



Publications as Data for AI

Nicholas Taylor and Rory Elliott

[Research Library](#)

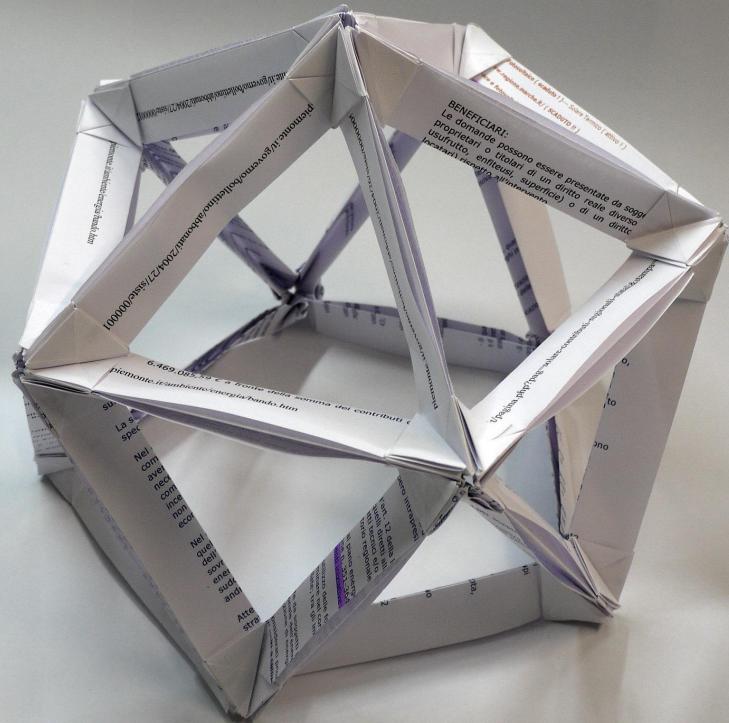
[DOE Fusion XIII](#)

17 September 2025

LA-UR-25-29019

Publications as Data Are Foundational to Powerful AI

"modular origami tv (nu) dome" by
[Simona](#) under CC BY-SA 2.0



The Problem of Publications as Data for AI

tom's HARDWARE US Edition RSS Sign in Search

Best Picks CPUs GPUs SSDs News 3D Printers More Forums

TRENDING Back to School Deals The death of Win 11 SE Where to Buy Switch 2

Tech Industry > Artificial Intelligence

Meta staff torrented nearly 82TB of pirated books for AI training — court records reveal copyright violations

By Jowi Morales published February 9, 2025

Did they think they could get away with it?

When you purchase through links on our site, we may earn an affiliate commission. [Here's how it works.](#)

Morales, Jowi. "[Meta staff torrent nearly 82TB of pirated books for AI training — court records reveal copyright violations](#)". Tom's Hardware, 9 Feb. 2025. Accessed 15 Aug. 2025.

ars TECHNICA SECTIONS FORUM SIGN IN

FOLLOW THE PAPER TRAIL

Anthropic destroyed millions of print books to build its AI models

Company hired Google's book-scanning chief to cut up and digitize "all the books in the world."

BENJ EDWARDS - JUN 25, 2025 2:00 PM 243

Credit: Alexander Sipatov via Google Images

Edwards, Benj. "[Anthropic destroyed millions of print books to build its AI models](#)". Ars Technica, 25 Jun. 2025. Accessed 15 Aug. 2025.

nature View all journals Search Log in

Explore content About the journal Publish with us

nature > news > article

NEWS | 09 December 2024

Publishers are selling papers to train AIs – and making millions of dollars

Generative-AI models require massive amounts of data – scholarly publishers are licensing their content to train them.

By Diana Kwon

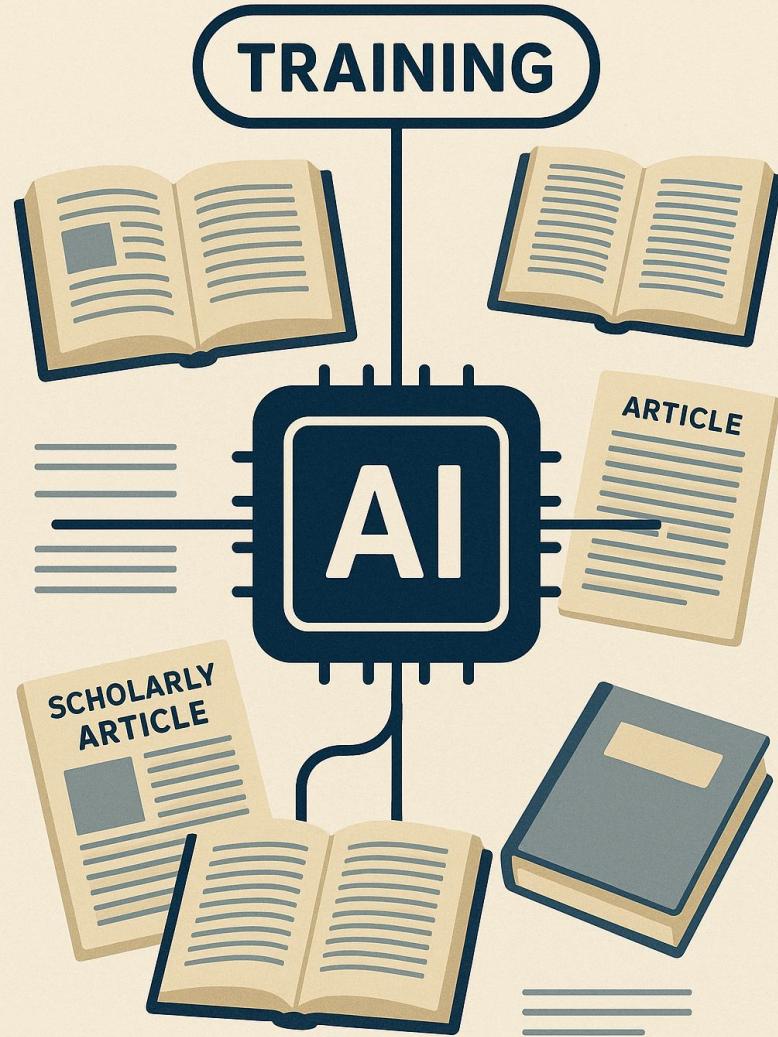
Since the explosion in popularity of generative [artificial intelligence](#) (AI), several scholarly publishers have forged agreements with technology companies looking to [use content to train the large language models \(LLMs\)](#) that underlie their AI tools. A new tracker aims to catalogue what deals are being made – and by whom.

Access options

Kwon, Diana. "[Publishers are selling papers to train AIs – and making millions of dollars](#)". Nature, 9 Dec. 2024. Accessed 15 Aug. 2025.

Why publications for AI?

- By *publications*, here I mean *scholarly* publications – journal articles, conference proceedings, books, technical reports
- Can be used to train new models, fine-tune or RAG with existing models
- Improve the quality, sophistication, and veracity of responses
- Imbue them with particular domain knowledge



Why publications for LANL AI?

- Apply domain knowledge to solve national security science challenges
- Leverage agentic capabilities to interact with scientific facilities and instruments
- Enable quicker iteration thru hypotheses, experiment design, execution
- Achieve better (or novel) results, more quickly than humans alone

["NSSB Exterior Photo"](#) by unknown
LANL photographer



Agenda

1. Overview
2. Addressing Misconceptions
3. Summarizing Challenges
4. Unencumbered Works
5. Find Unencumbered Works
6. Other Legal Approaches
7. Looking Forward

Addressing Misconceptions

"a little bit misconception
#мобилография #натюрморт" by
[sergej xarkonen](#) under [CC BY 2.0](#)



Can Read = Can Process w/ AI?

You would reasonably suppose...

- *I can download and read an article...*
- *Surely, I should be able to do what I want with it at that point...*
- *Including upload it to an LLM to summarize and query it...*
- *Or use it as part of a corpus to help train or fine-tune an AI model...*



"IMG_0147" by [derechoaleer](#)
under [CC BY-SA 2.0](#)

But, Content Licenses

- Unfortunately, this is inconsistent w/ terms of most e-resource subscriptions
- Your libraries negotiate contracts and try to secure AI permissions
- Publishers limit AI allowances; want to separately monetize
- But content-for-AI licensing service models are still immature



["sailor pulling rope"](#) by [Mikel Ortega](#)
under [CC BY-SA 2.0](#)

Courts Affirm Fair Use

- Courts have recognized transformative use of copyrighted works as fair use
- On that basis, District Court Judge ruled AI training on copyrighted books is fair
- Didn't explicitly extend to copyrighted articles, but reasonable inference?
- Surely, should allow for use of arbitrary copyrighted content for arbitrary AI use?

["Soup Columns"](#) by [AppleDave](#)
under [CC BY-NC 2.0](#)



But (Again), Content Licenses

- Content licenses can (and, often, do) also terminate fair use rights
- That is, activities otherwise permissible via fair use may not be for this content
- And automated retrieval of subscription content endangers institutional access
- Publishers have mechanisms to detect, and condition access on compliance



Why not negotiate better?

- Sounds like there's a logical solution:
negotiate better content licenses
- Challenged by monopsony and
non-fungibility of content
- Necessity for *specific content* limits
institutional leverage

["handshake pls ^"](#) by [Kaivr](#)
under [CC BY-NC-SA 2.0](#)



Summarizing Challenges

["installing cement road barriers"](#) by
[Jnzl's Photos](#) under [CC BY 2.0](#)



Legal Risk

- Copyright law unclear on permissibility of AI use cases for copyrighted materials
- Cases pending, few decided, and only at Federal District Court level
- Unauthorized use of copyrighted content for AI creates potential financial liability
- Undermines institutional financial outlook and credibility

["Close to the edge" by Chris](#)
under CC BY-NC-ND 2.0



Risks to E-Resource Access

- Our work relies upon e-resources for many mission-critical purposes beyond AI
- Noticed use of subscription publications for AI endangers continued access
- Systematic downloading of subscription publications endangers continued access
- Access restriction may be temporary or permanent and affect entire organization

[“Disconnected”](#) by [Sharon Drummond](#)
under [CC BY-NC-SA 2.0](#)



Immature Licensing Paths

- Publishers don't want to freely concede monetizable permission grants, but relative market demand isn't yet clear
- Consequently, most publishers and aggregators are proceeding cautiously
- May mean that suitable licensing options eventually become available



["Bay Bridge construction"](#) by [Dan Bluestein](#) under [CC BY 2.0](#)

Unencumbered Works

["grey metal chain in close up photography"](#) by [Matthew Lancaster](#)
under [Unsplash License](#)



Unencumbered Works?

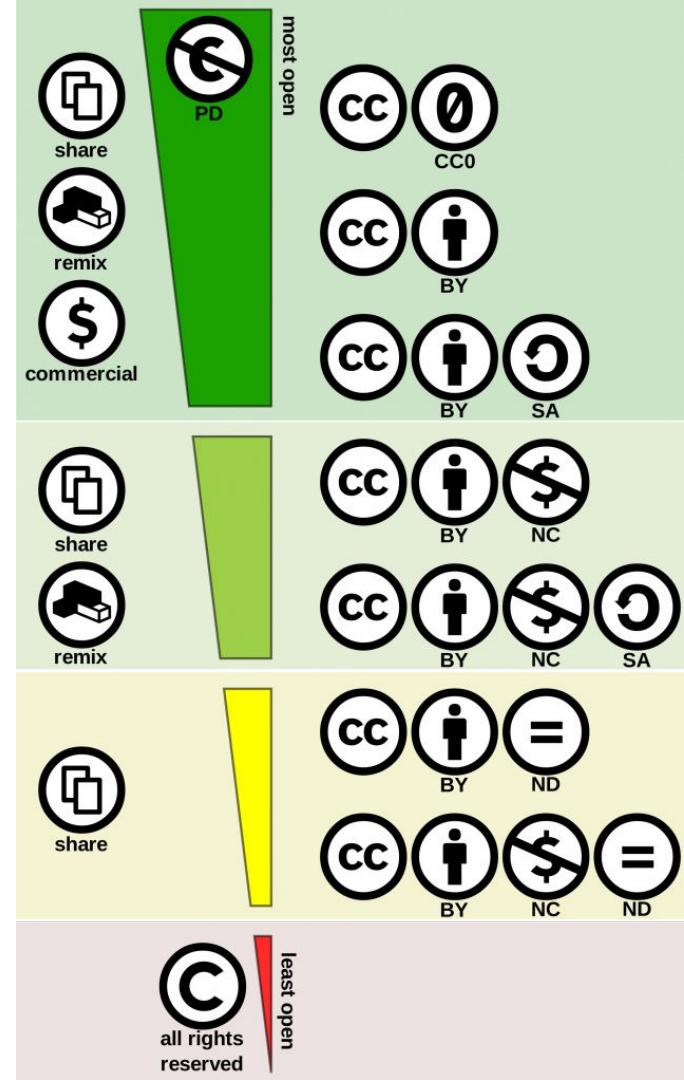
- *Unencumbered* works are without significant licensing or copyright restrictions that would prevent use with AI
- May typically be used, shared, or, even, adapted without needing permission
- Practically, two principal strategies:
 - Find *unencumbered versions* of specific publications of interest
 - Find alternate, *unencumbered publications* to those of interest

[“Hot air balloons in the sky” by Manny Becerra](#) under [Unsplash License](#)



Creative Commons (CC)

- Copyright framework for codifying permissions for open works
- Allows creators to define usage terms re: use, copying, sharing, adapting works
- Best for AI use:
 - **CCØ (Public Domain Dedication)** - creator waives all rights
 - **CC BY (Attribution)** - can copy, modify, use work, with attribution
- Conditional for AI use:
 - **CC NC (NonCommercial)** - cannot be used for commercial purposes



CC License Guidance

“The attribution (BY) and ShareAlike (SA) conditions, and NoDerivatives (ND) restriction are triggered only when works or adaptations of works are publicly shared.”

- [Using CC-Licensed Works for AI Training](#)

Attribution recommendations:

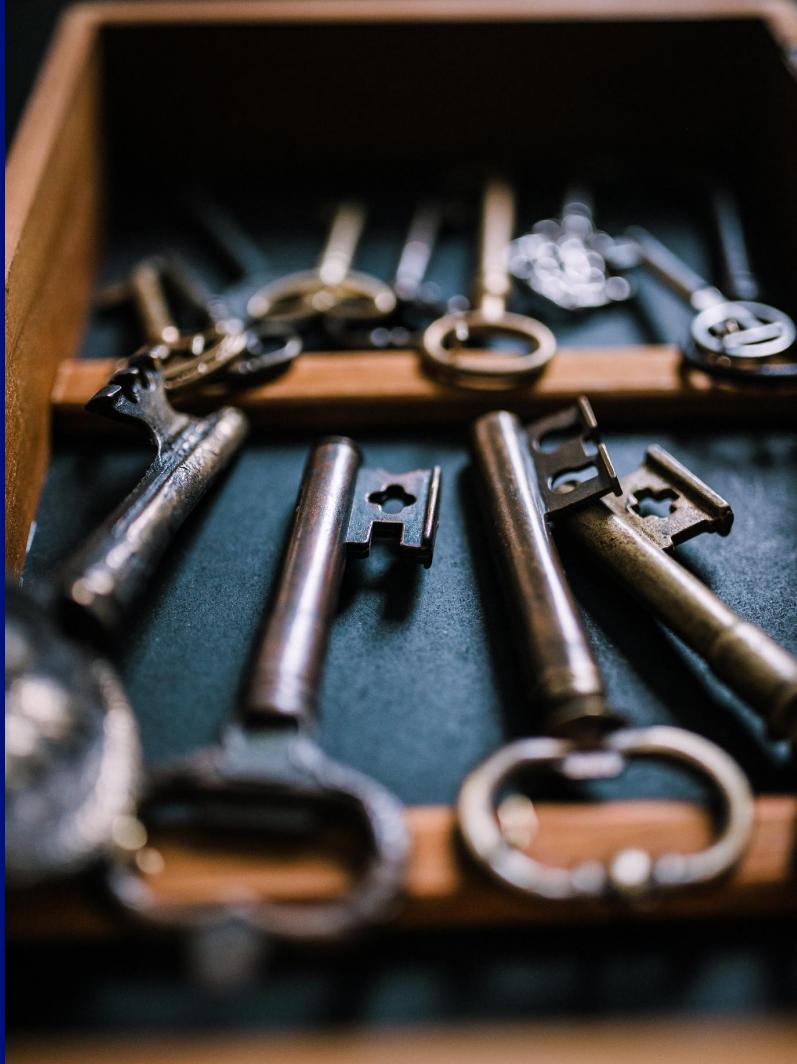
- Keep source metadata, including title, author, source, and license (TASL)
- Share if outputs (e.g., AI model or model outputs) are *publicly* shared

[“Directions” by Irene Bonacchi](#)
under [CC BY-ND 2.0](#)



Find Unencumbered Works

["Details of old keys in a box" by Ivan Radic](#) under CC BY 2.0



Sources for Unencumbered Works



"A large library filled with lots of books" by [Julio Lopez](#) under [Unsplash License](#)

Open Repositories



"Assorted paper stucked on wall" by [Joyce Hankins](#) under [Unsplash License](#)

Open Access Journals

Open Repositories

- Make publications and other open research outputs freely available and discoverable, often with rich metadata
- May offer APIs
- Tend to be more tolerant of automated downloading than journal platforms
- Example open repositories:
 - Pre-prints (e.g., [arXiv](#))
 - Government (e.g., [OSTI.GOV](#))
 - Aggregators (e.g., [CORE](#))

"A large library filled with lots of books" by [Julio Lopez](#) under [Unsplash License](#)



Open Access Journals

- Journals that make their articles freely available online, often (though not always) under CC licenses - check terms
- Journal may be full open access or *hybrid* - i.e., mix of open and paywalled articles
- Rarely offer APIs
- Typically no more tolerant of automated downloading than paywalled journals

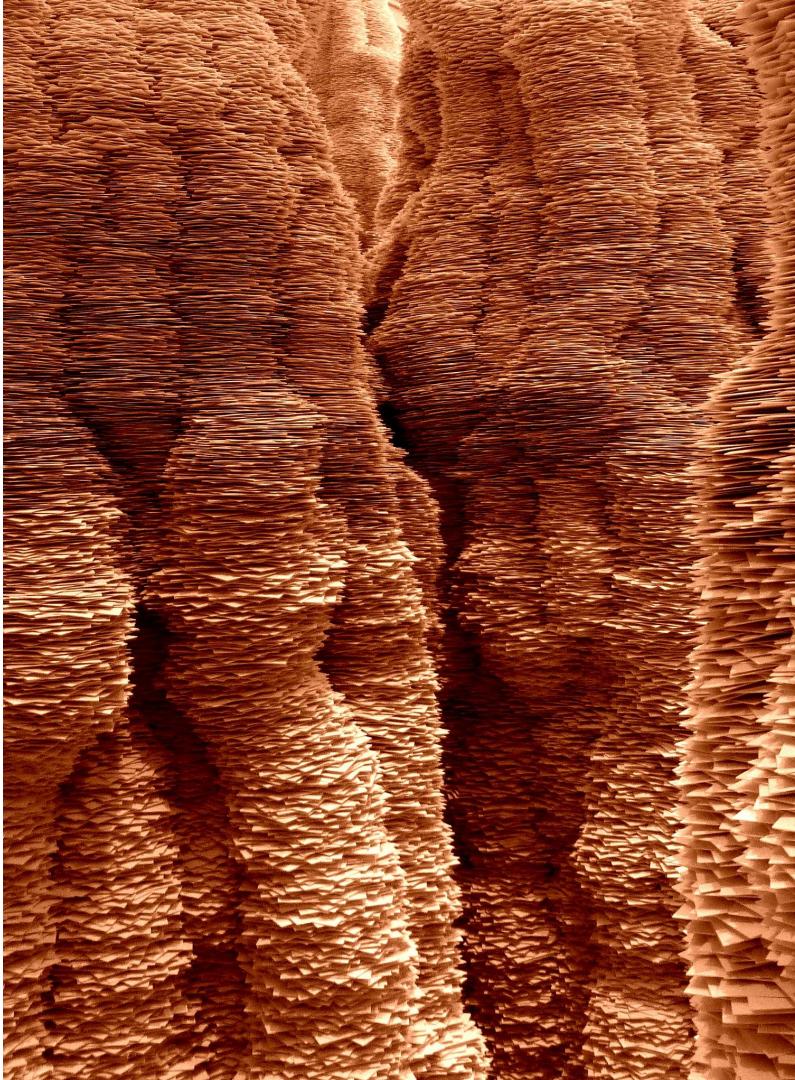
["A room filled with stacks of books on top of a hard wood floor"](#) by [Rebecca Hausner](#) under [Unsplash License](#)



Open Content Indexes

- Metadata aggregation services focused specifically on open scholarly works
- Key tool: [Unpaywall](#)
 - Harvests metadata for open works from 50k sources
 - Provides [REST API](#) and [bulk DOI query web interface](#)
- See also: [CORE](#)

["Untitled by Tara Donovan Nov 27, 2015, 11:57 AM"](#) by [F Delventhal](#)
under [CC BY 2.0](#)



Other Legal Approaches

["Supreme Court detail"](#) by [Dave Newman](#) under [CC BY 2.0](#)



Fair Use

A legal doctrine that allows for limited use of copyrighted material without permission from the rights holder, under certain circumstances

Rests on judicial analysis of 4 factors:

- Purpose and character of use (e.g., commercial vs. educational)
- Nature of the copyrighted work
- Amount, substantiality of portion used
- Effect of use on potential market

"goal poster" by [EvelynGiggles](#)
under [CC BY 2.0](#)

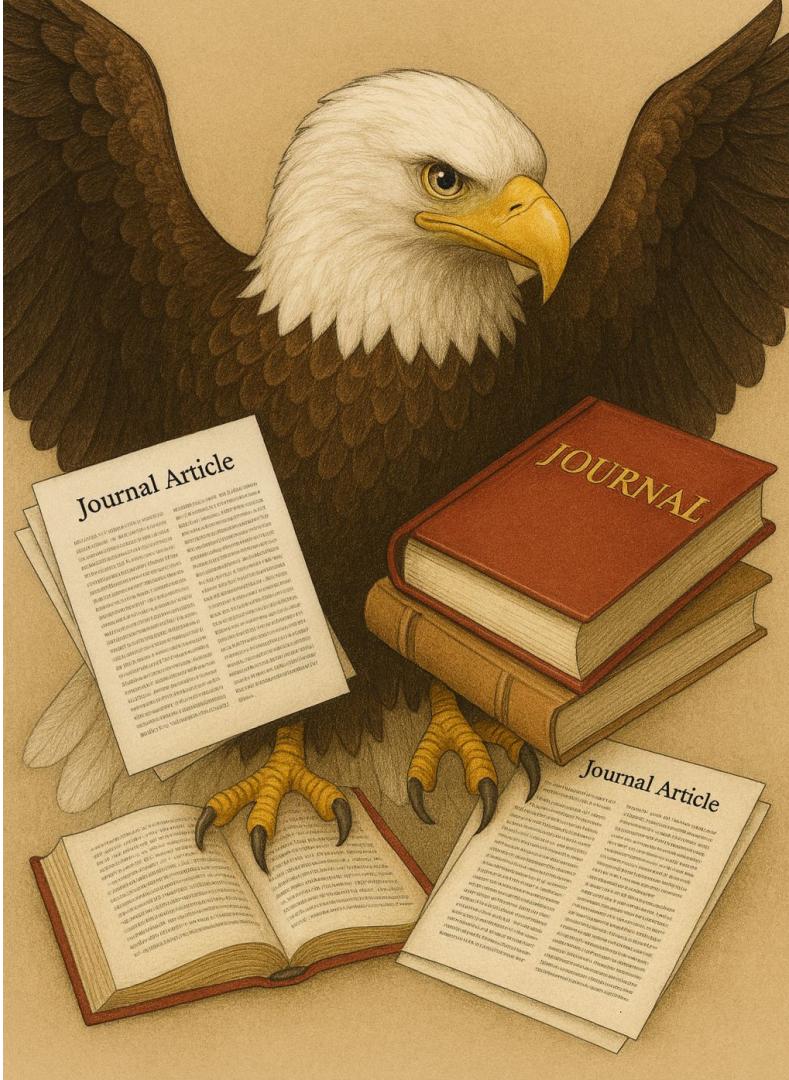


Government Use Rights

The U.S. Federal Government has by law a non-exclusive, irrevocable, worldwide, royalty-free license to exercise or authorize others to exercise all rights under copyright to use a federally funded work for Federal purposes

Also known as the *Federal Government License* or *Government Purpose License*

created using [ChatGPT](#) with prompt,
“create an image of a bald eagle
grasping journal articles and books”



Looking Forward

[“Balancing on the rail” by b.k under CC BY-SA 2.0](#)



What You Can Do

As a researcher using AI

- Request the resources you need, for the purposes they're needed
- Be prepared to be flexible
- Be a courteous downloader
- Abide by content access terms

As an author

- Retain copyright in author agreements, when possible
- Publish under permissive open access licenses (e.g., CC BY)
- Consider permissions you'd want in others' content; act accordingly for your own

What to Watch For

- Maturation of licensing options (market)
- Additional court rulings (law)
- Evolving community practices (norms)
- New technical frameworks for copyright owners to communicate preferences for content use with AI (code)
 - e.g., [CC Signals](#), [Cloudflare Permission-Based Model for AI Crawling](#), [IETF AIPREF](#)

["Choose your reality"](#) by [Raphael Labaca Castro](#) under [CC BY-SA 2.0](#)



Where to Learn More

- LANL Research Library offers public, curated resource guides on AI: lanl.libguides.com/ai_ml
- Including on AI Training Data: [Open and Legal Sources](#)
 - Features practical guidance on finding and using unencumbered works

 **Research Library**

Los Alamos National Laboratory Research Library | LibGuides | Artificial Intelligence/Machine Learning | AI Training Data: Open and Legal Sources | Home

AI Training Data: Open and Legal Sources

Home

- Creative Commons Licenses
- Open Access Articles & Books
- Technical Reports & STI
- Federal Government Licenses & Accepted Manuscripts
- Open Access Repositories
- APIs & OA-PMM
- Fair Use for AI?

Librarian



Rory Elliott

Contact:
ORCID: <https://orcid.org/0000-0002-6023-1105>

Data Sources and Allowed Uses for AI

A Library may subscribe to resources (e.g., journals, ebooks, or databases); however, that does not necessarily mean that someone may copy the content into an AI system for training, fine-tuning, or bulk ingestion. Many publishers explicitly prohibit such uses in their license agreements. In addition, the majority of the Research Library's licenses for electronic resources disallow systematic downloading or scraping.

Potential ramifications for breaking license terms can include:

- Termination of access to resources. The vendor may disable access to all or an institution's users.
- Libraries or institutions may be liable for damages or incur penalties.

This guide is to assist researchers who are trying to locate resources for training or fine-tuning AI models, focusing primarily on resources that are unencumbered or open - to allow for quicker use while following ethical best practices.

Open and Legal Sources for AI Training Data

The foundation of trustworthy, adaptable AI models relies on the data they're trained. One of the most readily available sources of training material comes from open access content and other unencumbered resources.



Definitions

- Accepted Manuscript: the peer-reviewed version of a paper before the publisher's formatting.
- Commercial dataset: a collection of data or datasets used for linguistic, machine learning, or AI purposes.
- Creative Commons (CC) Licenses: licenses that allow creators to specify how others may use their works.
- Fair Use: a legal doctrine that allows limited use of copyrighted material without permission for purposes such as teaching or research. Applicability is determined on a case-by-case basis.
- Fine-tuning: the process of taking a pre-trained AI model and continuing its training on a specialized dataset—such as scientific literature in a specific field—in order to adapt the model's capabilities to that domain.
- Hybrid journal: a subscription-based journal that allows authors to pay to make individual articles openly accessible under a license, while the rest of the content remains behind a paywall.
- Licenses: legal agreement or permission that defines how a user can access, use, or share a work. In the context of OA, Creative Commons licenses are the most common type.
- Open Access (OA): scholarly content made freely available online, typically without subscription or paywalls, and often accompanied by open licenses (e.g., Creative Commons) that specify how the work may be reused.
- Public domain: content that is not protected by copyright and may be freely used by anyone for any purpose, often because copyright has expired or never applied (e.g., U.S. federal government works).
- Training data: data used to teach a machine learning model to recognize patterns or make decisions. In the context of AI, it can include text, images, code, or any structured or unstructured content.
- Unencumbered works: works that are free from copyright restrictions, licensing fees, or legal barriers that would otherwise limit how the material can be reused. These works can typically be used, shared, or even adapted without needing permission.

Last Updated: Jul 1, 2025 4:14 PM | URL: <https://lanl.libguides.com/ATrainingData> | Print Page

Tags: AI, Artificial Intelligence, open access

Managed by Triad National Security, LLC for the US Department of Energy's NNSA. Copyright Triad National Security, LLC. All Rights Reserved.

[Login to LibApps](#)

Thank You!

"*Gratitude*" by [JustOneMoreBook!](#)
[Children's Book Podcast](#) under CC
BY-NC-SA 2.0

