MIT MEMEX: A proposal for a provenance-preserving scientific inference library

Ric Fulop and Neil Gershenfeld
(Dated: September 8, 2026)

MIT MEMEX offers a novel licensing-compliant path to scientific inference that preserves publishers' rights and puts MIT's existing computational-access and open-access rights to work. We propose a library-governed recursive language model (RLM) harness for discovering every MIT library paper and DSpace document, with full-text processing governed by each collection's permissions. Using an existing language model, the service would investigate external evidence recursively, preserve source-to-claim provenance, and verify calculations without training library content into model weights. Publisher-approved RAG licensing provides a practical foundation. An existing team-built Google Cloud demonstration supplies the implementation basis, with users querying through Parlay. Illustrative storage for 100 million papers costs approximately $4,700 per month, making the corpus affordable without a new training or GPU inference cluster. Explicit scenarios add 35–42 million scientific articles worldwide over the next decade, increasing both the value of maintained evidence and the strategic cost of unequal access. MIT MEMEX would strengthen institutional research capacity through current scholarship, inspectable reasoning, and reusable results while sustaining publisher attribution and compensation.

I. THE OPPORTUNITY

MIT's library could become an active research capability: a place to ask a complex question and obtain an answer whose evidence and calculations can be inspected. We propose *MIT MEMEX*, a shared inference service spanning MIT Libraries and DSpace@MIT. Inspired by Vannevar Bush's associative trails [1], it would connect research questions to exact passages, extracted data, and reproducible derivations. Its enduring asset would be this maintained scientific record, usable across generations of models.

The ambition is discovery of every library paper and DSpace document. Full-text inference would operate within the rights attached to each source and user. MIT already has relevant institutional open-access licenses and publisher computational-access agreements [3, 24]. A library-governed service would turn those permissions into research tools, adding agreements where needed.

II. FROM LANGUAGE MODELS TO RECURSIVE INVESTIGATION

A large language model (LLM) generates text from patterns learned during training and information supplied within a finite context window. It is prone to *hallucination*: plausible but false statements, numbers, or citations. Information encoded in its weights does not reliably preserve the source or qualifications needed to check a scientific claim.

Retrieval-augmented generation (RAG) supplies external documents at answer time [4]. Direct quotations and source links make evidence checkable, reducing hallucination risk. RAG is compatible with publisher licensing because documents can remain outside model weights and processing can follow agreed conditions. Publishers already offer permissions for specified RAG uses [16–18].

A. Retrieval and recursive investigation

In a basic RAG workflow, a retriever selects passages and an LLM uses them to answer. The external index may cover millions of documents even when the selected passages occupy only a small context window. RAG therefore already separates corpus size from model input size. Its limitation in this workflow is that evidence omitted from the retrieved selection cannot inform the answer. Iterative and agentic RAG can revise searches, use tools, and maintain working state; these capabilities are not exclusive to RLMs.

Alex L. Zhang, Tim Kraska, and Omar Khattab at MIT CSAIL introduced recursive language models in December 2025 [2]. An RLM places long inputs in an external execution environment that the model can examine through code and divide among further model calls. Retrieval supplies evidence; recursion organizes model work over that evidence. The two mechanisms can operate together. Experiments handled inputs up to 100 times the underlying model's context window on evaluated tasks, motivating the proposed architecture (Fig. 1). This result does not establish exhaustive reasoning over a scientific library.

B. Context capacity and external working state

The context window limits the material available to an individual model call. An RLM retains this limit while allowing the investigation to span a larger external collection. Source documents, extracted measurements, and intermediate calculations can remain in persistent storage. Each call receives a bounded portion relevant to its task, and the coordinating model receives selected results and source references. The potential benefits are less irrelevant material in each call, reuse of intermediate work, and selective token expenditure rather than repeatedly placing the entire source collection in a prompt. These

benefits depend on effective decomposition and evidence selection; recursion alone does not guarantee higher accuracy or lower cost.

For example, consider a question about how processing temperature affects a material's strength. Basic RAG might retrieve several relevant passages and summarize them. A MEMEX investigation could search for candidate studies, assign bounded calls to extract processing conditions and strength measurements, store those observations with page references, and use code to separate incompatible test conditions before synthesizing the findings. Follow-up steps could return to the original pages when a comparison is ambiguous. This is a proposed workflow, not a demonstrated corpus-wide meta-analysis.

A larger accessible corpus does not imply that every document has been read. If a question requires examining all records, that work still incurs retrieval, model, and computation costs. Recursive calls add latency, intermediate summaries can omit qualifications, and errors can propagate into the synthesis. Evaluation must therefore distinguish indexed coverage, evidence actually inspected, and correctness of the final claims. MIT MEMEX would preserve source versions, passage locations, and calculation records across these steps, with explicit limits on aggregate work. Provenance enables inspection and correction; it does not certify a model's interpretation.

A *harness* is the software providing a model with tools, working memory, and an execution loop. Summer 2026 results illustrate its importance: the same Claude Opus 5 model scored 30.2% on ARC-AGI-3 in the standard evaluation harness and 95.0%, 95.2%, and 95.5% in an RLM-enabled harness (Fig. 2) [5–7]. These action-efficiency scores compare complete systems with different budgets and orchestration, rather than isolating recursion alone. MIT MEMEX applies the central idea to scholarship: give an existing model a more capable research environment and authorized evidence.

III. A LIBRARY-SCALE RESEARCH ENVIRONMENT

A unified catalog would identify each document's DOI or repository handle, version, source, and permissions. Articles, accepted manuscripts, theses, supplements, and corrections would remain linked but distinct. Approved local stores and publisher interfaces would supply full text; unresolved or restricted items would remain discoverable only within their metadata permissions [8]. Embargoes and sensitive repository restrictions would persist.

Lexical search, semantic retrieval, and structured filters would support an Elasticsearch-compatible or comparable index. Recursive calls would inspect selected evidence while retaining tables and intermediate results outside the model context. Each investigation would have aggregate limits on time, tokens, reads, and concurrency. Broad collection coverage would therefore support selective investigation without requiring every query to scan every paper.

Extraction would preserve page locations, equations, tables, and links to the original image. Generated code would run in a restricted environment, and document text could never change access policy or authorize network transfers. The researcher would receive a source-linked artifact with the searched scope and unresolved gaps, making both conclusions and omissions visible.

Materials research illustrates the value (Fig. 3). Prior studies have extracted synthesis information and useful materials relationships from scientific text [9, 10]. A MEMEX investigation comparing carbon-vacancy energetics in tungsten carbide would distinguish formation energies from migration barriers and retain phase, method, and energy reference. It could then perform a calculation with recorded inputs and units. Observations, derived quantities, and proposed hypotheses would receive separate labels, preventing an attractive extrapolation from appearing to be an established result.

IV. LICENSING, PROVENANCE, AND TRUST

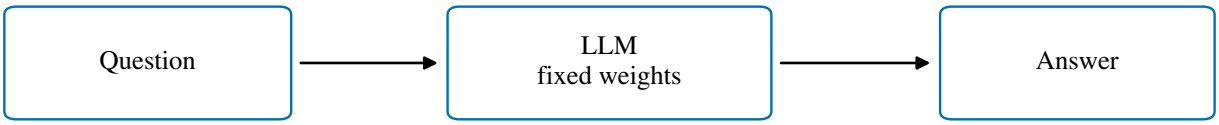
A. Existing rights and publisher precedents

MIT MEMEX's compliance case begins with rights MIT already holds. The faculty open-access license supports harvesting, indexing, and value-added services over covered articles [13]. MIT Libraries reports that its ACM agreement, through December 2028, includes computational text and data mining access to the ACM Digital Library and automatic DSpace deposit of MIT co-authored manuscripts [24]. Its Framework for Publisher Contracts calls for computational access as a standard provision [25]. The framework is a negotiating policy; actual grants establish the usable scope.

Publishers also recognize licensed inference. Wiley offers AI content rights and describes RAG licensing, while Copyright Clearance Center's Annual Copyright License covers specified internal RAG uses for participating content [16–18]. Springer Nature and Wiley describe conditions for particular AI uses [19, 20]. These precedents establish a practical route, although they do not make every subscription an AI license.

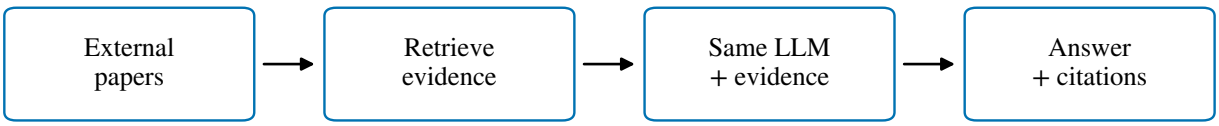
Where an agreement permits retrieval, processing, and synthesis, recursive execution can stay within that permission if every step and the complete workflow meet its conditions. Child calls would inherit the same users, purposes, processors, retention rules, and output limits. Aggregate controls would prevent many small requests from evading a total-use restriction. An explicit recursive-inference clause could resolve architecture-specific wording.

(a) LLM alone



Original papers may be absent at inference time.

(b) Retrieval-augmented generation (RAG)



Retrieved evidence augments the prompt; weights stay fixed.

(c) MIT MEMEX: RLM-enabled inference

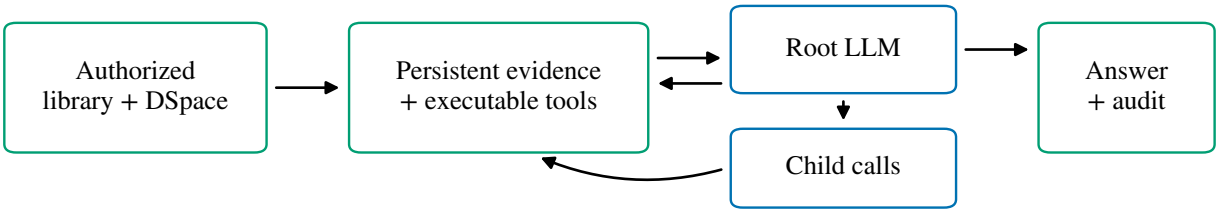


FIG. 1. Conceptual comparison of inference architectures. (a) An LLM can answer from its trained parameters and supplied prompt without having the original source documents available. (b) Conventional RAG adds retrieved evidence. (c) The proposed MIT MEMEX environment adds persistent evidence state, programmatic analysis, and recursive calls, with authorization enforced on every operation. Model weights remain fixed during a MEMEX investigation. Advanced RAG systems can share several of these capabilities; the distinction is the proposed integration, not an exclusive claim to citations or tools.

B. Enforcement throughout the investigation

MIT MEMEX would keep library documents outside model weights and exclude their content and interaction logs from provider training. A separate policy service would authorize each ingestion, embedding, model call, retained artifact, and export before execution (Fig. 4). Decisions would depend on the user, document version, operation, purpose, environment, and current license. Unknown permissions would disable the operation pending review.

Approved quotation and attribution rules would govern answers, with controls against reconstructing restricted works through repeated requests. Retention and termination requirements would cover source copies, indexes, caches, and backups. MIT Press's TDM terms illustrate why this matters: they require special access, prohibit site scraping for TDM, restrict unauthorized training, and require project-end deletion [15]. DSpace documents likewise have differing rights, including thesis ownership and third-party components [14].

This design supports compliant operation by enforcing each agreement's boundaries. MIT's public policies provide a starting point; the executed collection agreements still require mapping to the proposed service [11, 12]. Local execution can reduce third-party disclosure, but processing permissions must cover it. The service would activate existing grants first and obtain amendments for remaining uses.

The *Bartz v. Anthropic* litigation reinforces the distinction between training and acquiring or retaining source copies. The June 2025 order and subsequent settlement concern particular uses, works, and claims [21, 22]. They do not establish that AI companies lost access to all copyrighted science. MEMEX's advantage is access to exact, current, authorized evidence at inference time; proprietary training-corpus overlap and comparative coverage require measurement.

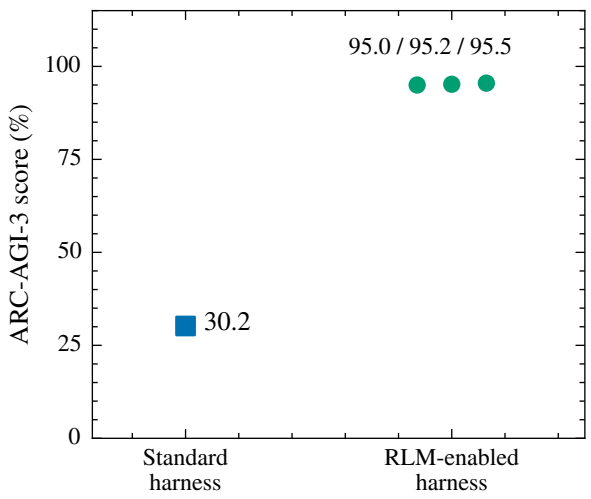


FIG. 2. Same underlying model, different harnesses. Claude Opus 5 scored 30.2% in ARC Prize's standard evaluation harness [7]; the reported RLM-enabled harness scored 95.0%, 95.2%, and 95.5% [5, 6]. Horizontal offsets distinguish runs.

C. Reproducible evidence

Policy checks and specified calculations can be deterministic when inputs and software versions are fixed. Model reasoning can vary. MEMEX would distinguish a recorded replay, using saved responses and tool inputs, from a fresh rerun in which the model reasons again. Replay would preserve only snapshots and records whose retention is permitted.

A provenance graph, consistent with W3C PROV [23], would connect claims to source passages and calculations, including contrary evidence and unresolved disagreements. Hashes verify source identity; separate checks assess whether a passage supports the claim. Researchers could inspect an extraction, change an assumption, and rerun a calculation. This would make the basis of a conclusion easier to challenge and reuse.

V. SCALE AND THE STRATEGIC STAKES

A. Growing evidence without corpus training

The baseline requires zero additional library-corpus weight-training tokens. It reuses pretrained models while paying the real costs of parsing, indexing, storage, inference, and verification. Updating the collection would add authorized evidence without retraining the LLM. Validated extractions and calculations could accumulate for later reuse where retention permits.

About 49% of access-classified scientific articles in the NCSES 2022 dataset were closed access: 1,520,454 of 3,104,987 articles [26]. This is reader-access status, not an AI-use license. No verified industry-wide count of scientific papers removed from commercial training corpora is available from the sources reviewed here.

NCSES reports approximately 3.5 million science and engineering articles worldwide in 2024 [27]. Taking that flow as an assumed 2027 starting point, Fig. 5 estimates ten-year additions as

$$N(10; g) = 3.5 \times 10^6 \sum_{k=0}^{9} (1+g)^k. \tag{1}$$

Annual growth of 0%, 2%, or 4% yields 35.0, 38.3, or 42.0 million new articles through 2036. At an illustrative 10,000 tokens per article, these represent 350–420 billion external-document tokens. They are sizing scenarios, not measured training tokens or guaranteed MIT holdings.

MIT's usable share would depend on collection coverage, processing rights, and technical accessibility. Under the 2% scenario, combined usable shares of 25%, 50%, or 75% yield 9.6, 19.2, or 28.7 million additions. Coverage reports should distinguish cataloged, accessible, authorized, and actually inspected evidence. The resulting advantage should be measured in correct research per researcher-hour, rather than inferred from document count.

B. Accumulating national-security exposure

Unequal access to scientific evidence can become a strategic liability as new research accumulates. Materials, semiconductors, energy, and manufacturing depend on current technical detail. America's AI Action Plan links scientific data and AI leadership to national security [28]. In its September 1, 2026 filing in the consolidated OpenAI copyright litigation, including the New York Times dispute, the Justice Department supported training-stage fair use and warned of advantages for foreign competitors unconstrained by respect for U.S. IP rights. It also recognized licensing for specialized, current, paywalled, or proprietary content [30]. This is the government's litigation position, not a court ruling.

USTR's 2026 review places China among the countries on its Priority Watch List [29]. Uneven enforcement may reduce competitors' acquisition costs or expand their evidence coverage. The relevant comparison is documented access and conduct: national legal exceptions differ, and China also has training-data compliance requirements [31]. National origin alone cannot establish a model's infringement or training corpus.

Growing open access does not necessarily remove the exposure. Closed-access output rose from 716,290 articles in 2003 to 1,520,454 in 2022 while its share fell from 77% to 49% [26]. Figure 6 assumes that share continues falling, from 49% in 2027 to 35% in 2036. At 2% annual

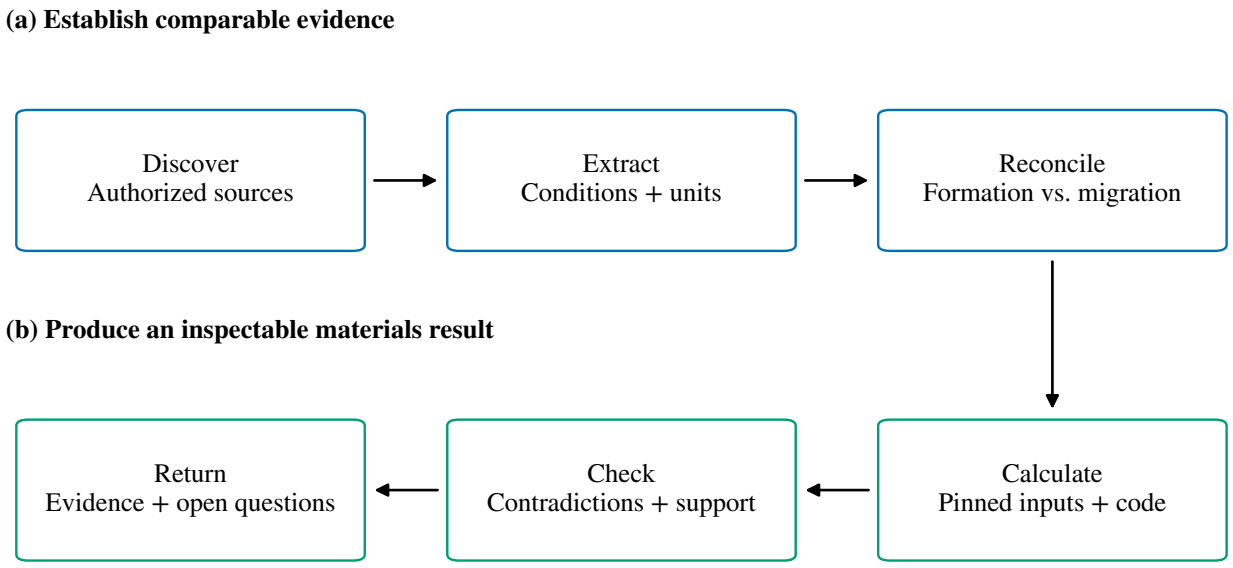


FIG. 3. Illustrative materials investigation proposed for MIT MEMEX. (a) The workflow first preserves conditions and definitions. (b) It then calculates, checks, and returns a research artifact. A source-reported observation, a calculated quantity, and a proposed hypothesis receive different labels in the final artifact. This schematic contains no measured materials values and claims no demonstrated performance gain.

publication growth, the decade still adds approximately 16 million closed-access articles.

If 10%, 25%, or 50% of that flow were usable by a foreign comparator but unavailable to a U.S. workflow, cumulative gaps would reach 1.6, 4.0, or 8.0 million papers. These scenarios exclude an existing backlog and later gap closure through licensing or open access. They quantify potential evidence exposure, not measured U.S.–China differences or intelligence gains. The liability accumulates even if the annual gap declines.

MIT MEMEX offers a practical response: make existing computational rights operational, expand agreed coverage, and retain scientific evidence within approved infrastructure. Sensitive research questions would also require separate research-security controls. A repeatable university-publisher arrangement could strengthen U.S. research capacity while preserving creators' rights.

VI. DEPLOYMENT THROUGH PARLAY

The team has already created a Google Cloud demonstration of MEMEX RLM. MIT MEMEX would extend this implementation to the library corpus, using managed object storage for documents, an authenticated retrieval service, and the existing model runtime. Users would query through Parlay, which would provide the inference interface and route authorized evidence to the RLM harness. The institutional addition is a maintained library corpus and its access rules; a new foundation model or dedicated GPU inference cluster is unnecessary under this arrangement.

TABLE I. Illustrative corpus storage at 2.5 MB per document, using decimal MB/TB. Monthly estimates use regional Standard Cloud Storage in Iowa at approximately $0.020/GiB-month [33]. These cover object storage only.

Documents	Storage	Monthly cost
10 million	25 TB	$470
50 million	125 TB	$2,300
100 million	250 TB	$4,700

Google Cloud Storage holds the papers independently of a compute instance [32]. Storage grows with the collection, while request-serving resources scale with usage. An existing runtime can serve the corpus through an authenticated connector. The connector must carry the user's entitlement into retrieval and restrict delivery to approved processors, preserving the licensing controls described above.

For an explicit sizing estimate, assume a mean source document of 2 MB and an additional 25% for extracted text, metadata, and provenance. The resulting allowance is 2.5 MB per document, or 2.5 TB per million documents. Table I gives three collection sizes. These are planning scenarios, not a measured inventory of MIT holdings; scanned theses, supplements, and retained versions can raise the average. At 1–5 MB per source doc-

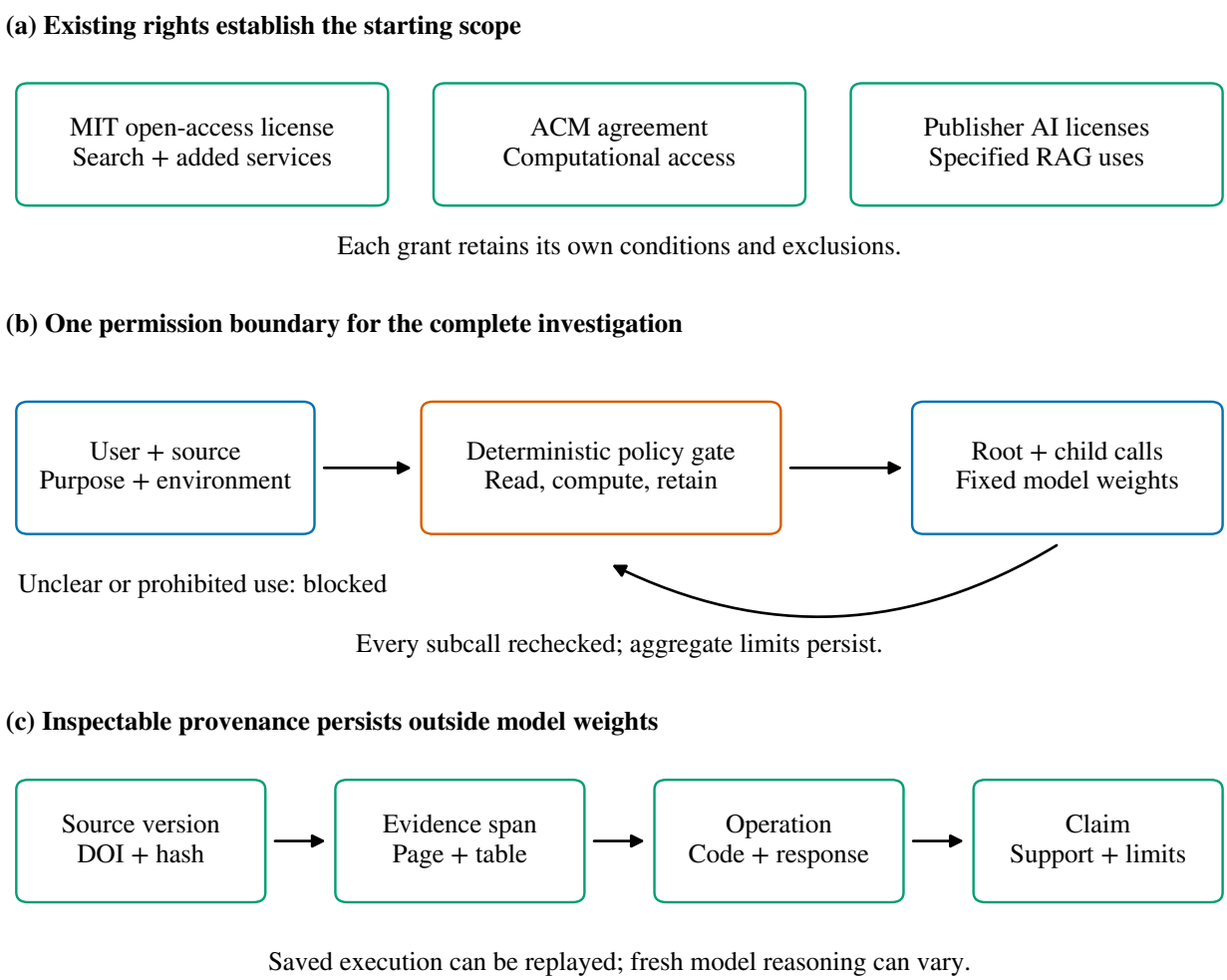


FIG. 4. Proposed rights and provenance controls. (a) MIT's institutional open-access license, current ACM computational-access agreement, and publisher AI licenses establish distinct starting permissions; their scope is not interchangeable. (b) Deterministic policy checks gate each operation and recursive call, including retention, approved environments, and aggregate usage. (c) The evidence record links source versions to exact spans, recorded operations, and answer claims. Saved execution supports replay, while fresh model reasoning can vary.

ument, a 100-million-document collection would instead require 125–625 TB with the same allowance.

Parlay supplies query inference in this design, so the library does not need to duplicate that compute. Incremental operating costs still include ingestion and OCR where needed, search-index capacity, object requests, and data transfer. The table excludes these costs, model usage, and additional backups or retained versions. Existing services may absorb some of this work; corpus size alone cannot determine their throughput requirements. These estimates make a strong affordability case for the corpus infrastructure: even 100 million documents require only about $56,000 per year in modeled object-storage charges. Reusing the cloud implementation and Parlay concentrates institutional investment on connecting and maintaining authorized collections.

Publishers would gain an authorized channel for computational scholarship, with attribution and links to the version of record. Approved interfaces could replace uncontrolled scraping, and version-aware evidence could propagate corrections into subsequent research. Agreements could preserve subscriptions and define any additional inference entitlements, including compensation and retention. Privacy-preserving usage reports could demonstrate collection value without exposing research questions.

Routine service measurement would track claim support, scientific correctness, response time, and cost, alongside access isolation and deletion propagation. Comparing the same model and corpus under standard, iterative RAG, and RLM harnesses would quantify the added value of recursion. MIT MEMEX would make the

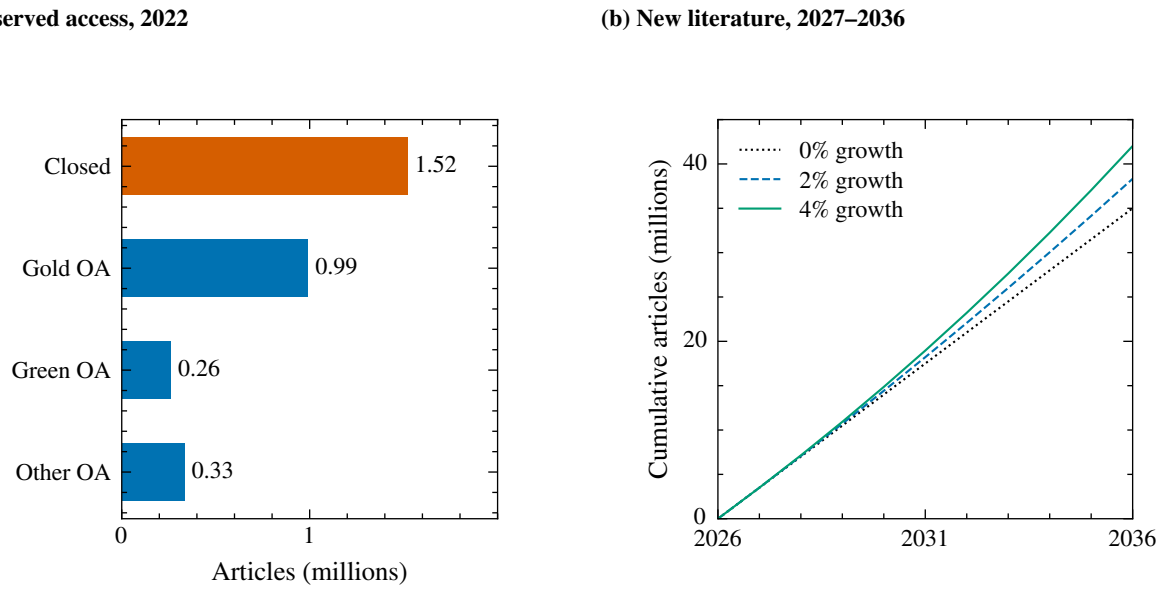


FIG. 5. Observed publication scale and explicit decade scenarios. (a) NCSES 2022 counts for publications with classified access status: closed access, Gold OA, Green OA, and Other OA; categories are mutually exclusive, and unclassified publications are excluded [26]. Reader-access status is not an AI-use license. (b) Cumulative new articles for 2027–2036, assuming an initial annual flow of 3.5 million, anchored to the rounded observed 2024 total [27], and annual growth of 0%, 2%, or 4% thereafter. Lines connect calculated annual scenario points; these are not confidence intervals or forecasts. None of these counts represents papers lost by AI companies or guaranteed MIT coverage.

library a durable scientific memory available through researchers' existing interface, with current evidence and inspectable calculations across the full authorized collection.

DATA AND CODE AVAILABILITY

The existing team demonstration provides the implementation basis; this proposal reports no library-scale performance experiment. Six figures, drawing code, benchmark values, publication counts, scenario assumptions, and a source and claim register accompany it. Diagrams show proposed behavior. No licensed MIT library corpus was ingested.

[1] V. Bush, As we may think, The Atlantic Monthly **176**, 101 (1945), original essay.

[2] A. L. Zhang, T. Kraska, and O. Khattab, Recursive language models, arXiv:2512.24601 (2025), version 3, May 11, 2026.

[3] MIT Libraries, Open access policies at MIT, accessed September 8, 2026.

[4] P. Lewis et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems **33**, 9459 (2020), arXiv:2005.11401.

[5] Prime Intellect, Prime Agent: A self-improving RLM agent (August 2026), accessed September 8, 2026.

[6] S. Karten et al., Prime Agent: A self-improving RLM harness, arXiv:2608.23552 (2026).

[7] ARC Prize Foundation, Claude Opus 5: ARC-AGI results (July 24, 2026), accessed September 8, 2026.

[8] MIT Libraries, Open metadata policy, accessed September 8, 2026.

[9] E. Kim et al., Materials synthesis insights from scientific literature via text extraction and machine learning, Chemistry of Materials **29**, 9436 (2017), doi:10.1021/acs.chemmater.7b03500.

[10] V. Tshitoyan et al., Unsupervised word embeddings capture latent knowledge from materials science literature, Nature **571**, 95 (2019), doi:10.1038/s41586-019-1335-8.

[11] MIT Libraries, Using licensed resources, accessed September 8, 2026.

[12] K. Zimmerman, Legal considerations for your AI project, MIT Libraries (October 20, 2023).

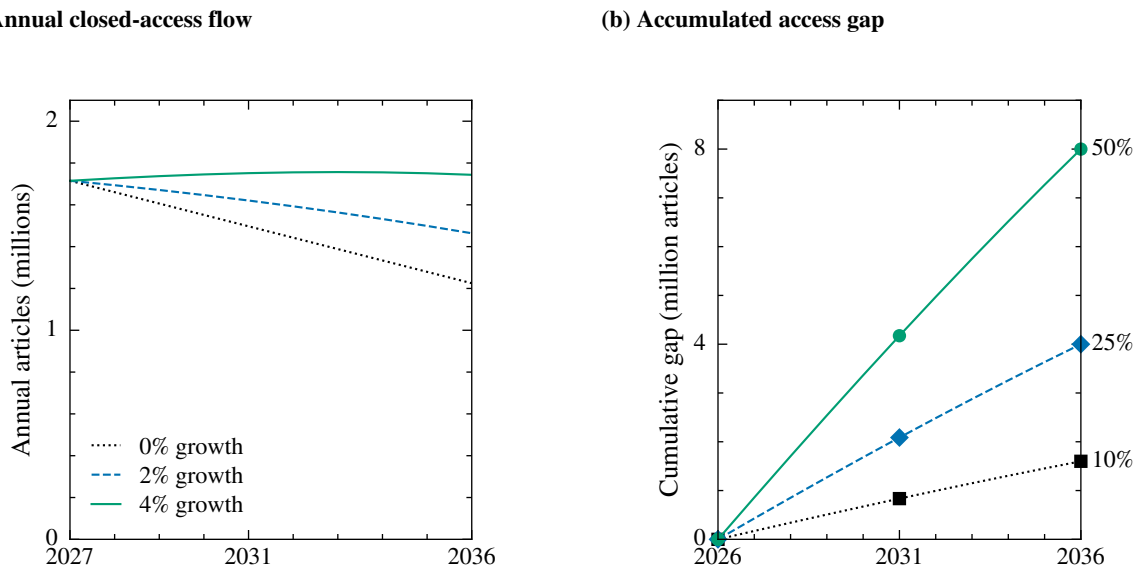


FIG. 6. Estimated strategic exposure despite increasing open access. Both panels assume a linear fall in closed-access share from 49% in 2027 to 35% in 2036, starting from an assumed 3.5 million annual publications. The initial share is anchored to the rounded 2022 NCSES observation, and the annual flow to its rounded 2024 global total; their use in 2027 is a scenario assumption [26, 27]. (a) Annual closed-access volume under 0%, 2%, and 4% publication growth. (b) With 2% growth, cumulative access gaps if 10%, 25%, or 50% of new closed articles are usable by a foreign comparator but unavailable to a U.S. workflow. The dotted, dashed, and solid curves in (a) denote 0%, 2%, and 4% output growth; in (b) they denote 10%, 25%, and 50% access asymmetry, respectively, with endpoint labels identifying those shares. No existing backlog or later backfill is included. Lines connect annual calculations; scenarios are not forecasts, confidence intervals, measured U.S.–China differences, or findings of infringement.

[13] MIT Libraries, FAQ on the OA policy, accessed September 8, 2026.
[14] MIT Libraries, Theses & copyright, updated July 30, 2026; accessed September 8, 2026.
[15] MIT Press, Text- and/or data-mining, Direct Customer Service Desk, accessed September 8, 2026.
[16] Wiley, Wiley AI Rights, accessed September 8, 2026.
[17] J. Jarrett, Breaking down emerging segments in the AI content licensing landscape, Wiley, accessed September 8, 2026.
[18] Copyright Clearance Center, Annual Copyright License: Overview for rightsholders (2026), accessed September 8, 2026.
[19] Springer Nature, AI guidance for our researchers and communities, accessed September 8, 2026.
[20] Wiley, Licensing: Permission for AI use cases, accessed September 8, 2026.
[21] *Bartz v. Anthropic PBC*, Order on Fair Use, No. 3:24-cv-05417-WHA, Dkt. 231 (N.D. Cal., June 23, 2025), court order.
[22] Association of American Publishers, Anthropic publisher settlement FAQs (October 2025); settlement information at the court-authorized settlement website, accessed September 8, 2026.
[23] W3C, PROV-Overview, W3C Working Group Note (April 30, 2013).
[24] MIT Libraries, Open access publishing support: Association for Computing Machinery, accessed September 8, 2026.
[25] MIT Libraries, MIT Framework for Publisher Contracts, accessed September 8, 2026.
[26] National Center for Science and Engineering Statistics, Open-Access Publishing in a Global Context, NSF 25-347 (2025), access-classified publication counts for 2022.
[27] National Science Board, The State of U.S. Science and Engineering 2026: Research publications, NSB-SEP-2026-1 (2026).
[28] The White House, America's AI Action Plan (July 2025).
[29] Office of the United States Trade Representative, USTR releases 2026 Special 301 Report on intellectual property protection and enforcement (April 2026).
[30] United States, Statement of Interest, *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (SHS) (OTW), filed September 1, 2026; associated-case docket No. 1:25-cv-03483, Dkt. 316, pp. 1–4 and n. 13, public copy of filed document.
[31] U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, pre-publication version (May 2025), Sec. IV.G.
[32] Google Cloud, Cloud Storage overview, accessed September 8, 2026.
[33] Google Cloud, Cloud Storage pricing, regional Standard storage, Iowa, accessed September 8, 2026.