

The books in the libraries they took

A plain-language guide to the pirate-library catalogs in the AI copyright cases: what the court record says each company downloaded, how a book is matched to its catalog entry and its U.S. copyright registration, and why the registration date matters.

Figures from dataset `payauthors-2026-10-01-m64` (built Oct. 5, 2026). Court record reviewed through Oct. 5, 2026.

Read online, with live charts and links: payauthors.com/white-paper

In brief

- Court orders and the companies' own filings show that several AI developers downloaded large pirate book collections: Library Genesis (LibGen), Z-Library (through the Pirate Library Mirror and Anna's Archive) and Books3. Other downloads are alleged and not yet decided.
- Those collections published catalogs of what they held, and dated copies of those catalogs survive. PayAuthors holds 31 of them, from November 2015 to October 2026, and the U.S. Copyright Office's registrations for literary works from 1978 through 2025.
- Matching a book to those catalogs and to its registration, then comparing the registration date with the download date, answers a narrow question the law cares about: was the work in the collection, and was the paperwork on it before the download?
- Counting each book once, the catalogs tied to Meta list 616,230 registered works that pass that date test; OpenAI 422,952; NVIDIA 266,677. For Anthropic, 9,584 such works are outside its \$1.5 billion settlement by our screen. These are our counts from public records, not court findings.
- A match is a lead, not proof. It does not show that a company downloaded a particular file, trained on it, or infringed anyone's copyright.

22.4M

catalog records,
across overlapping
sources (records, not
unique books)

31

dated catalogs:
LibGen 25, Z-Library
4, Books3 1, Internet
Archive 1

8.03M

U.S. copyright
registrations for
literary works, 1978–
2025

1.52M

registrations matched
to at least one catalog
record (candidates)

1. Why a catalog matters now

For three years, the lawsuits over AI training asked a broad question: did these companies copy pirated books? For several of them, the court record now answers part of it. A federal judge found that Anthropic downloaded millions of books from LibGen and the Pirate Library Mirror. Meta has admitted in its own court filings that it downloaded LibGen and torrented data from Anna's Archive. A discovery order in the OpenAI litigation describes as undisputed that an OpenAI employee downloaded pirated books from LibGen in 2018.

That moves the fight to narrower questions, book by book. Which works were in the collection a company took? Who owns them? Were they registered with the Copyright Office, and when? Anthropic's settlement shows how much those details matter. It covered 482,460 works on a final Works List, at about \$3,000 per work before fees and costs. Books with no ISBN, foreign editions, books registered too late and books whose catalog entry and registration did not line up were left off, even when they were in the downloaded collections. Authors who opted out, and authors whose books were never on the list, are now suing on their own.

The answer to "was my book in that collection, and on what date?" sits in the collections' own catalogs. This paper explains what those catalogs are, how PayAuthors reads them, and what a match can and cannot tell you.

2. The collections, in plain words

A **shadow library** is a website that offers copies of books and articles without the rights holders' permission. The large ones publish a **catalog**: a database listing every file they offer, with its title, author, publisher, year, ISBN, file type and a digital fingerprint of the file (its MD5 checksum). Some of them post the whole catalog as a downloadable file, a **dump**, and archivists have saved copies over the years. A dated dump is a record of what the library listed on that day.

Library Genesis (LibGen)

One of the oldest of the large collections, with separate non-fiction and fiction catalogs. It hands out its files in BitTorrent bundles of 1,000 files each. The Internet Archive preserved LibGen's catalog dumps over many years; PayAuthors holds 25 of them, from Nov. 14, 2015 to June 3, 2025. Between May 2017 and February 2021 only two survive (April 2019 and February 2020), and for fiction only the February 2020 one.

Z-Library, the Pirate Library Mirror and Anna's Archive

Z-Library is a large shadow library that overlaps with LibGen. In 2022 a group calling itself the **Pirate Library Mirror (PiLiMi)** copied Z-Library and released it in two parts: July 2022 (2.18 million records) and September 2022 (3.82 million). **Anna's Archive** is a search engine and mirror that combines LibGen, Z-Library and other collections; its index records the date each later Z-Library book was added, from August 2022 to August 2023. PayAuthors holds four Z-Library catalogs.

Books3

A collection of about 196,640 books, released in October 2020 as part of a training data set called The Pile. The texts are reported to have come from a private torrent site, Bibliotik. PayAuthors holds the published list of 196,608 Books3 file names and a list of 129,635 ISBNs extracted from the texts, not the texts themselves.

The Internet Archive's lending library

The Internet Archive is a nonprofit library, not a pirate site. Its lending library of scanned print books is included because pending AI cases allege that AI developers obtained its lending books; no court has found that. In 2024 an appeals court held that the Archive's scanning and lending of the 127 books in that case was not fair use. A listing there is not a finding that anyone did anything wrong.

PayAuthors works from catalog metadata only. It does not download, hold or open any book file.

3. What the court record says, company by company

Coverage of these cases often blurs what a judge decided with what a plaintiff claims. We label every statement by its source, and we ask reporters who quote us to keep the label.

COURT ORDER

What a judge found or ordered, or a fact an order calls undisputed.

ALLEGATION

What a complaint or brief claims. Not decided.

GOVERNMENT FILING

A statement to the court by a government agency that is not a party.

COMPANY'S OWN STATEMENT

A defendant's filing, answer or paper, about itself.

OUR ANALYSIS

Our counts and matches from public catalogs and Copyright Office records. Not a court finding.

Anthropic (2021–2022)

- **COURT ORDER** Downloaded Books3 (early 2021), more than five million books from LibGen (June 2021) and more than two million from the Pirate Library Mirror (July 2022).

Bartz v. Anthropic, N.D. Cal., order of June 23, 2025

- **COURT ORDER** \$1.5 billion settlement for the 482,460 works on its Works List, approved July 20, 2026. 350 owners opted out, covering 1,802 works. First payments are expected by Nov. 15, 2026; two appeals are pending.

Bartz, Dkts. 680, 619-3, 692

- ◆ **COMPANY FILING (ARGUMENT)** Anthropic argues the opt-out authors' suits come too late: a three-year limit, it says, bars claims over the 2021 and 2022 downloads. The court has not ruled; a hearing is set for Dec. 17, 2026.

Cambronne v. Anthropic, Dkt. 228 (Sept. 24, 2026)

Meta (2022–2024)

- **COURT ORDER** Meta first used a shadow library in October 2022, when it downloaded LibGen, and downloaded Anna's Archive in early 2024. On the named authors' record, the court ruled that training was fair use, a ruling it said "stands only for the proposition that these plaintiffs made the wrong arguments and failed to develop a record."

Kadrey v. Meta, N.D. Cal., Dkt. 598 (June 25, 2025)

- ◆ **COMPANY'S OWN STATEMENT** Meta admits downloading LibGen data in or about October 2022, torrenting Anna's Archive data including its Z-Library and LibGen portions, using some of it for Llama, and that some data was uploaded while it torrented. It says any uploading was de minimis.

Sullivan v. Meta, Dkt. 85; *Elsevier v. Meta*, Dkt. 78

- **ALLEGATION** Publishers allege Meta torrented 134.6 terabytes between April and July 2024 and uploaded 40.42 terabytes to others. Textbook authors allege about 4,500 users obtained LibGen files from Meta in 2022.

Elsevier complaint ¶ 99; *Sullivan* complaint ¶ 51

Where it stands: several suits in San Francisco; four are set for trial on May 24, 2027.

OpenAI (2018–2020)

- **COURT ORDER** A discovery order describes as undisputed that “in 2018, an OpenAI employee downloaded pirated copies of books from Library Genesis.”

Authors Guild v. OpenAI, S.D.N.Y., Dkt. 782 (Nov. 24, 2025)

- **ALLEGATION** Two employees torrented about 35 terabytes, all of LibGen’s fiction and non-fiction, between September 2019 and January 2020; data sets called LibGen1 and LibGen2 were renamed Books1 and Books2; OpenAI deleted its LibGen data in 2022.

Class plaintiffs’ Rule 56.1 statement, MDL Dkt. 1987 (Sept. 17, 2026)

- GOVERNMENT FILING The Justice Department told the court that training models on copyrighted text is fair use; it did not address how training copies were obtained. The news plaintiffs answered on Sept. 28, 2026.

In re OpenAI, No. 1:25-md-03143, Dkts. 1682, 2071

Where it stands: consolidated in New York before Judge Sidney H. Stein; OpenAI’s summary judgment motion is pending.

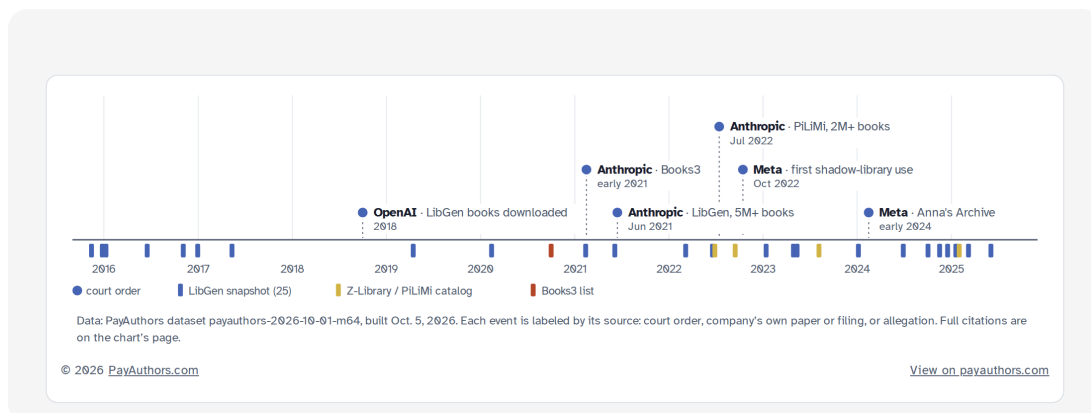
NVIDIA and others

- **COURT ORDER** NVIDIA: claims over the Pirate Library Mirror, Bibliotik and BitTorrent go forward (May 5, 2026). This is a ruling on the pleadings, not a finding of copying.

Nazemian v. NVIDIA, N.D. Cal.

- ◆ **COMPANY’S OWN STATEMENT** NVIDIA’s answer admits downloading data from Anna’s Archive’s data sets page and from libgen.rs, without dates (June 9, 2026). Bloomberg’s BloombergGPT paper lists Books3 among its training data (2023). DeepSeek’s DeepSeek-VL paper says it used e-books from Anna’s Archive (2024).

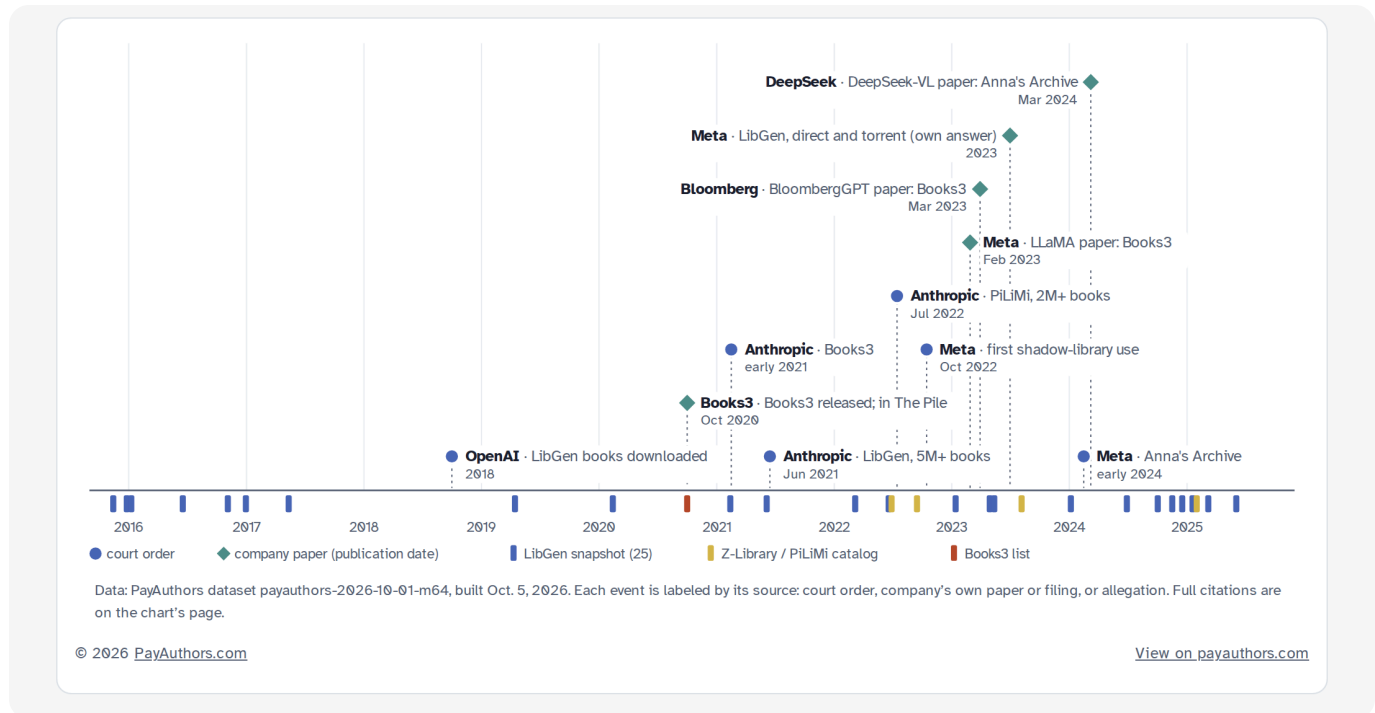
- **ALLEGATION** Google: publishers and authors allege Gemini was trained on Google Books copies and on web text that includes pirate sites (July 10, 2026). Apple, xAI and Perplexity: authors say there is a reasonable basis to believe each used Anna’s Archive.



Court orders only: the downloads a judge has found or called undisputed. Live version and citations: payauthors.com.

4. The timeline

The chart below puts each documented download on the same axis as our dated catalogs. The blue ticks along the bottom are LibGen snapshots; gold ticks are Z-Library catalogs; the red tick is the Books3 file list. A dotted line drops from each event to the time axis.



Timeline at a glance: documented downloads, 2016–2024, with our dated catalogs. The full timeline, with allegations and how precisely each date is known, is at payauthors.com/timeline.

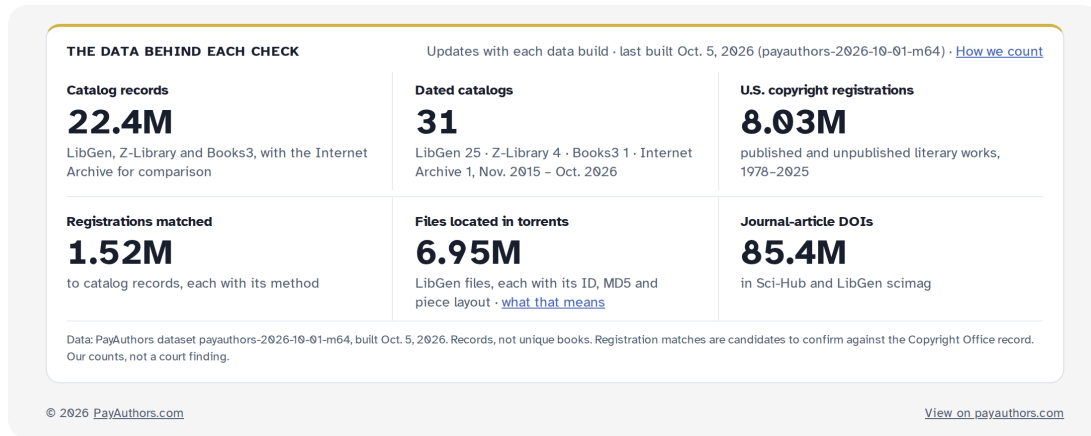
Two things make a catalog useful against this timeline. First, a snapshot dated before a download shows what the library listed by then. Our LibGen snapshot of June 5, 2021, for example, falls in the month the court found Anthropic downloaded LibGen. Second, each LibGen entry, and each Z-Library entry added after the Pirate Library Mirror, carries the library's own record of when the file was added, so a report can show that a book was listed before a given download even between snapshots.

What a snapshot cannot show is which files a company actually took. Connecting a specific book to a specific download takes the downloader's own evidence: a file manifest, file fingerprints, a log or testimony. That comes from discovery, not from a catalog. A catalog entry's MD5 fingerprint is what lets that comparison be made, byte for byte, once the company's records are produced.

5. From a name to a registration: how a match is made

- 1 Search the name.** An author's name (or an ISBN) is searched across every catalog. A record matches when every part of the name is found in one author name on the record, so "Ann" finds Ann and Anna but not Joanne. Other spellings are separate searches. When the Copyright Office's own records tie a pen name to a legal name ("pseud." or "writing as" on a registration), the free check suggests the other name, and a catalog record filed under the pen name is matched to registrations in the legal name; those matches are labeled and held for a person to confirm.
- 2 Gather the catalog lines.** Entries that describe the same edition are combined into one line. Each line lists the dated snapshots that show it, the library's file ID, the file type and the MD5 fingerprint, and, where we could locate it, which BitTorrent bundle carried the file. We located 6.95 million LibGen files in their torrents this way, from the torrents' own metadata. Where the torrent file survives, the line also shows how the file sits in that torrent's pieces: its size, the piece length, how many pieces it spans, and the largest share of the file inside one piece. A torrent is checked piece by piece, so a file that lies inside one piece is whole in every verified copy of that piece. The metadata does not show who sent which parts.
- 3 Find the registration.** Each book is matched to candidate U.S. Copyright Office registrations, strongest evidence first: the same ISBN on both records (confirmed by a name or most of the title); the same full title plus an author's or claimant's surname; the same main title plus surname; the same title words plus surname; or the same title plus the publisher on a company registration. Translations, study guides, adaptations and conflicting volume or edition numbers are set aside.
- 4 Check it.** Each candidate is tested against the Office record: the title agrees, a surname of the catalog author appears on the record, no volume or edition number conflicts, and the catalog edition is dated close to the registration. Matches that pass are marked "Record matched." The rest are reviewed one by one in a software review and stay marked "Not yet checked" until they are confirmed. When someone buys a report, a person at PayAuthors checks every match that affects its counts against the full Office record, and a person's decision overrides the software's.
- 5 Compare the dates.** For each confirmed registration, the report compares the registration date with each documented download date and says whether the paperwork was in place before the download. Only works that pass are counted.

The full Copyright Office record behind each match, including the publisher, the claimant and its address, any transfer of rights, and the rights-and-permissions contact the Office lists, is shown in the Registration Details report, with a link to the Office's own page so anyone can check it.



The data at a glance. Sources and dates for every figure: payauthors.com/methodology.

6. Why the registration date matters

U.S. copyright law sets two rules that turn on registration, and both are about paperwork, not about whether the copying happened.

Registration before suit. A U.S. work generally cannot be sued on until the Copyright Office registers it or refuses it (17 U.S.C. § 411(a)). Works first published outside the United States are exempt from that rule.

Registration before the infringement. Statutory damages and attorney’s fees are available only if the work was registered before the infringement began, or within three months of first publication (17 U.S.C. § 412). For these cases, that means: was the work registered before the download?

The takeaway for an author is simple: was the paperwork on the book before the download? If it was, the author can ask for statutory damages, which do not require proving any loss.

Statutory damages run from \$750 to \$30,000 per work, and up to \$150,000 per work if the infringement is found to be willful (§ 504(c)). They are set per work, not per copy: one award for all of one infringer’s infringements of a work in a case. An author can instead seek actual damages and the infringer’s profits (§ 504(b)), which must be proven and in rare cases can be higher. A court may also award costs and attorney’s fees (§ 505).

The same timing now shapes the cases. Anthropic’s Works List required registration within five years of publication and either before the settlement’s Download Date (Aug. 10, 2022) or within three months of publication. The textbook-author classes proposed against Meta and OpenAI are defined the same way: works registered within five years of first publication and before the defendant’s copying, or within three months of publication. Neither class has been certified.

Each company is timed from its **earliest known download** of a work. Courts treat repeated infringement of the same kind as one continuing infringement that began with the first act (*Derek Andrew, Inc. v. Poof Apparel*

Corp., 9th Cir. 2008), so a registration made between two downloads by the same company does not count for that company.

Our date screen is deliberately cautious. We test each registration against a date we can be sure is no later than the download: the later of the download’s known start and the library’s own record of when that copy was added. A registration dated inside a download period, or a comparison with no usable date, is shown but never counted.

7. What the numbers say

For each company, we built lists of the registered works in the collections it is found or alleged to have downloaded, as those collections stood at the time. Copies, formats, editions and translations of one book count once, and a book on two of a company’s lists counts once.

COMPANY	COLLECTIONS	WORKS ON ITS LISTS	ON TWO OR MORE LISTS	REGISTERED BEFORE THE DOWNLOAD
Anthropic	Books3, LibGen and PiLiMi works not on the Works List by our screen	9,700	1,635	9,584
Meta	LibGen, Anna’s Archive and Books3	618,350	145,360	616,230
NVIDIA	Anna’s Archive (admitted; date alleged) and Books3 (alleged)	267,561	29,419	266,677
OpenAI	LibGen 2018 (court order), 2019–2020 torrenting and Books3 (alleged)	422,952	332,983	422,952

■ **OUR ANALYSIS** Registrations matched to dated catalog records; each match is a candidate to confirm against the Office record. These are counts of works, not damages estimates. If a court awards statutory damages, it sets them per work, from \$750 to \$30,000, or up to \$150,000 for willful infringement, and whether any work qualifies is for a court. Figures for different companies are not a combined total. Anthropic works released in its settlement cannot be claimed again against Anthropic.

Anthropic’s lists are the works its settlement left out: 5,461 LibGen works, 1,629 PiLiMi works and 4,515 Books3 works that are registered and pass the date screen but are not on the Works List by our screen. Their authors were left outside the settlement for reasons that have nothing to do with whether the books were copied.

8. How accurate it is

We test the matcher against works lists filed in court, which pair each title with its registration number.

- **Anthropic opt-out works** (*Bartz*, Dkt. 619–3, Ex. J; 1,749 registrations): of those found in the catalogs, 94.0% are linked to the exact registration in the court record and 96.0% to some registration of the

work. About four in ten of the remaining misses are entries where the exhibit itself pairs a title with another work's registration.

- **Textbook works lists** (7,001 registrations, mainly *Cengage Learning v. Google*, S.D.N.Y.): 93.1% exact, 96.9% some registration of the work.
- **How often a counted match is right:** reports count only matches that passed every automatic check or were confirmed in a review. A person is checking a random sample of 226 counted matches against the Office record, blind to how each was made; that result will be published on our Methodology page. A first sample of 100 matches, drawn from all candidate matches including ones reports hold back and checked in a software review, found 91% were the same work.

A match is right when the registration covers the catalog book's text. A registration covers the work in any printing or format, so a later paperback or ebook with the same text is the same work; the edition matters only where the text differs. The 100 reviewed matches are published, each with its verdict and a link to the Office record, so anyone can recheck them. These checks are ours, not an independent audit.

9. What this cannot show

- **A listing is not a download.** A catalog shows what a library offered, not what any company took. A match does not show that a company downloaded a particular file, that a book was used to train a model, or that anyone infringed a copyright.
- **Gaps in the record.** Catalogs between snapshots are lost; entries deleted before a surviving snapshot do not appear at all. About 178,000 LibGen entries carry no author and cannot be found by name.
- **Matching misses some books and catches some it shouldn't.** Other spellings, pen names the Office records do not tie to a legal name, and registrations under a different form of a name are missed; common names bring up other people's books. Every match is a candidate to confirm.
- **Registration data has edges.** It covers literary-work registrations (TX and TXu) from 1978 through 2025. Audiobooks, renewals and earlier registrations are not covered. A missing match does not mean a book is unregistered.
- **Foreign works are not in the counts.** Works first published abroad can be sued on without registering, but statutory damages still need timely registration. Unregistered foreign works are left out of our counts, though their authors may still claim actual damages.
- **Deadlines.** Copyright claims generally must be brought within three years (§ 507(b)). How that applies to downloads in 2021 and 2022 is being argued now. The date screen does not address whether a claim is timely.

None of this is legal advice. Ownership, defenses such as fair use, and any deadline are questions for a lawyer.

10. Sources and how to check us

Every source is public. The LibGen dumps come from the Internet Archive's preserved copies; the Z-Library catalogs from the Pirate Library Mirror index and Anna's Archive's metadata as re-hosted for research; the

Books3 lists from a published repository; the registrations from the Copyright Office's own public data, published in January 2026. Each source file is listed on our [Methodology](#) page with its archive item, size and checksum.

- **Preservation.** We keep a copy of every original source file (about 61 GB of catalog dumps, file lists and Office data), with checksums recorded at download. Every report names the dataset build it came from, and its results are frozen at purchase.
- **Each file identified without holding it.** A catalog's MD5 fingerprint identifies one exact file. A copy produced in discovery can be checked against it byte for byte, without anyone at PayAuthors opening a book file.
- **Every decision recorded.** Each match decision, whether automatic, by the software review or by a person, is stored with its basis, so any one can be inspected.

Read further: [Methodology](#) (sources, matching rules, accuracy, checksums) · [What a match means](#) · [Timeline](#) · [Lawsuits, company by company](#) · [News, graded by source](#) · [A sample report](#).

11. For reporters

Keep the grade. An allegation reported as a finding is the most common error in coverage of these cases. Each item on our timeline and news page carries its source.

Use the date. "Registered before the download" is the test the law, the settlement and the proposed classes use. Our reports show it work by work.

A match is a lead. Say "listed in the catalog" or "in the collection," not "downloaded by" a company, unless a court record or the company says so.

Dates to watch. First Anthropic settlement payments by Nov. 15, 2026; hearing on Anthropic's motion to dismiss the opt-out suits, Dec. 17, 2026; trial in four suits against Meta, May 24, 2027.

Press codes. Reporters can get a free press code for their outlet, good for 10 Registration Details reports, from a work e-mail address on our [press page](#). Live, credited charts can be embedded from the same page.

How to cite: Source: PayAuthors.com analysis of dated LibGen, Z-Library and Books3 catalog records, the Internet Archive's lending-library catalog, and U.S. Copyright Office registrations (payauthors.com).

© 2026 PayAuthors.com. Charts, figures and tables may be shared unaltered with credit and a link: payauthors.com/terms#reuse.

PayAuthors is a research service. It uses catalog metadata only and holds no book files. It is not a law firm, does not refer clients to lawyers, and is not affiliated with or endorsed by the U.S. Copyright Office. Reports are research support, not legal advice. Inquiries: support@payauthors.com.