Inside one company's secret plan to 'destructively scan every book in the world'

Court filings reveal how AI companies raced to obtain more books to feed chatbots, including by buying, scanning and disposing of millions of titles.

23 minutes ago

(Illustration by Alexis Arnold/The Washington Post; iStock)

11 min Summary

By Aaron Schaffer, Will Oremus and Nitasha Tiku

In early 2024, executives at artificial intelligence start-up Anthropic ramped up an ambitious project they sought to keep quiet. "Project Panama is our effort to destructively scan all the books in the world," an internal planning document unsealed in legal filings last week said. "We don't want it to be known that we are working on this."

Within about a year, according to the filings, the company had spent tens of millions of dollars to acquire and slice the spines off millions of books, before scanning their pages to feed more knowledge into the AI models behind products such as its popular chatbot Claude.

Details of Project Panama, which have not been previously reported, emerged in more than 4,000 pages of documents in a copyright lawsuit brought by book authors against Anthropic, which has been valued by investors at $183 billion. The company paid $1.5 billion to settle the case in August but a district judge's decision last week to unseal a slew of

documents in the case more fully revealed Anthropic’s zealous pursuit of books.

The new documents, along with earlier filings in other copyright cases against AI companies, show the lengths to which tech firms such as Anthropic, Meta, Google and OpenAI went to obtain colossal troves of data with which to “train” their software.

The Anthropic case was part of a wave of lawsuits brought against AI companies by authors, artists, photographers and news outlets. Filings in the cases show top tech firms in a frantic, sometimes clandestine race to acquire the collected works of humanity.

Books were viewed by the companies as a crucial prize, the court records show. In a January 2023 document, one Anthropic co-founder theorized that training AI models on books could teach them “how to write well” instead of mimicking “low quality internet speak.” A 2024 email inside Meta described accessing a digital trove of books as “essential” to being competitive with its AI rivals.

But court records suggest the companies didn’t see it as practical to gain direct permission from publishers and authors to use their work. Instead, Anthropic, Meta and other companies found ways to acquire books in bulk without the authors’ knowledge, court filings allege, including by downloading pirated copies.

On several occasions, Meta employees raised concerns in internal messages that downloading a collection of millions of books without permission would violate copyright law. In December 2023, an internal email said the practice had been approved after “escalation to MZ,” an apparent reference to CEO Mark Zuckerberg, according to filings in a copyright lawsuit brought by book authors against the company. Meta declined to comment for this story.

In one newly released legal filing, Anthropic disclosed that co-founder Ben Mann personally downloaded a haul of fiction and nonfiction from a “shadow library” of books and other copyright-infringing content called LibGen over an 11-day stretch in June 2021. A screenshot of his web browser included in the filings showed him downloading files with file-sharing software.

A year later, Mann hailed the July 2022 debut of a new website called the Pirate Library Mirror, which claimed to have a massive database of books and had stated that “we deliberately violate the copyright law in

most countries.” Mann sent a link to the site to other Anthropic employees with the message, “just in time!!!”

Anthropic said in legal filings that the company never trained a commercial AI model that generated revenue using its LibGen data and never used the Pirate Library Mirror to train any complete AI model.

Ed Newton-Rex, a former AI executive and music composer who now runs a nonprofit asserting creators’ rights, said the disclosures underscore that AI companies owe creators a greater debt than they’ve paid so far. “We urgently need a reset across the AI industry, such that creatives start being paid fairly for the vital contributions they make,” he said.

Google, Microsoft and ChatGPT-maker OpenAI are also facing copyright lawsuits from book authors making similar allegations. (The Washington Post has a content partnership with OpenAI.)

Most of the legal cases against AI companies are still ongoing and James Grimmelmann, professor of digital and information law at Cornell Tech, said the questions they raise are still unsettled law. But in two early rulings, judges have found that tech companies’ use of books to train AI models without an author’s or publisher’s permission can be legal under a doctrine in copyright law known as “fair use.”

In June, District Judge William Alsup found that Anthropic was within its rights to use books for training AI models because they process the material in a “transformative” way. He likened the AI training process to teachers “training schoolchildren to write well.” The same month, District Judge Vince Chhabria found in the Meta case that the book authors had failed to show that the company’s AI models could harm sales of their books.

But companies can still get in trouble for how they went about acquiring books. In Anthropic’s case, the book-scanning project passed muster, but the judge found the company might have infringed on authors’ copyright when it downloaded millions of pirated books free before launching Project Panama.

Alsup granted class-action status to authors whose books were included in a pair of shadow libraries — huge troves of digitized books shared online without authorization — that Anthropic had downloaded and stored for future use. Rather than face a trial, the company agreed to pay $1.5 billion to publishers and authors without admitting wrongdoing. Authors whose books were downloaded can claim their share of the settlement, estimated to be about $3,000 per title.

“This case has been settled, but the court’s landmark June 2025 ruling remains intact,” Anthropic’s deputy general counsel, Aparna Sridhar, said in an email to The Washington Post. “Judge Alsup held that AI training was ‘quintessentially transformative’: Anthropic’s AI models trained on works not to ‘replicate or supplant them — but to turn a hard corner and create something different.’ The issue we settled on was about how some materials were acquired, not whether we could use them to develop” AI models.

An image disclosed in court filings shows a book warehouse that was alleged to play a role in Anthropic's Project Panama, its project to scan, digitize and destroy millions of books. (Obtained by The Post)

Buy, cut, scan, recycle

When Anthropic embarked on its Project Panama operation to buy and scan physical books, it turned to a Silicon Valley veteran. The company hired Tom Turvey, a Google executive who had helped to create the search giant’s famous but legally contested Google Books project two decades earlier.

Anthropic initially considered buying books from libraries or used bookstores like New York City landmark the Strand, known for its “18 miles” of used books, according to the filings. The store was “interested in providing used books,” according to a document detailing a March 2024 Anthropic content acquisition meeting.

Anthropic employees also discussed approaching U.S. libraries including the New York Public Library or “a new library that is

chronically underfunded,” according to the documents.

It’s not clear which, if any, of the proposals Anthropic implemented. Reached via email, a spokesperson for the Strand said the book shop did not end up selling any books to Anthropic. NYPL did not respond to a request for comment.

Anthropic eventually bought millions of books, often in batches of tens of thousands, according to the filings. It relied on booksellers including used book retailers Better World Books and U.K.-based World of Books.

The ultimate number of books scanned and their cost are redacted in the documents, but a project proposal by a vendor that ultimately worked with Anthropic noted that the AI company was “seeking an experienced document scanning services vendor to convert from 500,000 to two million books over a six-month period.”

Better World Books and World of Books did not respond to requests for comment on Monday.

The document describes how the scanning company’s “hydraulic powered cutting machine” would “neatly cut” books, whose pages would later be “scanned on high speed, high quality, production level scanners.” Finally, it notes, the scanning company will “schedule with the recycling company to pick up the completed books.”

‘Doesn’t feel right’

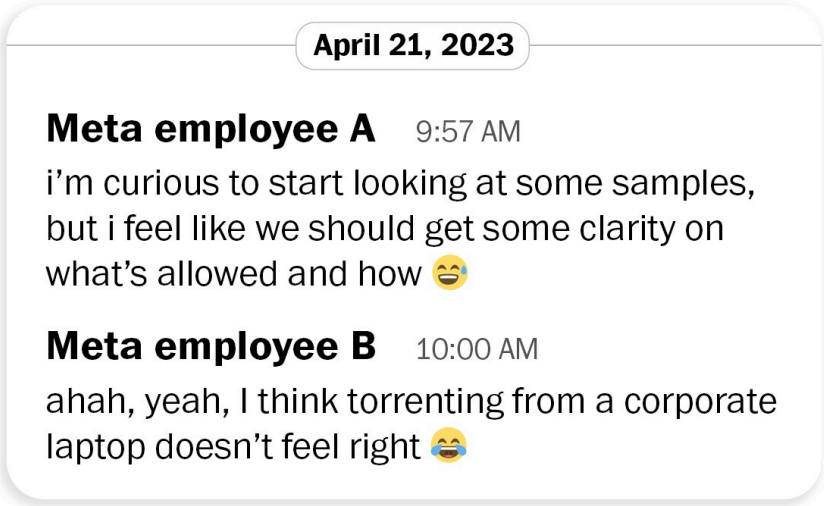
Documents released in the copyright suit against Meta show employees at the social network giant were also hungry for more data and were willing to take legal risks to obtain it. While Chhabria, the judge, sided with Meta on its use of books to train AI models, he allowed the authors to proceed with allegations that Meta illegally distributed copies of pirated books. The plaintiffs are seeking class-action status for those claims in the Northern District of California.

In their lawsuit, the authors alleged that Meta higher-ups considered paying for books to train their AI models but opted to instead download millions of books free from “torrent” platforms that facilitate online piracy. The way platforms are designed often rewards users who upload material with faster downloads of large collections of files.

Internal documents, some of which have been previously reported, showed Meta employees expressing concern that what they were doing was risky or wrong — and discussing how to cover their tracks.

“Torrenting from a corporate laptop doesn’t feel right,” one engineer wrote in 2023, according to the documents. The same employee later

shared a concern with the company’s legal team that using torrent sites could entail sharing pirated works with others, which “could be legally not OK.”



An excerpt from a conversation between two Meta employees that was disclosed in court filings. (Washington Post illustration; Court filings obtained by The Post)

The December 2023 email from the court filings makes clear that use of LibGen had been approved, apparently by Zuckerberg, referred to by his initials. “After a prior escalation to MZ, GenAI has been approved to use LibGen for Llama 3 … with a number of agreed upon mitigations,” it said, before listing legal and policy risks to using the data.

“If there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues,” the email went on.

By April 2024, internal communications showed the company was moving to download LibGen and other shadow libraries. Chat logs show one employee asked another to clarify why they were using servers rented from Amazon for torrenting rather than those owned by Facebook. The reply: “Avoiding risk of tracing back” the activity to the company.

In a filing last month, Meta’s attorneys wrote that the company “denies that it distributed Plaintiffs’ works when it downloaded training data … using torrents.”

In a separate lawsuit originally filed in 2023, book authors have accused OpenAI and Microsoft of also breaching copyright law in their own pursuit of books for AI training. OpenAI, where Mann and Anthropic CEO Dario Amodei worked before co-founding the start-up, has acknowledged downloading LibGen but told the court it deleted the files before the release of ChatGPT.

“OpenAI fired the starting gun that led to the rampant piracy by AI companies and the strip-mining of all of humanity’s expression,” said Justin A. Nelson, an attorney at Susman Godfrey LLP who is representing book authors in both the OpenAI and Anthropic cases. OpenAI declined to comment for this story.

Earlier this month, two major publishers asked a court to let them join a group of writers and illustrators in a copyright suit against Google that was originally filed in 2023.

Grimmelmann, the Cornell Tech law professor, said that AI companies “talked themselves into a fallacy” on the use of copyrighted data. The breakthroughs behind ChatGPT and similar tools began in academic research, where using copyrighted material for training is broadly accepted, he said, but researchers continued the practice even as AI models were commercialized.

“By the time the tension became clear, they had made huge investments in incorporating copyrighted data into their pipelines, and were locked in a fast-paced high-stakes competition to release newer and better models,” Grimmelmann said.

Anthropic’s decision to begin acquiring and scanning physical books instead of downloading shadow libraries “turned out to be a smart call,” he added. “This would be a good example of the company taking a more restrained approach and achieving legal compliance.”

By Aaron Schaffer
Aaron Schaffer is a researcher on The Post's Politics & Government desk. @aaronjschaffer

By Will Oremus
Will Oremus writes about the ideas, products and power struggles shaping the digital world for The Washington Post. Before joining The Post in 2021, he spent eight years as Slate's senior technology writer and two years as a senior writer for OneZero at Medium. @willoremus

By Nitasha Tiku
Nitasha Tiku is The Washington Post's tech culture reporter based in San Francisco.
@nitashatiku