**A.I. brings shadow libraries into the spotlight**

**Deal Book, International New York Times, 22 July 2023**

[Large language models](https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html), or L.L.M.s, the artificial intelligence systems that power tools like ChatGPT, are developed using enormous libraries of text. Books are considered especially useful training material, because they’re lengthy and (hopefully) well-written. But authors are starting to push back against their work being used this way.

This week, [more than 9,000 authors](https://authorsguild.org/news/more-than-8000-authors-sign-the-authors-guild-letter-calling-on-ai-industry-leaders-to-protect-writers/), including Margaret Atwood and James Patterson, called on tech executives to stop training their tools on writers’ work without compensation.

That campaign has cast a spotlight on an arcane part of the internet: so-called shadow libraries, like Library Genesis, Z-Library or Bibliotik, that are obscure repositories storing millions of titles, in many cases without permission — and are often used as A.I. training data.

**A.I. companies have acknowledged in research papers that they rely on shadow libraries.**[OpenAI’s GPT-1](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf) was trained on BookCorpus, which has over 7,000 unpublished titles scraped from the self-publishing platform Smashwords.

To [train GPT-3](https://arxiv.org/pdf/2005.14165.pdf), OpenAI said that about 16 percent of the data it used came from two “internet-based books corpora” that it called “Books1” and “Books2.” According to a [lawsuit by the comedian Sarah Silverman](https://llmlitigation.com/pdf/03416/silverman-openai-complaint.pdf) and two other authors against OpenAI, Books2 is most likely a “flagrantly illegal” shadow library.

**These sites have been under scrutiny for some time.** The Authors Guild, which organized the authors’ open letter to tech executives, [cited studies](https://www.judiciary.senate.gov/imo/media/doc/Preston%20Testimony.pdf) in 2016 and 2017 that suggested text piracy depressed legitimate book sales by as much as 14 percent.

[Efforts to shut down these sites](https://news.bloomberglaw.com/ip-law/rampant-shadow-libraries-drive-calls-for-anti-piracy-action) have floundered. Last year, the F.B.I., with help from the Authors Guild, charged two people [accused of running Z-Library](https://www.vice.com/en/article/n7z5m7/feds-arrest-two-russians-behind-worlds-largest-library-of-pirated-books) with copyright infringement, fraud and money laundering. But afterward, some of these sites were [moved to the dark web](https://www.vice.com/en/article/v7vnn4/shadow-libraries-are-moving-their-pirated-books-to-the-dark-web-after-fed-crackdowns) and torrent sites, making it harder to trace them. And because many of these sites are run outside the United States and anonymously, actually punishing the operators is a tall task.

**Tech companies are becoming more tight-lipped about the data used to train their systems.** This week, Meta researchers published a [paper on Llama 2](https://scontent-sjc3-1.xx.fbcdn.net/v/t39.2365-6/10000000_6495670187160042_4742060979571156424_n.pdf?_nc_cat=104&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=GK8Rh1tm_4IAX9f0Yy6&_nc_ht=scontent-sjc3-1.xx&oh=00_AfAsEzSFQ-t1u1IG1dDqiiKaiHwZADGEpOCvneSc5Yt2_g&oe=64BBD830), the company’s L.L.M., that described using only a “new mix of data from publicly available sources.” In a [research paper on GPT-4](https://arxiv.org/pdf/2303.08774.pdf) published in March, OpenAI explicitly noted that it wasn’t revealing anything about how it trained the L.L.M., citing “the competitive landscape” and “safety considerations.”