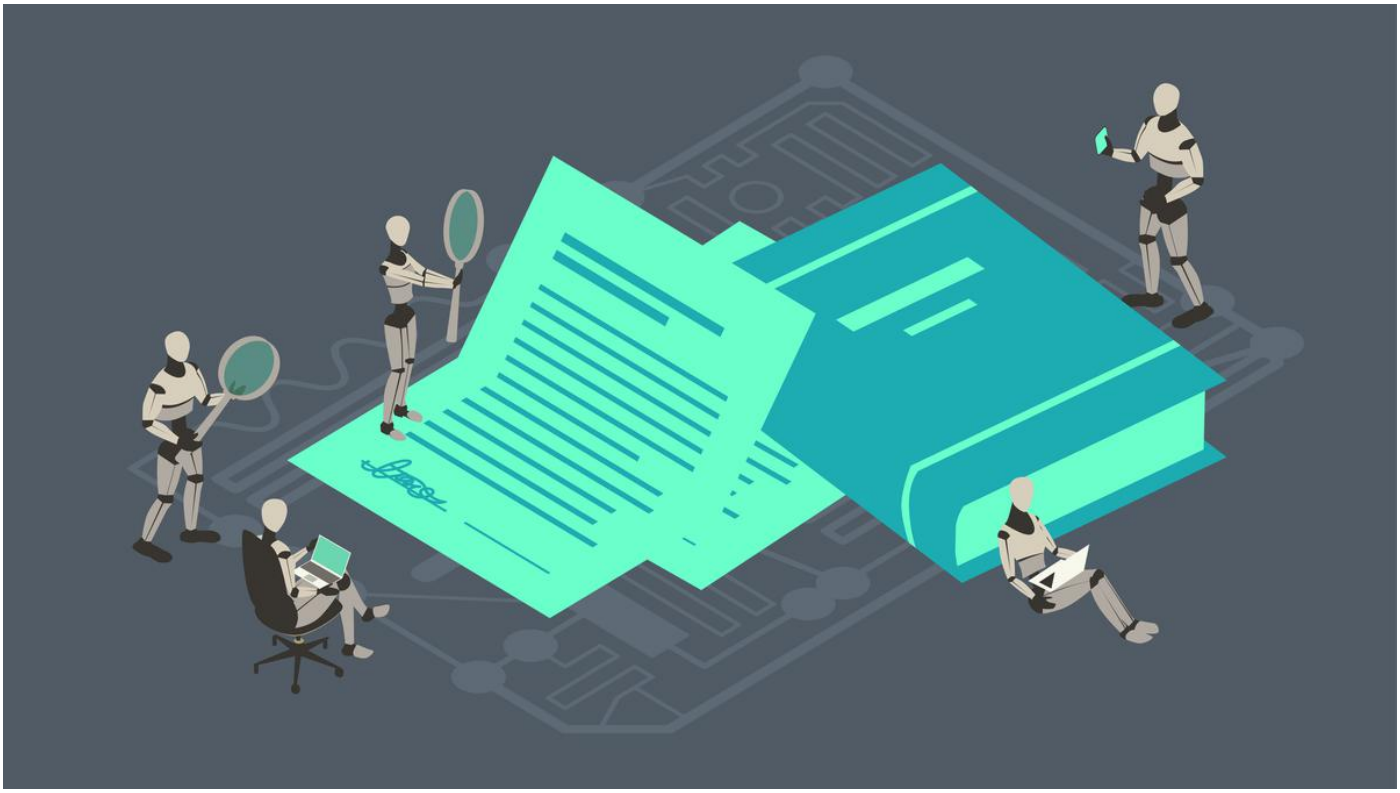
Explained | News media versus OpenAI's ChatGPT

Why have major news agencies and newspapers like The New York Times, Reuters and CNN blocked the GPT bot? What are crawlers and how do they help large language models?

August 28, 2023 10:40 pm | Updated August 29, 2023 10:35 am IST

ANUJ SRIVAS



For representative purposes. | Photo Credit: Getty Images

The story so far: A group of news media organisations, including *The New York Times*, Reuters, CNN and the Australian Broadcasting Corporation, recently shut off OpenAI's ability to access their content. The development comes in the wake of reports that The New York Times is planning on suing the artificial intelligence (AI) research company over copyright violations, which would represent a considerable escalation in tensions

between media companies and the leading creator of generative artificial intelligence solutions.

What does OpenAI do?

The company is best known for creating 'ChatGPT', which is an AI conversational chatbot. Users can ask questions on just about anything, and ChatGPT will respond pretty accurately with answers, stories and essays. It can even help programmers write software code. The hype around ChatGPT — specifically, the breathtaking advancements in the field of AI required to create it — has propelled OpenAI into becoming a $30 billion company.

What started the face-off between news outlets and OpenAI?

Software products like ChatGPT are based on what AI researchers call 'large language models' (LLMs). These models require enormous amounts of information to train their systems. If chat bots or digital assistants need to be able to understand the questions that humans throw at them, they need to study human language patterns. Tech companies that work on LLMs like Google, Meta or Open AI are secretive about what kind of training data they use. But it's clear that online content found across the Internet, such as social media posts, news articles, Wikipedia, e-books, form a significant part of the dataset used to train ChatGPT and other similar products. This data is put together by scraping it off the Internet. Tech companies use software called 'crawlers' to scan web pages, hoover up content and put it together in a dataset that can be used to train their LLMs.

This is what news outlets took a stand against last week when *The New York Times* and others blocked a web crawler known as GPT bot, through which OpenAI used to scrape data. They told OpenAI that the company can no longer use their published material and their journalism, to train their chat bots.

Why are media companies upset?

Search engines like Google or Bing also use web crawlers to index websites and present relevant results when users search for topics. The only difference is that search engines represent a mutually beneficial relationship. Google, for instance, takes a snippet of a news article (a headline, a blurb and perhaps a couple of sentences) and reproduces them to make its search results useful. And while Google profits off of that content, it also directs a significant amount of user traffic to news websites.

OpenAI, on the other hand, provides no benefit, monetary or otherwise, to news companies. It simply collects publicly available data and uses it for the company's own purposes.

"Anyone who wants to use the work of *Wall Street Journal* journalists to train artificial intelligence should be properly licensing the rights to do so from Dow Jones," Jason Conti, general counsel for News Corp.'s Dow Jones unit, said in a statement earlier this year.

But it's also true that some news outlets probably view ChatGPT as a potential competitor that will profit off their journalism. After all, if you ask ChatGPT to describe the coffee and food served by the best cafes on Manhattan's Upper East Side, the answer probably comes from some AI-generated mixture of reporting done by *The New York Times*' features team and reviews put out by food-centric publications.

What is the way forward?

Looking ahead, there are two key questions to be answered. If your data was used to train ChatGPT without permission or compensation, have your rights been violated? And just how much can companies like OpenAI pay out before it makes the whole endeavor financially unfeasible? Tech gurus like to argue that the value of online content only exists in the aggregate. Or in other words, ChatGPT could still exist as a high-quality product without CNN's reporting. But if all media publications across the world refused to provide access to OpenAI, it's likely that the final product would be of lower quality. And, of course, if every single creator of online content turned down OpenAI, then ChatGPT would almost certainly not exist.

However, at the same time, it's clear that OpenAI does believe some data is worth paying for. Last month it signed a licensing arrangement with The Associated Press, in a deal that would allow the company to use the news agency's archival content as a training dataset. But what happens when people refuse to accept payment and sue OpenAI for copyright infringement, the way a group of novelists did last year? The legal battles ahead will have interesting implications for journalism, intellectual property and the future of artificial intelligence.

(Anuj Srivas is a freelance writer.)