

Worked for 20s >

Think of it as a **private corporate library**, not a public ebook service.

The scanned file sits in Anthropic's internal "research library" or "generalized data area." Anthropic can allow authorized personnel and internal computer systems to search it, inspect it, select it for a training dataset, or use it for other internal research. The court record says the files were not shown, sold, or shared outside the company, but it does not disclose the precise access controls or which employees can open a complete book.

 Copyright Alliance +1

A normal Claude user does **not** receive access to that digital book. You cannot ordinarily ask Claude to open the scan, display its pages, or lend you the ebook. During training, software processes the text so the model can learn patterns from it. The trained model is separate from the source library: it does not ordinarily function as a bookshelf containing retrievable ebook files, although models can sometimes reproduce portions of training material—which would raise separate legal questions.

So, in practical terms:

Anthropic controls access to its particular scan. It decides which internal people and systems can use that file.

The author or publisher still controls the copyright. They retain the legal rights to publish, license, sell, and distribute authorized editions. Anthropic cannot simply put its scan online, sell copies, or give the full file to Claude customers merely because it owns the internal replacement copy. The court's ruling depended partly on the absence of external distribution.  Copyright Alliance +1

The public must obtain the book elsewhere—from a publisher, bookstore, library, used-book seller, or another lawful digital source. Anthropic destroyed only the copy it purchased, not the copyright or other existing copies.

The unusual result is that Anthropic may possess a highly usable, searchable digital copy that outsiders cannot access, while the physical copy that previously could have been resold or donated no longer exists. That is why saying Anthropic does not “own the copyright” is legally important but does not eliminate the practical issue you are pointing to: Anthropic exclusively controls access to **its** scan.

    ...  Sources

do other AI companies follow this practice? why not scan the books like the internet archive does, and make them available to everyone?

Worked for 22s >

There is **no comparably well-documented evidence that another major AI company has run an Anthropic-scale program of buying, cutting apart, and scanning millions of physical books.** Meta and OpenAI have faced allegations concerning books obtained from online datasets or shadow libraries, rather than physical books they purchased and destroyed. Google has scanned tens of millions of library books, but generally used non-destructive methods and returned the books.  U.S. Copyri... +3

Other unidentified companies may be buying books for scanning. Recent booksellers have reported unusual bulk orders, but the available reporting has not conclusively connected most of those purchases to named AI developers.  The Atlantic +1

Why not follow the Internet Archive model? There are three main reasons.

First, making a copyrighted scan publicly readable is legally much riskier than using it internally. Copyright includes separate rights to reproduce, distribute, and publicly display a work. Even assuming that internal AI training is fair use, that does not give the company permission to offer the complete ebook to the public.

The Internet Archive tried a limited version called controlled digital lending: it would lend only one digital copy for each corresponding physical copy it held. The federal appeals court nevertheless ruled in 2024 that this system was not fair use for the copyrighted books at issue, partly because free digital loans substituted for publishers' licensed ebooks. The Internet Archive subsequently stopped lending many affected books and ended the litigation.  Justia Law +1

Second, Internet Archive-style scanning is slower and more expensive. Its equipment holds an intact book in a cradle while cameras photograph facing pages, minimizing stress to the binding. Destructive scanning removes the spine and permits loose pages to pass through automated document scanners. It sacrifices the physical copy for much greater throughput.  Internet Archive ... +1

Third, Anthropic regards the resulting collection as a **proprietary corporate asset**. A private, searchable corpus can improve its models and give it an advantage over competitors. Publishing that corpus would give away the dataset while also creating substantial copyright liability.

A lawful public-interest alternative would resemble **Google Books**, rather than unrestricted Internet Archive lending:

- scan books without destroying them;
- allow full access to public-domain or expressly licensed books;
- provide search results and limited snippets for unlicensed copyrighted books;
- direct readers to libraries, publishers, and booksellers for access to the full work.

Courts upheld Google Books largely because its public service was search-oriented and did not let users read substantial portions of unlicensed copyrighted books.  Justia Law +1

So Anthropic could have preserved the physical books and created a public discovery system. What it generally could not do—without permission—was place the complete copyrighted scans online for everyone to read.

    ...  Sources

so it's safe to say that Anthropic is an outlier in their choice here

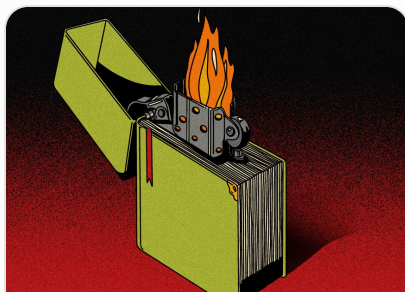
Yes—with an important qualification. Among major AI companies whose practices are publicly documented, Anthropic appears to be an outlier in the **scale and deliberate design** of its program: purchasing millions of physical books, cutting off their bindings, scanning them, recycling the originals, and retaining a private digital library.

It is not safe to say Anthropic is necessarily the **only** company doing this. Recent bulk purchases suggest other AI developers or data contractors may be acquiring physical books, but the buyers and intended uses are often concealed, and there is not yet equivalent evidence linking another named major AI laboratory to an Anthropic-scale destructive-scanning program.  The Atlantic +1

Anthropic is therefore:

- an outlier in **publicly proven physical destruction at industrial scale**;
- not an outlier in using copyrighted books as private AI-training material;
- possibly not unique, given the secrecy of data-acquisition supply chains.

So the most defensible phrasing is: **“Anthropic is the clearest and largest publicly documented example, rather than conclusively the only company engaged in the practice.”**



 The Atlantic

Someone Is Mysteriously Snapping Up Used Books Around the World

Today



 The Guardian

'More than just objects': Australian booksellers raise alarm over 'horrific' destruction of rare titles to feed AI

4 days ago

      Sources