The Atlantic

Sign In Subscribe

More From Artificial Intelligence

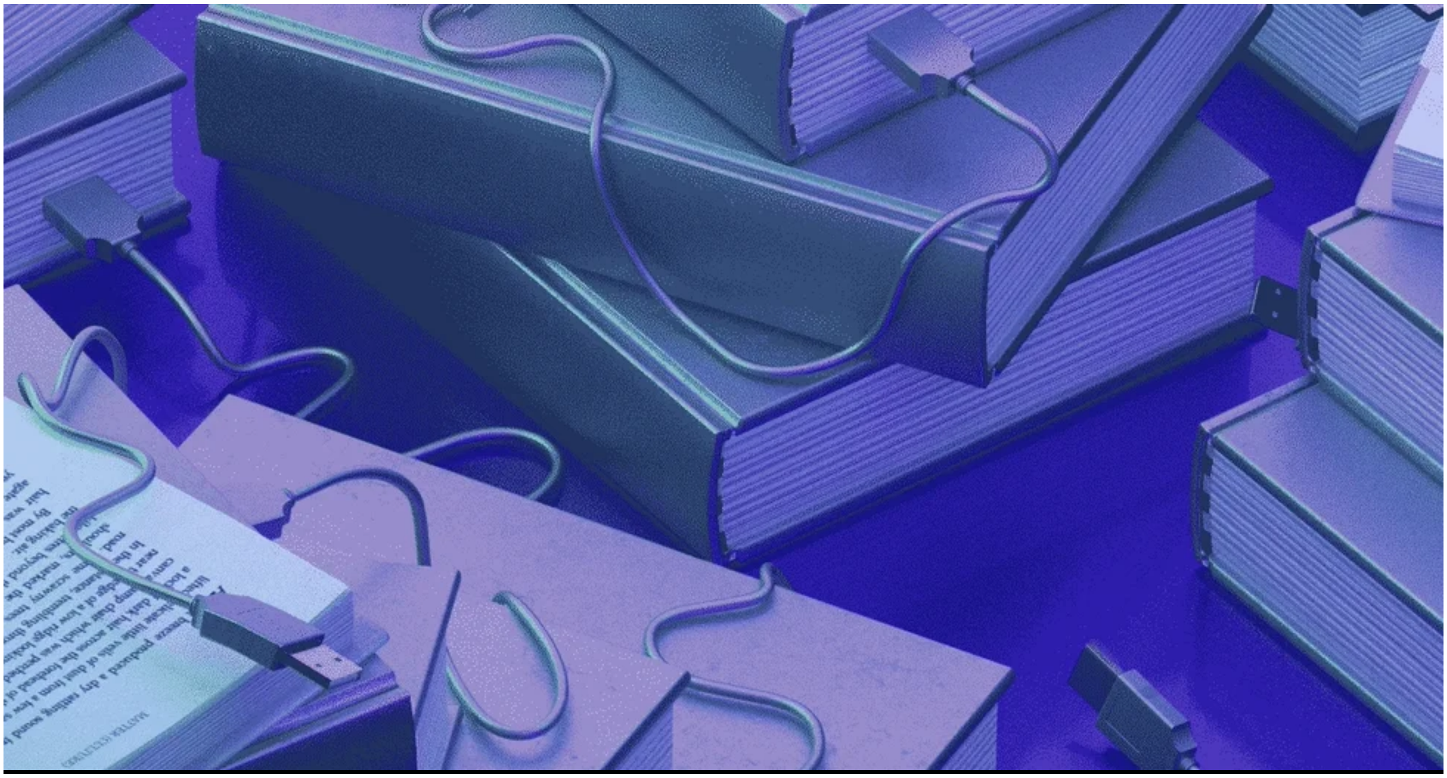
Explore This Series

TECHNOLOGY

The Unbelievable Scale of AI’s Pirated-Books Problem

Meta pirated millions of books to train its AI. Search through them here.

By Alex Reisner



Listen

- 1.0x +

0:00

15 15

11:01

Produced by ElevenLabs and News Over Audio (Noa) using AI narration. Listen to more stories on the Noa app.

Updated at 5:40 p.m. ET on March 21, 2025

Editor’s note: This analysis is part of The Atlantic*’s investigation into the Library Genesis data set. You can access the search tool directly here. Find* The Atlantic*’s search tool for movie and television writing used to train AI here.*

When employees at Meta started developing their flagship AI model, Llama 3, they faced a simple ethical question. The program would need to be trained on a huge amount of high-quality writing to be competitive with products such as ChatGPT, and acquiring all of that text legally could take time. Should they just pirate it instead?

Meta employees spoke with multiple companies about licensing books and research papers, but they weren’t thrilled with their options. This “seems unreasonably expensive,” wrote one research scientist on an internal company chat, in reference to one potential deal, according to court records. A Llama-team senior manager added that this would also be an “incredibly slow” process: “They take like 4+ weeks to deliver data.” In a message found in another legal filing, a director of engineering noted another downside to this approach: “The problem is that people don’t realize that if we license one single book, we won’t be able to lean into fair use strategy,” a

Court documents released last night show that the senior manager felt it was "really important for [Meta] to get books ASAP," as "books are actually more important than web data." Meta employees turned their attention to Library Genesis, or LibGen, one of the largest of the pirated libraries that circulate online. It currently contains more than 7.5 million books and 81 million research papers. Eventually, the team at Meta got permission from "MZ"—an apparent reference to Meta CEO Mark Zuckerberg—to download and use the data set.

This act, along with other information outlined and quoted here, recently became a matter of public record when some of Meta's internal communications were unsealed as part of a copyright-infringement lawsuit brought against the company by Sarah Silverman, Junot Díaz, and other authors of books in LibGen. Also revealed recently, in another lawsuit brought by a similar group of authors, is that OpenAI has used LibGen in the past. (A spokesperson for Meta declined to comment, citing the ongoing litigation against the company. In a response sent after this story was published, a spokesperson for OpenAI said, "The models powering ChatGPT and our API today were not developed using these datasets. These datasets, created by former employees who are no longer with OpenAI, were last used in 2021.")

Until now, most people have had no window into the contents of this library, even though they have likely been exposed to generative-AI products that use it; according to Zuckerberg, the "Meta AI" assistant has been used by hundreds of millions of people (it's embedded in Meta products such as Facebook, WhatsApp, and Instagram). To show the kind of work that has been used by Meta and OpenAI, I accessed a snapshot of LibGen's metadata—revealing the contents of the library without downloading or distributing the books or research papers themselves—and used it to create an interactive database that you can search here:

There are some important caveats to keep in mind. Knowing exactly which parts of LibGen that Meta and OpenAI used to train their models, and which parts they might have decided to exclude, is impossible. Also, the database is constantly growing. My snapshot of LibGen was taken in January 2025, more than a year after it was accessed by Meta, according to the lawsuit, so some titles here wouldn't have been available to download at that point.

LibGen's metadata are quite disorganized. There are errors throughout. Although I have cleaned up the data in various ways, LibGen is too large and error-strewn to easily fix everything. Nevertheless, the database offers a sense of the sheer scale of pirated material available to models trained on LibGen. *Cujo*, *The Gulag Archipelago*, M generative-AI models on copyrighted work without a license, because LLMs "transform" the original material into new work. The defense raises thorny questions and is likely a long way from resolution. But the use of LibGen raises another issue. Bulk downloading is often done with BitTorrent, the file-sharing protocol popular with pirates for its anonymity, and downloading with BitTorrent typically involves uploading to other users simultaneously. Internal communications show employees saying that Meta did indeed torrent LibGen, which means that Meta could have not only accessed pirated material but also distributed it to others—well established as illegal under copyright law, regardless of what the courts determine about the use of copyrighted material to train generative AI. (Meta has claimed that it "took precautions not to 'seed' any downloaded files" and that there are "no facts to show" that it distributed the books to others.) OpenAI's download method is not yet known.

Meta employees acknowledged in their internal communications that training Llama on LibGen presented a "medium-high legal risk," and discussed a variety of "mitigations" to mask their activity. One employee recommended that developers "remove data clearly marked as pirated/ stolen" and "do not externally cite the use of any training data including LibGen." Another discussed removing any line containing *ISBN*, *Copyright*, ©, *All rights reserved*. A Llama-team senior manager suggested fine-tuning Llama to "refuse to answer queries like: 'reproduce the first

RECOMMENDED READING

The Extreme Discomfort of Sharing Salary Information

JOE PINSKER

There Are Two Kinds of Happy People

ARTHUR C. BROOKS

The Universe as We Understand It May Be Impossible

It is easy to see why LibGen appeals to generative-AI companies, whose products require huge quantities of text. LibGen is enormous, many times larger than Books3, another pirated book collection whose contents I revealed in 2023. Other works in LibGen include recent literature and nonfiction by prominent authors such as Sally Rooney, Percival Everett, Hua Hsu, Jonathan Haidt, and Rachel Khong, and articles from top academic journals such as *Nature*, *Science*, and *The Lancet*. It includes many millions of articles from top academic-journal publishers such as Elsevier and Sage Publications.

Read: These 183,000 books are fueling the biggest fight in publishing and tech

LibGen was created around 2008 by scientists in Russia. As one LibGen administrator has written, the collection exists to serve people in "Africa, India, Pakistan, Iran, Iraq, China, Russia and post-USSR etc., and on a separate note, people who do not belong to academia." Over the years, the collection has ballooned as contributors piled in more and more pirated work. Initially, most of LibGen was in Russian, but English-language work quickly came to dominate the collection. LibGen has grown so quickly and avoided being shut down by authorities thanks in part to its method of dissemination. Whereas some other libraries are hosted in a single location and require a password to access, LibGen is shared in different versions by different people via peer-to-peer networks.

Many in the academic world have argued that publishers have brought this type of piracy on themselves, by making it unnecessarily difficult and expensive to access research. Sci-Hub, a sibling of LibGen, was launched independently in 2011 by a Kazakhstani neuroscience student named Alexandra Elbakyan, whose university didn't provide access to the big academic databases. In that same year, the hacktivist Aaron

Elbakyan personally. The court granted an injunction, directed the sites to shut down, and ordered Sci-Hub to pay Elsevier $15 million in damages. Yet the sites remained up, and the fines went unpaid. A similar story played out in 2023, when a group of educational and professional publishers, including Macmillan Learning and McGraw Hill, sued LibGen. This time the court ordered LibGen to pay $30 million in damages, in what TorrentFreak called "one of the broadest anti-piracy injunctions we've seen from a U.S. court." But that fine also went unpaid, and so far authorities have been largely unable to constrain the spread of these libraries online. Seventeen years after its creation, LibGen continues to grow.

Read: There's no longer any doubt that Hollywood writing is powering AI

All of this certainly makes knowledge and literature more accessible, but it relies entirely on the people who create that knowledge and literature in the first place—that labor that takes time, expertise, and often money. Worse, generative-AI chatbots are presented as oracles that have "learned" from their training data and often don't cite sources (or cite imaginary sources). This decontextualizes knowledge, prevents humans from collaborating, and makes it harder for writers and researchers to build a reputation and engage in healthy intellectual debate. Generative-AI companies say that their chatbots will *themselves* make scientific advancements, but those claims are purely hypothetical.

One of the biggest questions of the digital age is how to manage the flow of knowledge and creative work in a way that benefits society the most. LibGen and other such pirated libraries make information more accessible, allowing people to read original work without paying for it. Yet generative-AI companies such as Meta have gone a step further: Their goal is to absorb the work into profitable technology

This article has been updated to include a comment from OpenAI.

ABOUT THE AUTHOR

Alex Reisner

Alex Reisner is a contributing writer at *The Atlantic.*

Explore More Topics

Mark Zuckerberg, Meta Platforms, OpenAI, Russia

MOST POPULAR

1 Americans Are Buying an Escape Plan

ATOSSA ARAXIA ABRAHAMIAN

2 One Word Describes Trump

JONATHAN RAUCH

3 The Careless People Won

CHARLIE WARZEL

4 What the Press Got Wrong About Hitler

TIMOTHY W. RYBACK

5 What the JFK File Dump Actually Revealed

KAITLYN TIFFANY

6 Elon Musk’s Anti-Semitic, Apartheid-Loving Grandfather

JOSHUA BENTON

8 Search LibGen, the Pirated-Books Database That Meta Used to Train AI

ALEX REISNER

9 The Unbelievable Scale of AI's Pirated-Books Problem

ALEX REISNER

10 Why the Past 10 Years of American Life Have Been Uniquely Stupid

JONATHAN HAIDT

MORE FROM ALEX REISNER

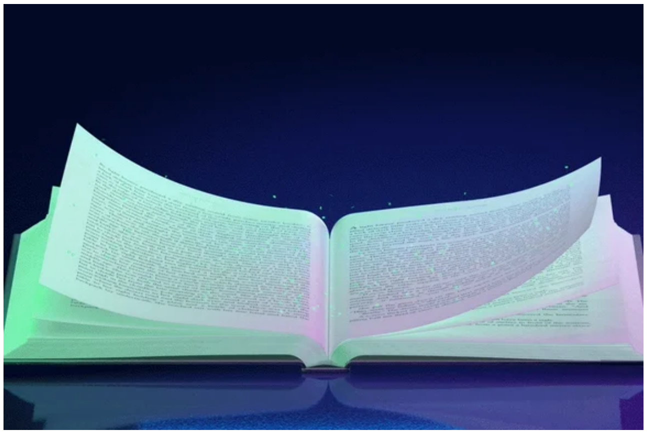


Illustration by Matteo Giuseppe Pani / The Atla…

Search LibGen, the Pirated-Books Database That Meta Used to Train

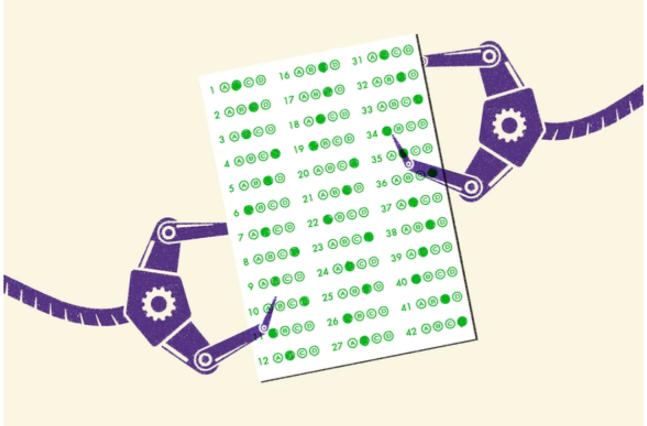


Illustration by The Atlantic. Source: Getty.

Chatbots Are Cheating on Their Benchmark Tests

Illustration by The Atlantic. Source: Getty.

Search the Hollywood AI Database

Use this search tool to see how writing from 139,000 movies and TV shows has trained generative AI.

MORE FROM ARTIFICIAL INTELLIGENCE

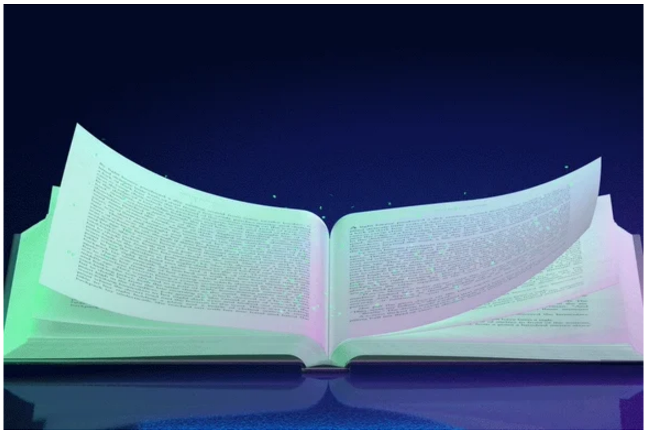


Illustration by Matteo Giuseppe Pani / The Atla…

Search LibGen, the Pirated-Books Database That Meta Used to Train AI

Millions of books and scientific papers are captured in the collection’s current iteration.

ALEX REISNER

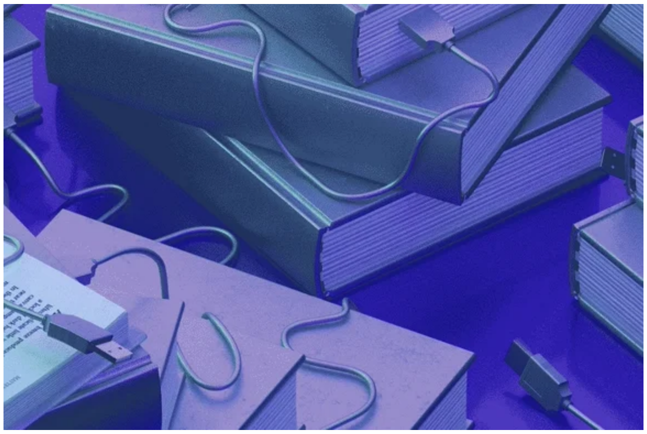


Illustration by Matteo Giuseppe Pani / The Atla…

The Unbelievable Scale of AI’s Pirated-Books Problem

Meta pirated millions of books to train its AI. Search through them here.

ALEX REISNER

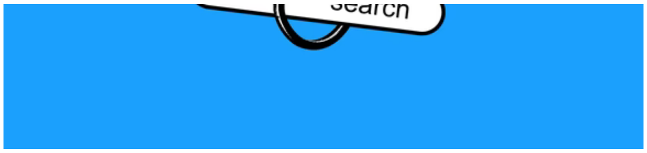


Illustration by Colin Hunter / The Atlantic. So…

Was Sam Altman Right About the Job Market?

Tech companies are unleashing AI products that do much more than answer questions.

MATTEO WONG

Illustration by The Atlantic. Sources: pressure…

DOGE's Plans to Replace Humans With AI Are Already Under Way

The civil service is being turned over to machines.

MATTEO WONG

The Atlantic Daily

Get our guide to the day's biggest news and ideas, delivered to your inbox every weekday and Sunday mornings. See more newsletters

Enter your email

Sign Up

Your newsletter subscriptions are subject to The Atlantic's Privacy Policy and Terms and Conditions.

Illustration by The Atlantic. Source: Getty.

Chatbots Are Cheating on Their Benchmark Tests

ain on questions they're later ow do we know if they're

Illustration by The Atlantic. Sources: Sebastia…

How Sam Altman Could Break Up Elon Musk and Donald Trump

Two of the most powerful tech executives in the world are desperate for the president's approval.

MATTEO WONG

Ideas That Matter

Subscribe and support more than 160 years of independent journalism.

Subscribe

THIS IS YOUR LAST FREE ARTICLE. | SIGN IN

Subscribe Now

Contact

Podcasts

Subscription

Follow

Privacy Policy Your Privacy Choices Advertising Guidelines Terms & Conditions Terms of Sale Responsible Disclosure Site Map