Los Alamos NATIONAL LABORATORY

Publications as Data for AI

Nicholas Taylor and Rory Elliott

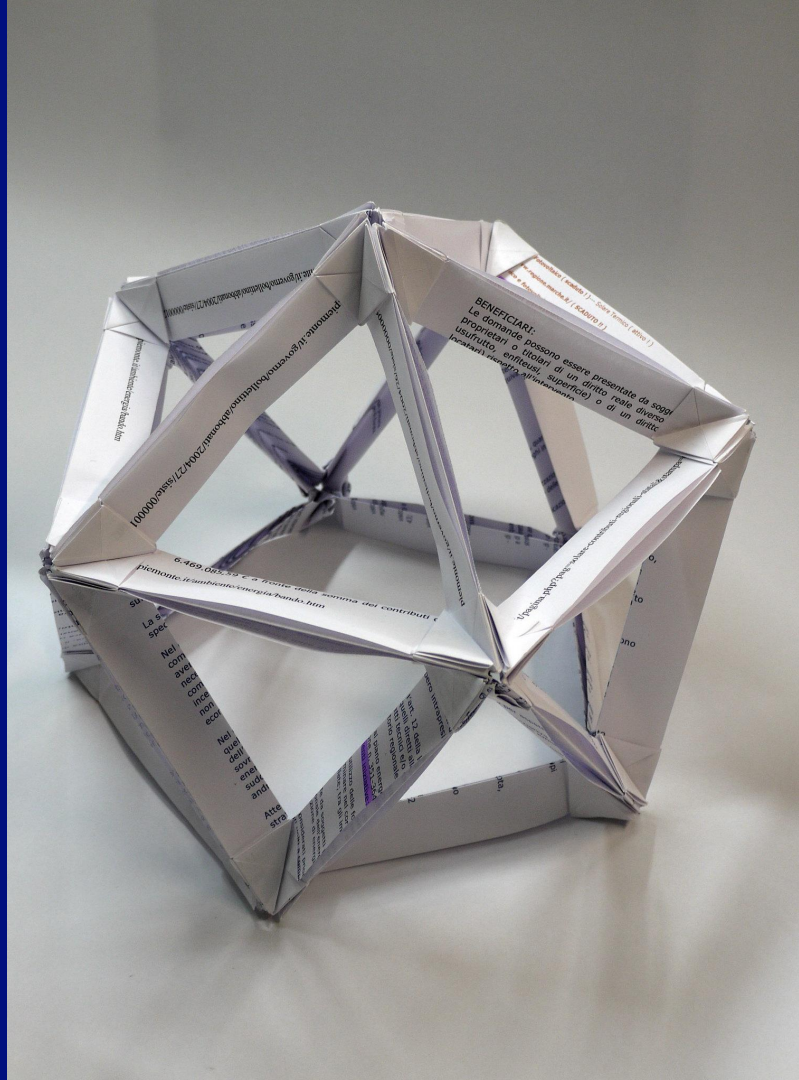
Research Library

DOE Fusion XIII

17 September 2025

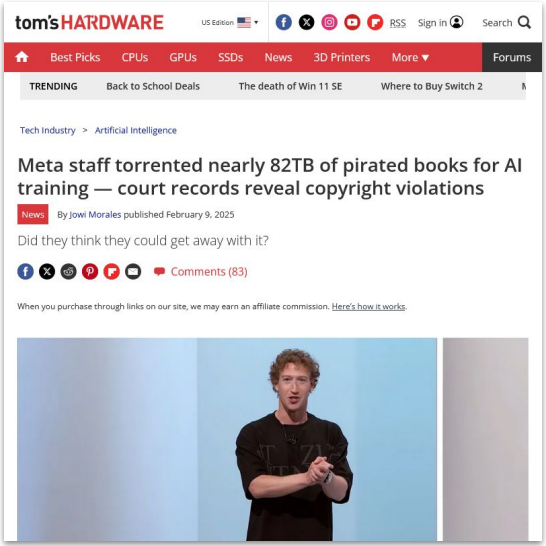
LA-UR-25-29019

Publications as Data Are Foundational to Powerful AI



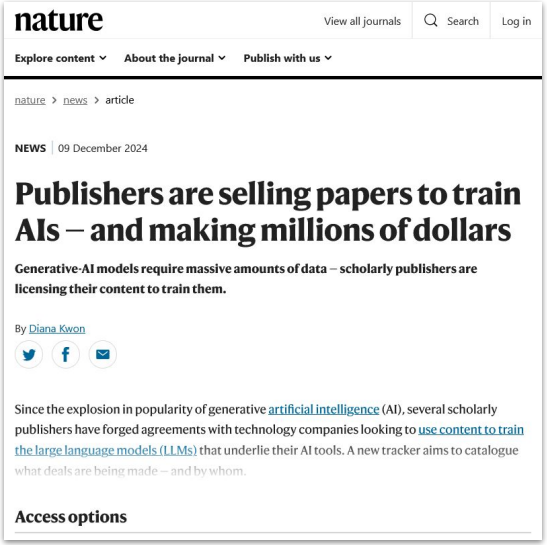
LOS ALAMOS NATIONAL LABORATORY Research Library

The Problem of Publications as Data for AI



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.

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

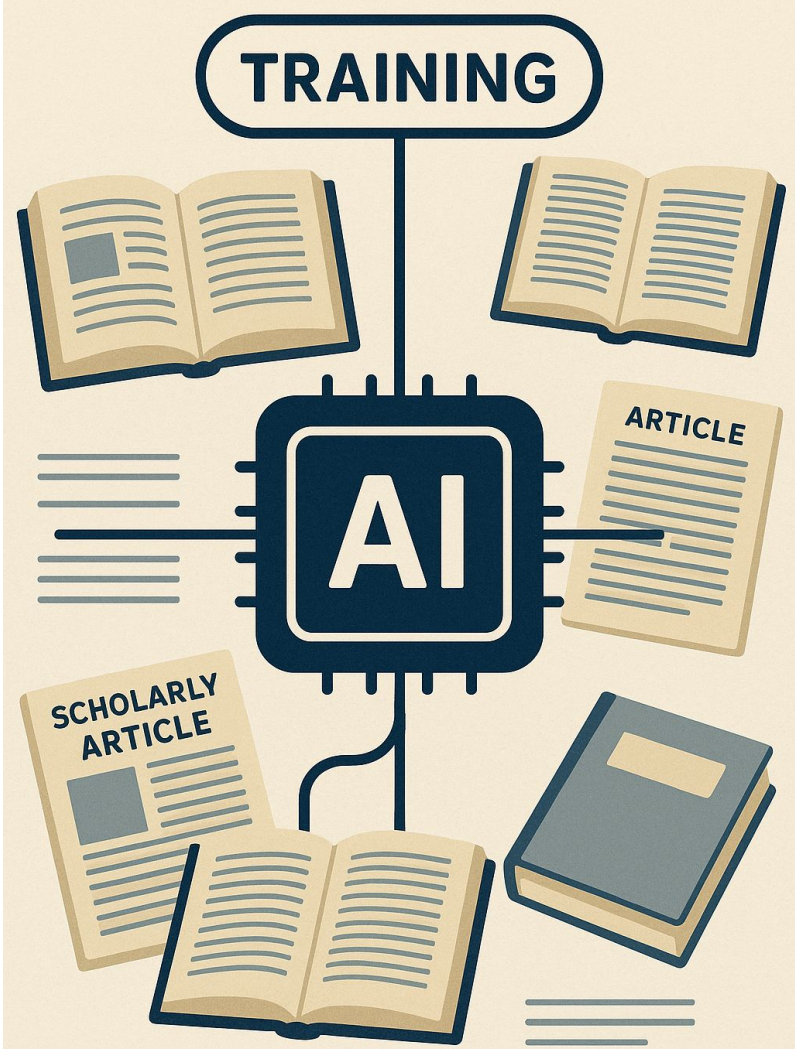


Kown, Diana. "Publishers are selling papers to train AIs – and making millions of dollars". Nature, 9 Dec. 2024. Accessed 15 Aug. 2025.

LOS ALAMOS NATIONAL LABORATORY
Research Library

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



created using ChatGPT with prompt, “generate an image of scholarly articles and books being using to train artificial intelligence”

LOS ALAMOS NATIONAL LABORATORY
Research Library

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

LOS ALAMOS NATIONAL LABORATORY
Research Library

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

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…*

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

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?

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

Summarizing Challenges

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

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

LOS ALAMOS
NATIONAL LABORATORY
Research
Library

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

Unencumbered Works

"grey metal chain in close up photography" by Matthew Lancaster under Unsplash License

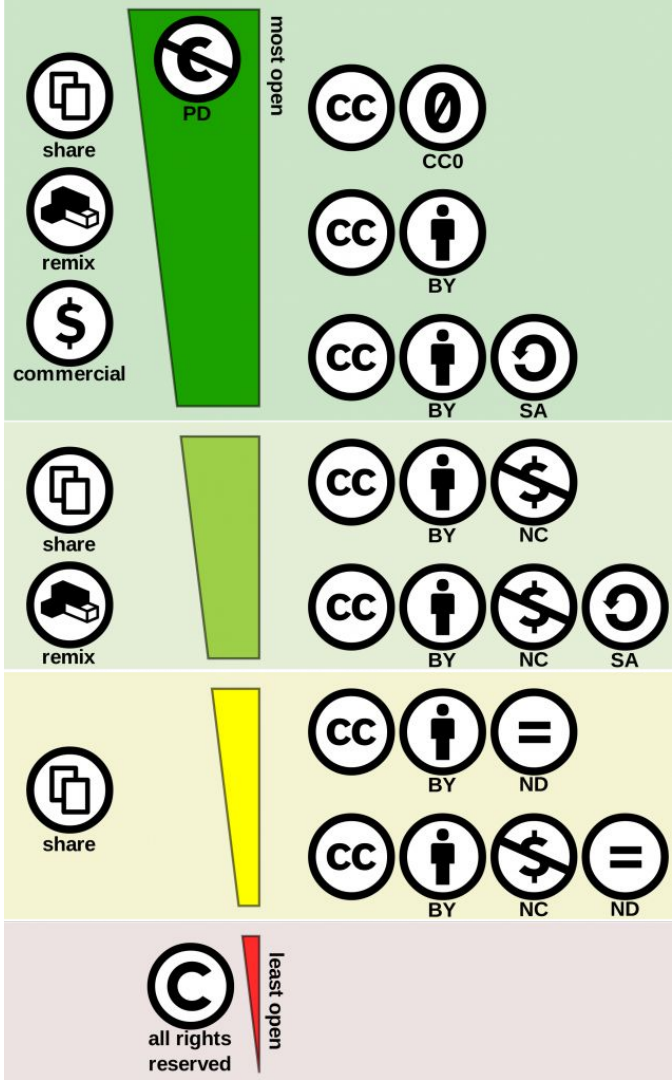
LOS ALAMOS NATIONAL LABORATORY Research Library

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

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



"Creative commons license spectrum" by Shaddim under CC BY 4.0

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

Find Unencumbered Works

LOS ALAMOS NATIONAL LABORATORY
Research Library

Sources for Unencumbered Works

"A large library filled with lots of books" by Julio Lopez under Unsplash License

Open Repositories

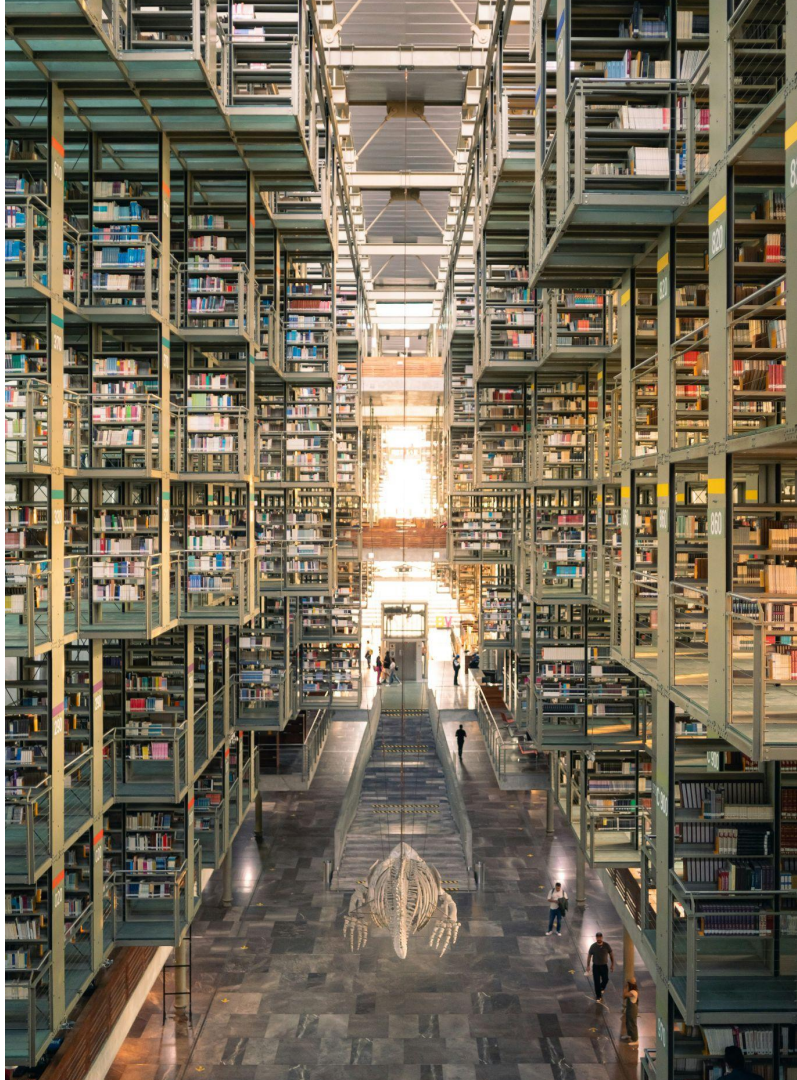
"Assorted paper sticked on wall" by Joyce Hankins under Unsplash License

Open Access Journals

LOS ALAMOS
NATIONAL LABORATORY
Research
Library

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)

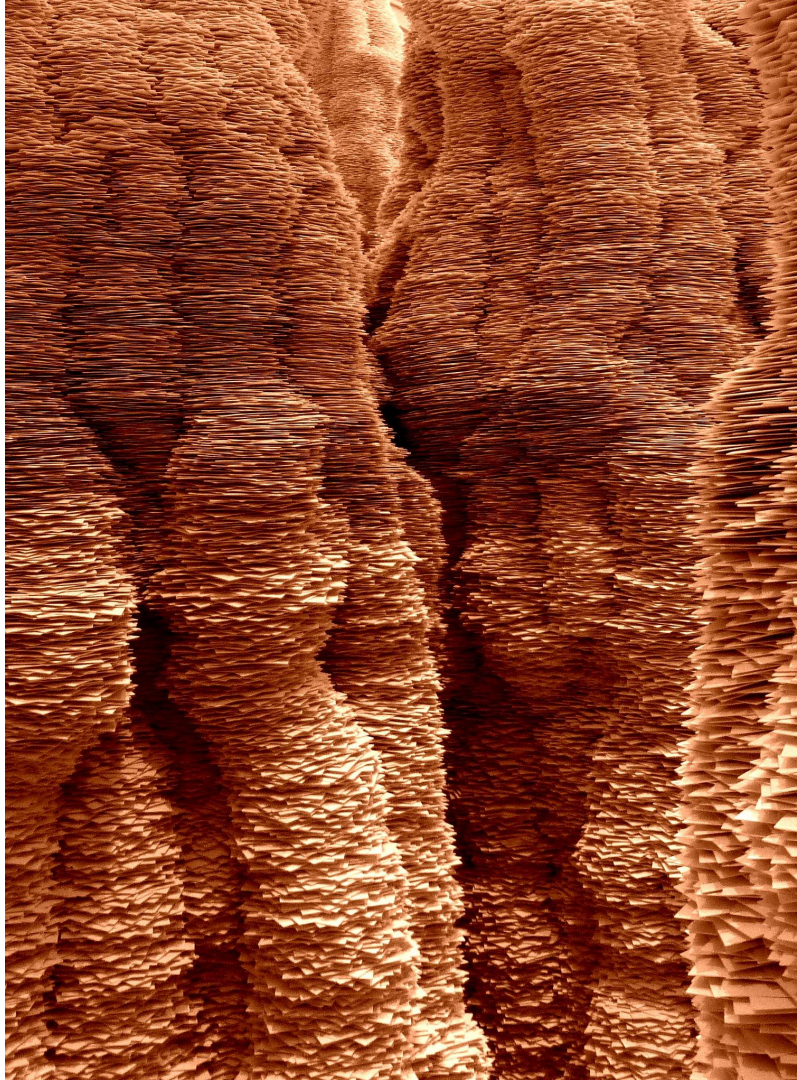


Open Access Journals

- 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

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



Other Legal Approaches

LOS ALAMOS NATIONAL LABORATORY
Research Library

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

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

LOS ALAMOS NATIONAL LABORATORY Research Library

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

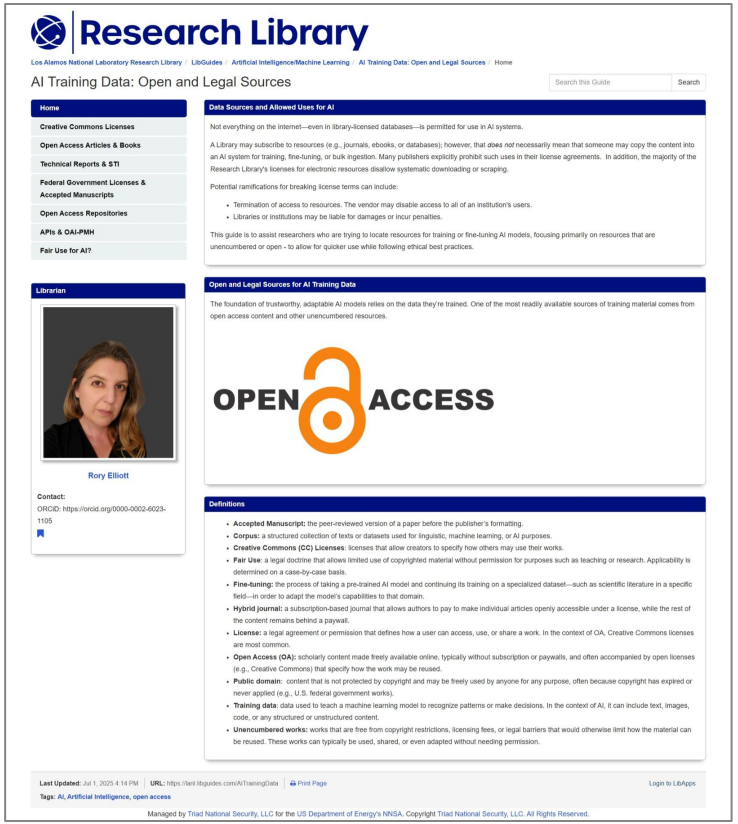
- Retain copyright in author agreements, when possible
- Publish under permissive open access licenses (e.g., CC BY)
- Consider permissions you'd want in others' content; act accordingly for your own

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

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



lanl.libguides.com/AITrainingData

LOS ALAMOS NATIONAL LABORATORY Research Library

Thank You!

LOS ALAMOS NATIONAL LABORATORY
Research Library