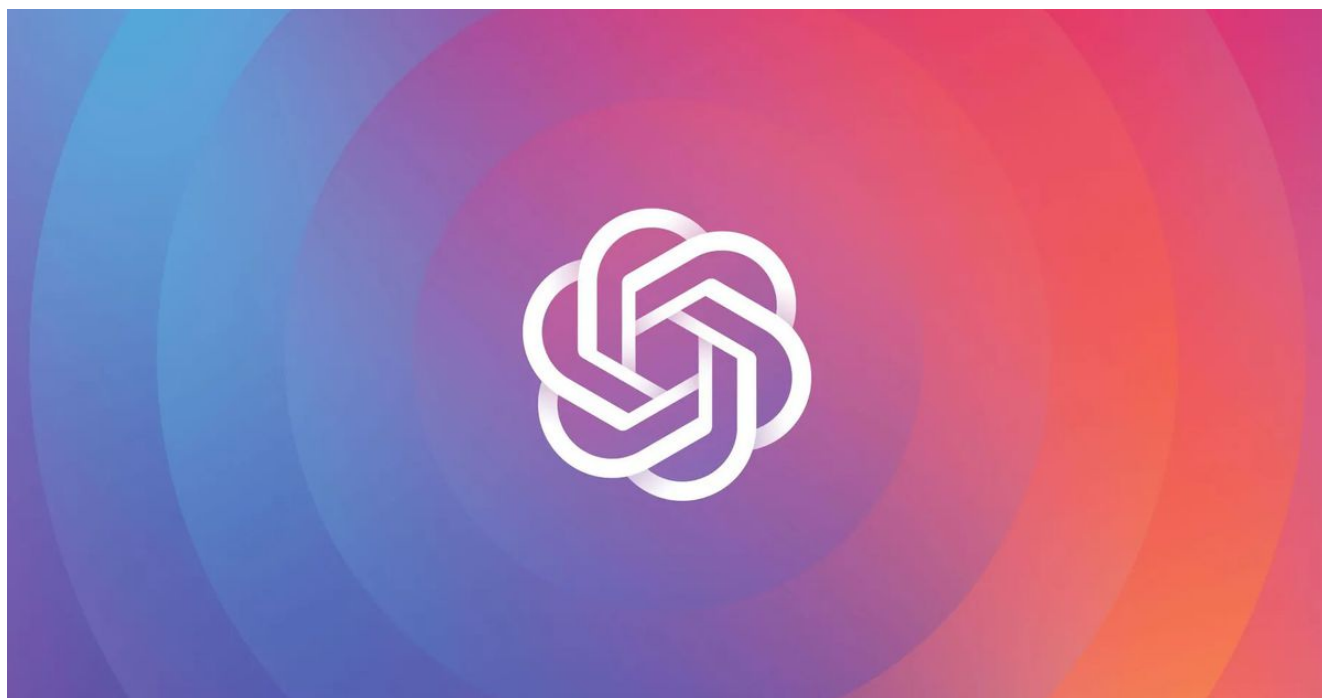


OpenAI's news publisher deals reportedly top out at \$5 million a year



As news publishers ink deals with AI companies to train their models with news stories, the price businesses like OpenAI are willing to pay for copyrighted information is coming to light.

The Information reports that OpenAI offers between \$1 million and \$5 million a year to license copyrighted news articles to train its AI models. That's one of the first indications of how much AI companies plan to pay for licensed material. It sits alongside a recent report saying Apple is looking to partner with media companies to use content for AI training and is offering at least \$50 million over a multiyear period for data. *The Verge* reached out to OpenAI for comment on the numbers.

The numbers appear roughly similar to some earlier non-AI licensing deals. When Meta launched the Facebook News tab – since discontinued in Europe – it allegedly offered up to \$3 million a year to license news stories, headlines, and

previews. But it's not clear whether the total payouts would equal some of the bigger numbers we've seen. Google announced in 2020 that it would invest \$1 billion in total to partner with news organizations, for instance. Under pressure from a new law, Google also recently agreed to pay Canadian publishers a total of \$100 million annually in exchange for linking to their articles.

Today's large language models have, insofar as we know what's in their training data, mainly been trained on information from the internet. While some AI models do not disclose how they got their training data, information is often available on which datasets or web crawlers were used. Pricing for training datasets varies by provider, size, and the content of a dataset. Some data providers, like LAION, are open source and completely free and are used by models like Stable Diffusion. AI developers also often set up web crawlers that take data around the internet to help train their models. (AI developers still have to hire people to vet, tag, and sometimes clean up training data, which significantly adds to operating costs.)

But this practice now faces major challenges. For one thing, OpenAI's GPT crawler has been blocked from accessing data by some companies, including *The New York Times* and *The Verge's* parent company, Vox Media. For another, several organizations argue that training on their data constitutes copyright infringement. *The New York Times*, among others, has sued OpenAI and Microsoft for copyright infringement, alleging that ChatGPT and Microsoft's Copilot can generate output almost verbatim to its work.

Striking partnerships lets AI companies avoid these issues, and it's become a more common practice over the past year. Publishers like Axel Springer – the parent company of *Politico* and *Business Insider* – and *The Associated Press* have signed deals with OpenAI to license stories to train models like GPT-4 and develop technology for news gathering.

OpenAI and Apple aren't the only AI developers hoping to work with news organizations. Google reportedly demoed an AI tool called Genesis that takes facts and spits out news stories to executives from *The New York Times*, *The Wall Street Journal*, and *The Washington Post*. Some news organizations, meanwhile, have used generative AI tools in newsrooms with mixed results.

<https://www.theverge.com/2024/1/4/24025409/openai-training-data-lowball-nyt-ai-copyright>