Opinion: OpenAI's content deal with Financial Times is an attempt to avoid legal challenges—and an AI 'data apocalypse'

May 9 2024, by Mike Cook

Credit: Unsplash/CC0 Public Domain

OpenAI's new ["strategic partnership"](#) and licensing agreement with the Financial Times (FT) follows similar deals between the US tech

company and publishers such as Associated Press, German media giant Axel Springer and French newspaper Le Monde.

OpenAI will license the FT's content to use as training data for its products, including successors to its AI chatbot ChatGPT. The AI systems developed by OpenAI are exposed to this data to help them improve their performance in terms of use of language, context and accuracy. The FT will receive an undisclosed payment as part of the deal.

This is happening against a global backdrop of legal challenges by media companies alleging copyright infringement over the use of their content to train AI products. The most high-profile of these is a case brought by the New York Times against OpenAI. There is also a fear among tech companies that, as they build more and more advanced products, the internet will no longer have enough high-quality data to train these AI tools.

So, what will this deal mean for the FT? There's still a lack of detail on partnerships like this one, apart from the fact the FT will be paid for its content. However, there are hints of other potential benefits.

In a statement, the FT Group's chief-executive, John Ridding, emphasized that the paper was committed to "human journalism." But he also acknowledged that the news business can't stay still: "We're keen to explore the practical outcomes regarding news sources and AI through this partnership … We value the opportunity to be inside the development loop as people discover content in new ways."

The FT has previously said it would "experiment responsibly" with AI tools, and train journalists to use generative AI for "story discovery."

OpenAI is probably keen to announce this partnership because it hopes it

will help solve the most acute problems facing its flagship products. The first is that these generative AI tools sometimes make things up, a [phenomenon known as hallucination](#). Using reliable content from the FT and other trusted sources should help with that.

The second problem is that it could help offset the legal scrutiny that OpenAI faces. Signing official deals with news sources provides the tech company with some reputational damage control, as it shows them trying to make good with the world of journalism. It also potentially provides more legal security going forward.

The licensed content from the FT—and other media sources—could provide ChatGPT and the upcoming GPT-5 with more specific, referenced responses to users. Gemini, Google's ChatGPT competitor, already attempts to do this by providing Google searches that support the claims it makes. Getting results directly from the source means OpenAI has more reliable evidence to search through and be trained on.

This appears to follow the trend of ["retrieval-augmented generation"](#) (RAG) that is becoming more popular in the AI world. RAG is a technique whereby a large language model (the technology that sits behind AI chatbots such as ChatGPT) can be provided with a database of knowledge which can be searched to support what the chatbot already knows. This is a bit like taking an exam with a textbook open in front of you.

This helps reduce the risk of hallucination, where the AI authoritatively produces a response that looks real but is actually made up. Having access to a database of trusted journalism helps offset the [reliability problems with AI products](#) as a result of them being trained on the open internet.

Partnership program

There's a subtext to this global media partnerships program that isn't about the law or ethics. OpenAI needs more and more data as time goes on to keep delivering big improvements through upgrades to its AI products. Yet these products are running out of high-quality training data from the open internet.

This is, at least in part, because there is now a proliferation of content made by AI on the web. This potentially undermines OpenAI's continual need to prove to its partners, governments and investors that it can deliver big improvements to its flagship products.

The New York Times lawsuit maintains that products such as ChatGPT threaten the business of media companies. Whatever the outcome of this case, it is in OpenAI's interests to keep its sources of training data, including media companies, productive and economically viable. The success of ChatGPT, at least for now, is very much tied to the success of the people and organizations producing the data that makes it useful.

PR from the AI industry has done much to foster the idea of inevitability: that AI, in the form of products such as ChatGPT, will transform industries—and people's lives in general. Yet technology fails all the time. The FT deal highlights the dynamic tension that exists between AI and the industries it is changing. ChatGPT now needs the trustworthy journalism that its own generative capabilities and training methods have helped to undermine.

The idea that generative AI has poisoned the internet is nothing new. Some AI researchers have likened the spread of AI-generated junk on the internet to how radioactive contamination of metals forced steel manufacturers in the 1950s to go diving for steel from wrecked ships that had been manufactured before the nuclear age. This pre-nuclear steel was needed for certain uses, such as in particle accelerators and Geiger counters.

In a similar way, for OpenAI and companies like it, training its products on data "scraps" does not seem like a viable way forward.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation