CHOICE WEBINAR · JULY 14, 2026

When AI Comes Knocking

A decision framework for archival collections

Leo S. Lo

Dean of Libraries · Special Advisor to the Provost for AI Literacy

Professor of Education · University of Virginia

IN THE NEXT 45 MINUTES

Three things you will leave with

1

The reframe

How to distinguish disposition decisions from access decisions — and why that distinction determines everything that follows.

2

The method

A step-by-step way to classify AI-related requests, from vendor contract clauses to direct training partnerships.

3

The community

How the Statement of Shared Practice turns individual decisions into collective action — and how your institution can join.

LIVE POLL · WHILE WE GET STARTED

Which best describes your role?

A **Special collections / archives**

B **Department head / administration**

C **Reference, instruction, or user services**

D **Other**

Answer in the poll panel →

01

THE SHIFT

People stopped searching. They started asking.

How People Use ChatGPT*

Aaron Chatterji[1,2] Tom Cunningham[1] David Deming[3] Zoë Hitzig[1,3]
Christopher Ong[1,3] Carl Shan[1] Kevin Wadman[1]

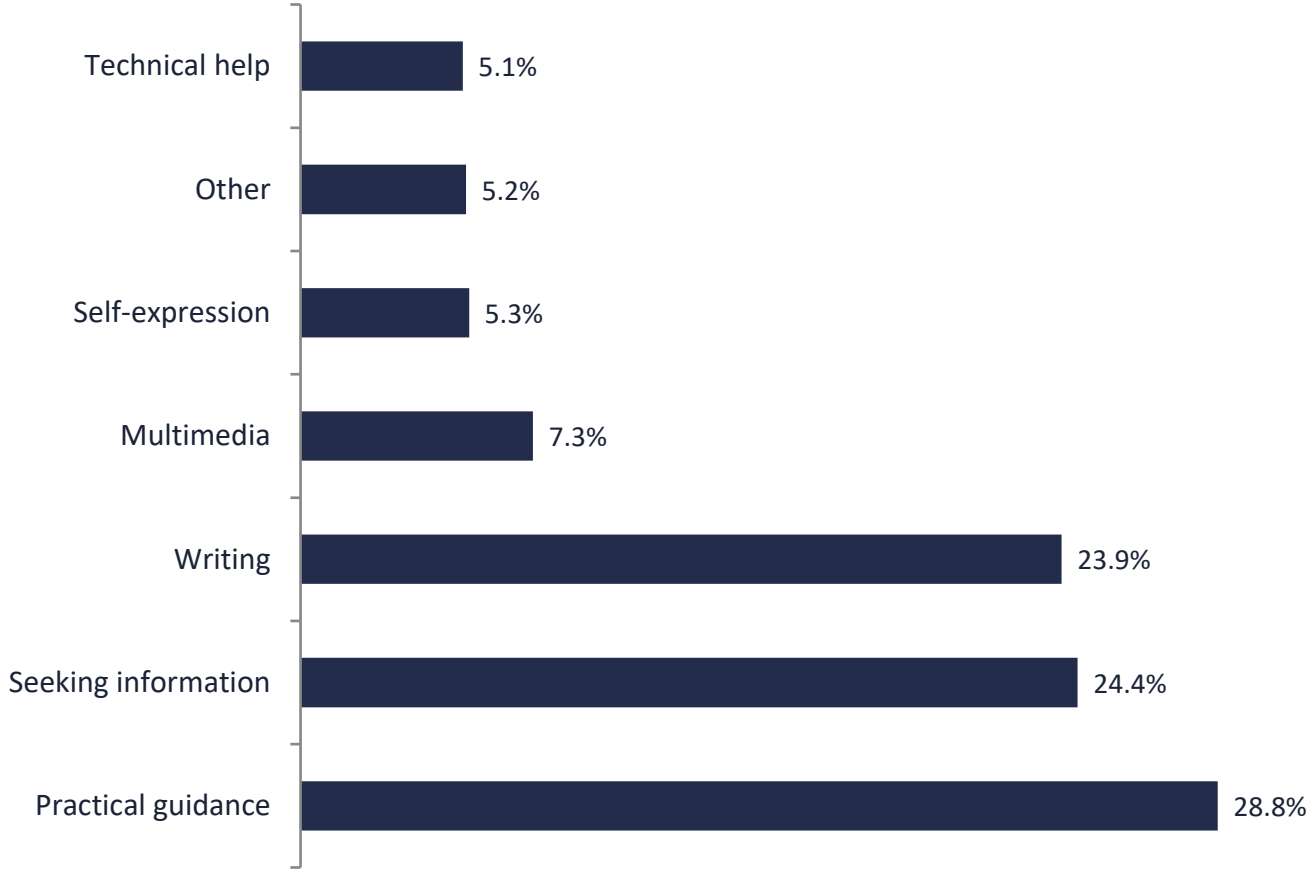
[1]OpenAI
[2]Duke University
[3]Harvard University

September 15, 2025

Abstract

Despite the rapid adoption of LLM chatbots, little is known about how they are used. We document the growth of ChatGPT's consumer product from its launch in November 2022 through July 2025, when it had been adopted by around 10% of the world's adult population. Early adopters were disproportionately male but the gender gap has narrowed dramatically, and we find higher growth rates in lower-income countries. Using a privacy-preserving automated pipeline, we classify usage patterns within a representative sample of ChatGPT conversations. We find steady growth in work-related messages but even faster growth in non-work-related messages, which have grown from 53% to more than 70% of all usage. Work usage is more common for educated users in highly-paid professional occupations. We classify messages by conversation topic and find that "Practical Guidance," "Seeking Information," and "Writing" are the three most common topics and collectively account for nearly 80% of all conversations. Writing dominates work-related tasks, highlighting chatbots' unique ability to generate digital outputs compared to traditional search engines. Computer programming and self-expression both represent relatively small shares of use. Overall, we find that ChatGPT provides economic value through decision support, which is especially important in knowledge-intensive jobs.

Seeking information is now a quarter of all chatbot use



Three uses dominate.

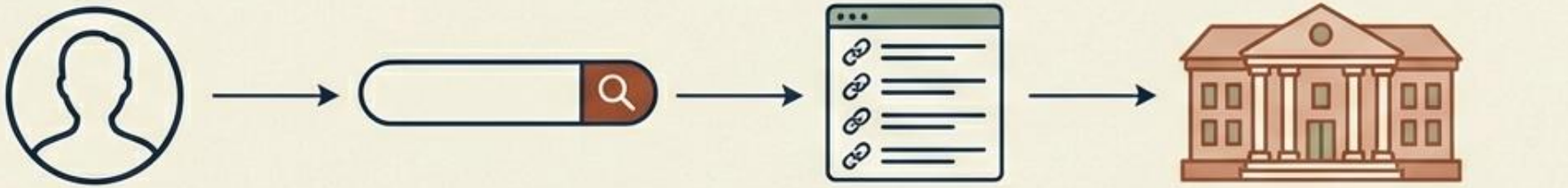
Guidance, information seeking, and writing account for nearly 80% of all conversations.

Systems answering those questions need authoritative sources. That is where we come in.

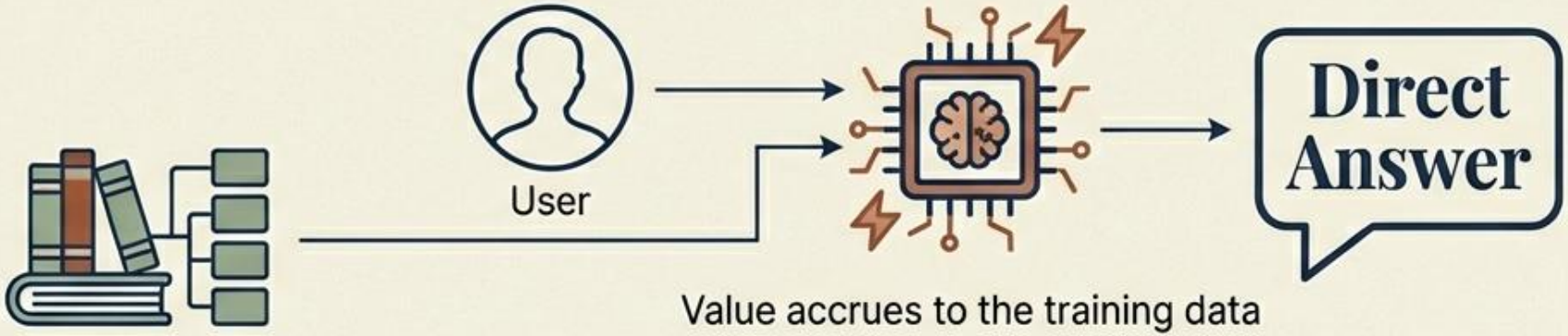
Chatterji et al., "How People Use ChatGPT," NBER Working Paper 34255 (2025). 1.1M sampled conversations.

The Demand: From Search to Answer Engines

Old Model (Search)



New Model (Answer Engine)



Core Argument: AI systems are evolving from retrieving links to delivering direct answers. Requests are targeting the most valuable, unique materials—specifically undigitized and unprocessed holdings. We will face more requests, more urgency, and more pressure to decide quickly.

NotebookLM

The open web is closing. The hunt has reached the archives.

$250M+

reported value of News Corp's five-year content licensing deal with OpenAI (Wall Street Journal).

25%

of data from the highest-quality sources was restricted from major training datasets in a single year (Data Provenance Initiative).

The market formed fast: OpenAI's reported opening offers to publishers were $1–5M a year in early 2024. The Associated Press, the Financial Times, The Atlantic, Time, Vox Media, and News Corp have all since licensed their archives. Publishers were the first wave.
Unique, unpublished collections are the next.

Libraries are already at the table

BODLEIAN LIBRARIES, OXFORD

Digitisation at scale, funded by OpenAI

A pilot research project, launched February 2025, on digitising public-domain collections at scale — funded by OpenAI as part of its five-year partnership with Oxford. Among its research questions: what impact will AI have on search and discovery of collections?

BOSTON PUBLIC LIBRARY

AI-enhanced digitization at scale

A multi-phase effort announced March 2025 to unlock hundreds of thousands of items previously inaccessible to the public — supported in part by OpenAI's NextGenAI initiative, with Harvard's Institutional Data Initiative, and centering library workers' expertise alongside the technology.

Your donors have noticed

Multiple ARL institutions report that donors are now requesting clauses in gift agreements guaranteeing their materials will not be used for AI training.

The question is reaching acquisition conversations, deeds of gift, and community relationships — whether or not an AI company ever contacts you.

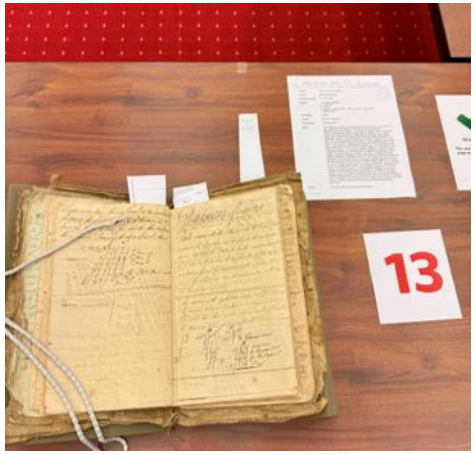
THE RISK

Institutions that cannot demonstrate a clear policy risk eroding donor confidence.

A written framework is now part of donor stewardship.

We experiment enthusiastically — on our terms

Transcription oLLMpics, January 21, 2026



1

Hands on, all staff

Library staff in teams of four, working with original Special Collections documents.

2

Four chatbots, head to head

Copilot, ChatGPT, Gemini, and Claude, tested on transcription of historical materials.

3

One hard rule

We disallow the AI services from ingesting these documents or chats into their training data.

Use the tools. Control the terms. That rule is the rest of this talk.

LIVE POLL · BEFORE WE GO FURTHER

Has your institution received a request involving AI and your collections?

A Yes — directly from an AI company

B Yes — through a vendor contract or renewal clause

C Not yet

D Not sure

Answer in the poll panel →

02

THE QUESTION

When an AI company asks to train on your collections — what do you say?

Training is a disposition decision,

not an access decision.

Simple definitions:

Access is granting permission to use the materials — by whom, how, and on what terms. You stay in charge: whatever you allow today, you can stop tomorrow.

Disposition is deciding what becomes of the materials. Once made, you usually can't take it back on your own.

Once trained, always trained

1

Absorption

Source materials are “baked” into the model weights, not stored alongside them.

2

No isolation

Once training is complete, isolating the influence of specific items is usually impossible.

3

No removal

You cannot un-learn data on demand, and you cannot verify complete removal.

Once trained, control is lost. We cannot un-bake the cake.

This is why disposition, not access, is the right frame for training requests.

THE STAKES, NAMED

Memory without origin

Even a correct answer that cannot be traced back to its source is memory without origin. If all the world’s knowledge worked that way, nothing could be verified against the record it came from.

Provenance is knowledge's paper trail - proof of where it came from.

Training keeps the knowledge and burns the trail.
It is how knowledge stays accountable.

The custodial relationship supersedes ownership

Archives hold materials in trust for donors, records creators, and communities of origin.

The SAA standard

SAA Core Values explicitly identify donor contracts and cultural protections as legitimate bases for access restrictions. Materials are entrusted to archivists, not owned outright.

The ethical line

The SAA Code of Ethics mandates that culturally sensitive material designated by communities should not be disseminated without express permission.

The conclusion

Applying governance to AI is not a departure from library values. It is the direct enforcement of existing use-based access conditions in a new technical environment.

THE QUESTION · COMMUNITY TRUST

Bulk ingestion breaches community governance — permanently

CARE Principles

Indigenous communities retain governance rights over their own records independently of institutional ownership or copyright.

The feedback loop of trust

Community trust → archival stewardship → institutional access → back to trust. Bulk AI ingestion cuts straight across the loop.

THE ULTIMATE COST

Oral histories ingested without consultation enter a commercial system generating unverifiable outputs. The community cannot audit, correct, or remove their voices.

This breach of the custodial relationship is permanent.

Retrieval is not training

PERMITTED DEFAULT

Retrieval & controlled models

- Systems that look at data without absorbing it.
- Source items remain stored and controlled by the archive — and can be removed.
- Example: retrieval-augmented generation (RAG).

PROHIBITED DEFAULT

General-purpose training

- Systems that absorb knowledge into model weights.
- Reversal is not realistic; data is widely redistributed.
- Example: training foundation models (GPT, Claude, Llama).

“Training” is not one thing — require the exact category

RISK: LOW

Indexing & retrieval

Pulling relevant passages for answers. Supports citation and removal.

RISK: MED–HIGH

Fine-tuning

Curated sets shape model behavior in a domain. Traceability weakens.

RISK: VARIABLE

Evaluation & benchmarking

Testing models against content. Sets often drift into training.

RISK: HIGH

General model training

Materials absorbed into weights. Pulling them back out is not realistic.

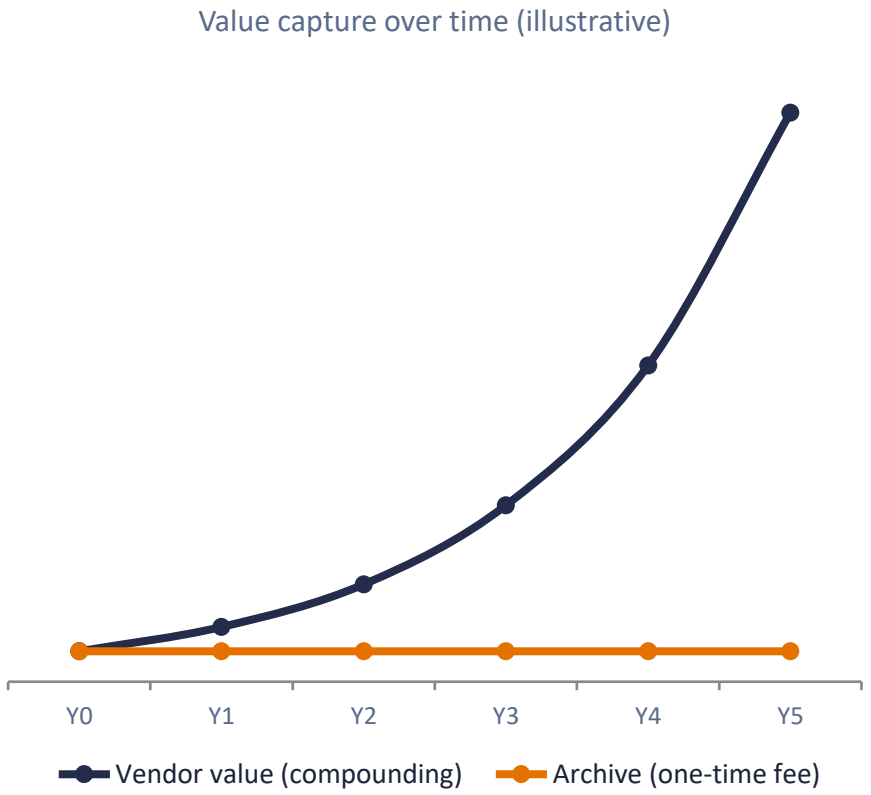
Vendors must specify the category of use — and how they will maintain that boundary.

Undigitized materials are not backlog. They are equity.

The asset. Libraries and archives hold the last scarce asset in the AI training landscape: undigitized, contextualized primary sources.

The one-time fee fallacy. A fixed check for perpetual training rights trades compounding future value for a single transaction. And vendors negotiate institution by institution, setting low baselines one deal at a time.

Once trained, the vendor's value compounds. The institution's leverage collapses.



03

THE PROTOCOL

The UVA Archival AI Protocol

The UVA Archival AI Protocol — one core rule

Irreversible models do not get access unless item-level provenance and meaningful attribution can be demonstrated in practice, and the archival organization retains contractually enforceable control to stop further use.

PILLAR 1

Provenance & attribution

If the AI cannot cite its source, it cannot use the archive.

PILLAR 2

Donor & community obligations

Legal rights are the floor, not the ceiling.

PILLAR 3

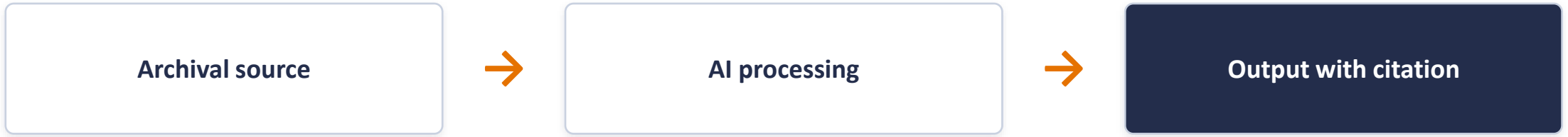
Institutional control

The archive retains the right to stop.

Version 1.1 · released as an adaptable model under CC BY 4.0 · doi.org/10.18130/5dqf-9w86

Provenance & attribution

If the AI cannot cite its source, it cannot use the archive.



- **Traceability.** Systems must record which collections are used.
- **Linking.** Outputs must link back to specific archival holdings.
- **Credit.** Mandatory credit to donors, creators, and communities.

Chain of custody, extended into the machine.

Donor, community & ethical obligations

Legal rights are the floor, not the ceiling.

We honor commitments made in deeds of gift and deposit agreements. AI uses will not proceed if they:

- ✕ Violate deed-of-gift terms or donor correspondence
- ✕ Conflict with cultural protocols or community norms
- ✕ Breach privacy expectations or place vulnerable individuals at risk

THE POLICY

Where obligations are unclear, we decline AI uses rather than risk trust.

Consent first — the default posture, not the exception.

Institutional control — the right to stop

The archive must retain the ability to pull the plug.

Right to cease

The organization can require a partner to stop ingesting content and remove archives from retrieval indexes.

Right to decommission

For narrow models trained mainly on archival collections, we require the right to destroy the model if permission is withdrawn.

Control that survives the signature — contractually enforceable, not symbolic.

Four steps for every request

1. **Rights & ethics check** Deeds of gift, cultural protocols, privacy. Any violation → decline.

2. **Identify the use type** Retrieval, fine-tuning, evaluation, or general training? Require the vendor to specify.

3. **Apply the control test** Retrieval: can we shut it down, do outputs link back? → proceed with control. Broad training: presumption against approval unless it is a narrow model with item-level provenance.

4. **Final control check** Cease and decommission rights in the contract → proceed with caution.

The framework in action: a vendor renewal

The clause: *“Vendor may use licensed and digitized content to develop and improve machine learning models.”*

1. **Rights and ethics check** The digitized collections include donor-restricted material. Proceed with caution.

2. **Identify use type** “Improve models” means training, not retrieval — and it names no category. Heightened scrutiny applies.

3. **Control test** No removal right. No item-level provenance. Presumption against approval.

The response: strike or narrow the clause. Offer retrieval-based access with provenance instead.

The minimum provenance standard

A complete citation

Source: Letter from J. Doe to R. Roe, 1924
Collection: MSS-1024 · Item: box3-folder2-item04
Location: Page 4, paragraph 2

VERIFIED HOLDING ✓

The requirement

If the AI cannot cite, it cannot use. If the system cannot produce citation fields, it must respond: “Insufficient evidence in the retrieved archival sources to answer.”

LIVE POLL · ONE MORE

Does your institution have a written policy or framework for AI training requests?

- **A** Yes
- **B** In development
- **C** No / don’t know

Answer in the poll panel →

04

THE COALITION

No institution knows what AI companies are asking across the field.

Together, we do.

They compare notes. We sign NDAs.

AI companies negotiate with each institution individually, often under nondisclosure agreements that prevent institutions from comparing terms.

The deeper asymmetry is **understanding**: without shared definitions, each institution evaluates requests alone, in vocabulary the vendor chose.

The coalition builds **shared practice** so each institution can decide independently, and decide well.

WHAT WE SHARE / WHAT WE DON'T

We share definitions, standards, and category-level visibility.

We do not share prices, terms, or decisions. Every institution decides alone.

The Statement of Shared Practice

Six voluntary commitments. A shared standard. A community of practice.

1. **Shared definitions — classify every request by category**
2. **Separate digitization funding from AI training rights**
3. **Require provenance, transparency, and removability**
4. **Prefer retrieval-based approaches**
5. **Apply heightened scrutiny to broad training requests**
6. **Share information through a confidential ledger**

THE SHARED LEDGER

Signatories contribute to an anonymized, category-level record of AI training requests: what institutions are receiving, from whom, on what terms.

You are not working through this alone.

What the Statement is not

Not a moratorium

It establishes a floor for responsible practice, not a ceiling. It favors AI approaches that preserve traceability.

Not a boycott

No agreement to refuse any party or set prices. Each institution retains full independent authority over every decision.

Not a legal contract

A professional commitment covering new requests only — special collections, archives, not published works. Withdraw anytime.

Twelve months, reassessed by February 2027. It expires unless affirmatively renewed.

Launched by 12. Growing every month.

12

founding institutions launched the Statement on April 3, 2026.

20+

signatories today — and open to any memory institution.

THE FOUNDING COHORT

University of Virginia (coordinating) · Duke · Florida State · Northwestern · Oklahoma State · Rice · Tulane · Florida · Rochester · Wisconsin–Madison · Washington University in St. Louis · Wayne State

Growing in every direction: Colgate became the first liberal arts college to sign; the University of Liège, the first European institution. Endorsement requires a senior leader's signature and is open to memory institutions.

BEFORE YOU GO

Monday morning

1

Sign the Statement of Shared Practice

library.virginia.edu/ai/shared-practice — your dean or director signs; you can start the conversation.

2

Pull your active vendor contracts

Search for “AI,” “machine learning,” and “improve services” clauses.

3

Decide who receives AI requests at your institution

And where they go next. Most institutions have never named that person.

RESOURCES

Take these with you

Statement of Shared Practice

Full text, FAQ, and endorsement form · library.virginia.edu/ai/shared-practice

UVA Archival AI Protocol

UVA's institutional framework, offered as an adaptable model · doi.org/10.18130/5dqf-9w86

Adoption kit & implementation toolkit

Free under CC BY 4.0 · doi.org/10.18130/jbeg-a995

"Memory Without Origin"

The essay behind the Protocol — the full rationale https://doi.org/10.18130/jyhw-yr80

Leo S. Lo

Connect with me on LinkedIn.
linkedin.com/in/leoslo

Thank you · Leo S. Lo · University of Virginia