EPHEMERAL MODELS, DURABLE CLAIMS: CLOSED-MODEL DEPENDENCE AND THE PRESERVATION OF AI RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are routinely used inside modern AI experiments: to generate or annotate data, to provide comparison points, and increasingly to evaluate other systems. When the exact model artifact can be retained locally, later researchers can often preserve that dependency. When access is mediated only through a provider-controlled service, they usually cannot. Using an end-to-end measurement pipeline validated on a blinded random sample (F1=0.967), we measure this dependence across 53,538 final archival papers from ACL, EMNLP, ICLR, ICML, and NeurIPS (2023–2026) and assess whether the referenced model artifacts persist. We find that 17,841 papers (33.32%) contain a covered closed model in at least one scientific role. Across the common five-venue window, prevalence rises from 14.96% in 2023 to 37.98% in 2025, while use of closed models as judges rises from 1.47% to 11.50%. At a September 24, 2026 lifecycle cutoff, seven of the 20 most frequently used normalized models are already retired; several others are deprecated, on explicit snapshot-shutdown paths, or named too coarsely to identify a unique served artifact. We characterize this exposure and derive practical preservation norms for durable scientific evidence, spanning model reporting, peer review, archival traces, and post-commercial research access.

1 INTRODUCTION

Modern AI papers increasingly use LLMs as components of the experimental workflow. A model can generate synthetic examples, annotate data, act as a comparison baseline, or score another system's outputs. These practices are not specific to closed models. The preservation problem arises from a narrower distinction: whether researchers can retain the exact computational artifact on which the experiment depended.

For open-weight or otherwise locally retainable models, authors can often preserve a versioned copy together with code, prompts, data, and evaluation scripts, subject to licensing and storage constraints. For a model available only through a provider-controlled API or product, the authors normally cannot do the same. They can record a name, a prompt, an access date, and perhaps a snapshot identifier, but continued access to the model itself remains outside the scholarly archive.

That difference matters after publication. If a benchmark is corrected, an old baseline may need to be recomputed. If an evaluation protocol is extended, later systems may need to be scored with the same judge. If a dataset was generated or annotated by a model, researchers may want to audit or replay its construction. A replacement model can support a new experiment, but it does not preserve the historical artifact behind the original evidence.

Prior work establishes the preservation concern; our focus is its prevalence and role structure across recent archival AI research. Johnson & Obradovich (2024) argue that commercial model retirement threatens replication and historical inquiry and propose continued research access and archival mechanisms. Balloccu et al. (2024) document contamination and evaluation-practice problems in a focused audit of 255 papers using GPT-3.5 and GPT-4. What remains unclear is how widely closed models have entered recent archival research, which scientific functions they serve, and whether the models that became common reference points have persisted.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study three questions. **RQ1:** How prevalent is scientific dependence on major closed-model families, and in which roles does it occur? **RQ2:** How has that dependence changed across time, venues, and research tracks? **RQ3:** Which closed models have become the most common scientific dependencies, and what are their lifecycle and identification states as of September 24, 2026?

We answer these questions with a frozen census of 53,538 final archival papers across five major ML/NLP venue families. We separate four non-exclusive scientific roles–DIRECT, BASELINE, JUDGE, and DATA GENERATION–from writing, coding, translation, and similar SUPPORTING USE assistance. We then join the measured model-level dependencies to independently curated first-party lifecycle evidence. The result is not a paper-level "irreproducibility rate"; it is a map of where the recent research record relies on computational artifacts whose identity or future availability is controlled outside the archive.

2 BACKGROUND

2.1 PRESERVING COMPUTATIONAL RESEARCH ARTIFACTS

Reproducibility efforts in machine learning have traditionally focused on code, data, environments, documentation, and experimental configuration (Pineau et al., 2021; Gundersen & Kjensmo, 2018). Software-citation principles similarly emphasize persistent, version-specific references (Smith et al., 2016); FAIR4RS treats software provenance as research metadata (Barker et al., 2022); and Software Heritage archives source code as part of the scholarly record (Di Cosmo & Zacchiroli, 2017). A related digital-preservation literature distinguishes *link rot*, where a referenced object disappears, from *content drift*, where the reference still resolves but its content changes (Klein et al., 2014; Jones et al., 2016).

Closed model services create a difficult case because the citing researcher often cannot archive the referenced executable object. A paper can preserve everything around the model and still depend on continued third-party serving. Provider documentation also makes clear that model names do not always denote immutable objects: OpenAI describes snapshots as a way to lock model behavior (OpenAI, 2026b), while Google documents changes to "latest" aliases over time (Google, 2026a).

2.2 CLOSED-MODEL LIFECYCLES AND SCIENTIFIC ROLES

In this study, *closed model* refers operationally to a model whose weights or equivalent executable artifact are not publicly available for researchers to retain and independently instantiate, and whose relevant public access is therefore mediated by a provider-controlled service. This is a property of access and preservation, not a claim that closed models are uniquely used in research. Our registry covers a broad, prespecified set of major commercial model families rather than every proprietary system ever released.

The preservation consequence depends on how a paper uses the model. Table 1 defines the four roles measured in the study. A baseline is a historical reference point; a judge is part of the measurement procedure; a generator participates in constructing experimental material; and Direct use places the model itself inside the experimental system. These roles are non-exclusive.

Counting BASELINE as scientific dependence is deliberate. Reproduction is not only the ability to rerun the focal method; it also includes preserving the reference conditions against which a claim was motivated and interpreted. If a baseline survives only as a number copied into a table, later researchers cannot recompute it on a corrected split, apply a repaired metric, or test a newly discovered confound. For example, RoBERTa's gains over BERT were evaluated against a baseline whose code and pretrained models were released, leaving the reference condition re-runnable rather than preserved only as a score (Devlin et al., 2019; Liu et al., 2019). We therefore count current and explicitly incorporated prior baseline results when they enter the paper's evidentiary comparison.

2.3 THREE WAYS A MODEL REFERENCE CAN FAIL

We distinguish three preservation failures (Figure 1). **Drift** occurs when a model name or alias continues to resolve but the served behavior changes. **Extinction** occurs when the relevant model or snapshot is no longer callable through the first-party service. **Underspecification** occurs when the

Table 1: Scientific roles measured in the study.

Role	What counts	Why access to the historical model matters
Direct	Model is an experimental object, target, backbone, agent, or system to which the paper applies a method.	Enables later interrogation of the same specimen under corrected or extended protocols.
Baseline	Current or incorporated prior model results form part of the paper's comparison.	Enables recomputation or extension of the same historical reference point.
Judge	Model scores, grades, ranks, critiques, or otherwise evaluates outputs.	Preserves the measurement instrument used to establish reported comparisons.
Data generation	Current authors use the model to generate, synthesize, annotate, label, or construct study material.	Enables auditing or replaying the process that produced experimental inputs or labels.

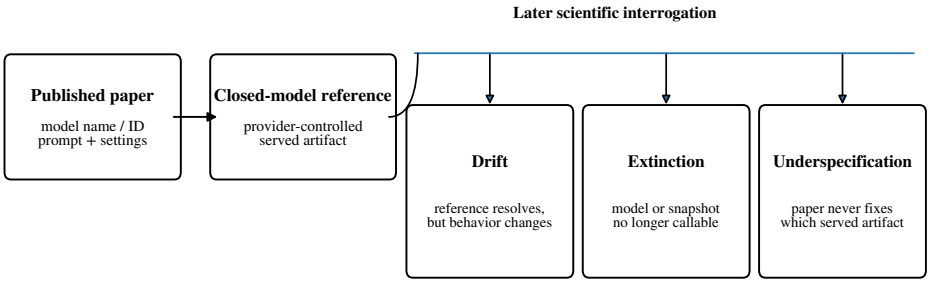

Figure 1: Three preservation failures for a provider-controlled model reference.

paper never identifies one fixed served artifact in the first place, for example by reporting only a broad product label. We use *epistemic durability* as shorthand for the resulting scientific property: whether later researchers can still identify, access, and interrogate the computational object that materially participated in a claim.

This framing is deliberately narrower than claiming full reproducibility. Preserving a model does not recover hidden training data, nondeterministic serving infrastructure, or every deployment detail. Conversely, preserving only an old score does not preserve the ability to rerun a corrected benchmark or apply the same judge to a new system. Work on LLM-as-a-judge already shows that judge choice and evaluation design can affect evaluation conclusions (Zheng et al., 2023; Huang et al., 2025); JudgeArena similarly motivates swappable judges and comprehensive judge metadata (Lushtaku et al., 2026).

3 METHODS

3.1 STUDY POPULATION AND ESTIMAND

We define the study population as the final archival proceedings of ACL (2023–2026), EMNLP (2023–2025), ICLR (2023–2026), ICML (2023–2025), and NeurIPS (2023–2025), including Main/Findings, Datasets & Benchmarks, and position tracks where applicable. The corpus is frozen at 53,538 papers in 34 venue/year/track strata. Our primary estimand is the fraction of archival papers with at least one covered closed model in a scientific role. This is a census of the defined proceedings population rather than a sample from it, so corpus summaries are reported as counts and prevalences rather than sampling-confidence intervals. Cross-venue temporal analyses use the common 2023–2025 window, while ACL and ICLR provide complete 2023–2026 within-venue trajectories.

The unit of prevalence analysis is the paper, while model identity and scientific role are retained at finer resolution. Roles are non-exclusive: the same paper may use several closed models and may

depend on them in several ways. Writing, coding, translation, literature assistance, and similar uses are recorded independently as SUPPORTING USE and are excluded from the primary SCIENTIFIC USE estimand.

3.2 MODEL REGISTRY AND MEASUREMENT PIPELINE

The pipeline uses a prespecified registry of 21 major closed-model retrieval families from OpenAI, Anthropic, Google, and other providers, mapped to 193 normalized model keys. The registry is intentionally broad but not exhaustive. We preserve three identity levels throughout measurement: a retrieval family used to recover lexical variants; a normalized semantic model used for aggregation; and the exact identifier reported in the paper when one is available. The last distinction is required for lifecycle analysis because a normalized lineage can contain snapshots with different serving histories.

Papers first pass through high-recall retrieval over extracted full text, using separator-tolerant matching and family-specific boundary/context guards for collision-prone aliases and version identifiers. Reference-only occurrences are excluded from candidacy, while a secondary extraction channel can rescue plausible parser losses; the full extraction and rescue rules are specified in Appendix D. Candidate evidence is then classified against the four roles in Table 1. Deterministic rules resolve cases where operational use, historical comparison, or non-use is sufficiently explicit. Importantly, role evidence remains bound to a normalized model rather than being inherited from another member of the same retrieval family. Cases in which use is apparent but the role is not, or several models within a family could inherit the same evidence, are escalated rather than forced into a deterministic positive label.

3.3 ADJUDICATION, FREEZING, VALIDATION, AND LIFECYCLE ANNOTATION

Adjudication is deliberately limited to unresolved model entities. The production adjudicator is OpenAI `gpt-oss-120b` (OpenAI, 2025), served locally with reasoning effort set to `low`; initial requests use structured JSON output, temperature 0, and seed 42. All unresolved entities from a paper are packed into one initial adjudication request. The model receives bounded evidence rather than the full paper or the deterministic classifier score. Tight *local* evidence is the only channel permitted to establish a scientific role; broader adjacent context can resolve use versus non-use but cannot donate a role. Positive role decisions must cite evidence for the same normalized model, and LLM output cannot overwrite a deterministic decision. Technical retries are permitted only for unusable generations (e.g., truncation or schema failure), not to seek a different semantic answer. Reconciliation must resolve every queued paper–model decision before the label dataset is frozen for analysis. Figure 2 summarizes the resulting measurement design.

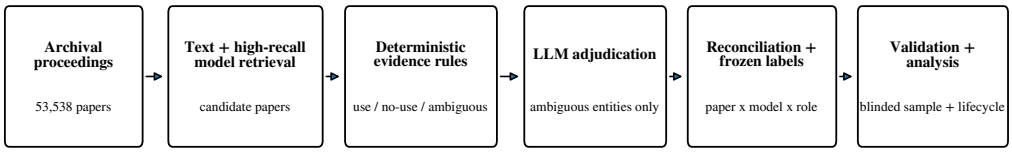

Figure 2: End-to-end measurement design. Deterministic decisions are retained; only unresolved entities enter LLM adjudication before the label dataset is frozen.

We validate the complete measurement procedure on a blinded simple random sample of 200 papers drawn from the full 53,538-paper population. Sampling is from the full population rather than conditioned on candidate status or system prediction, so the metrics evaluate extraction, retrieval, and downstream classification jointly. Final human annotation yields SCIENTIFIC USE precision/recall/F1 of 0.961/0.974/0.967 and four-role micro/macro-F1 of 0.888/0.884. Separately, we manually annotate lifecycle state from first-party provider documentation at a September 24, 2026 cutoff. Usage labels and lifecycle evidence are kept independent: lifecycle is never inferred from model spelling, and

normalized-model status is not converted into a paper-level "irreproducibility" label. Detailed retrieval, adjudication, reconciliation, validation, and lifecycle contracts are given in Appendix D–Appendix F.

4 RESULTS

4.1 OVERALL EXTENT OF SCIENTIFIC DEPENDENCE

Table 2 gives the corpus-level footprint. 17,841 of 53,538 papers (33.32%) contain at least one covered closed-model dependency in a scientific role. Most of this footprint reflects active involvement in the experiment rather than only inherited comparison numbers: 14,936 papers (27.90% of the full corpus; 83.72% of Scientific Use papers) contain DIRECT, JUDGE, or DATA GENERATION dependence, whereas 2,905 papers are baseline-only. Thus the prevalence result does not hinge on our broader decision to count incorporated historical baselines as scientific dependence: only 16.3% of Scientific Use papers are baseline-only.

Table 2: Corpus-level overview. Percentages use all 53,538 archival papers as denominator unless noted.

Quantity	Papers	Share
Scientific Use	17,841	33.32%
Active non-baseline Scientific Use	14,936	27.90%
Baseline-only Scientific Use	2,905	5.43%
Supporting Use	2,486	4.64%

The dependence is also frequently multi-role. Among Scientific Use papers, 61.6% have at least two of the four scientific roles and 25.1% have three or four. A paper may therefore rely on a closed model simultaneously as an experimental object, comparison point, generator, and evaluator. SUPPORTING USE appears in 2,486 papers (4.64% of the corpus), including 886 papers with no measured scientific role. We keep this category separate because assistance with writing, coding, translation, or tooling does not create the same evidentiary dependence as a model that participates in a published measurement or experimental construction.

4.2 TEMPORAL DYNAMICS ACROSS VENUES

Temporal change is visible in both complete within-venue trajectories and the common cross-venue window. Figure 3a follows ACL and ICLR, the two venues with complete archival coverage through 2026. ACL moves from 16.46% Scientific Use in 2023 to 62.61% in 2024 and then remains near two thirds of papers (67.17% in 2025; 65.19% in 2026). ICLR follows a different trajectory: 4.00%, 16.73%, 27.14%, and 33.43% across the same four years. The result is not simply that language research adopted model-mediated workflows earlier; a general-ML venue continues to gain closed-model dependence throughout the observed period.

For a cross-venue comparison, Figure 3b uses the common set of all five venue families observed in 2023–2025. Scientific Use rises from 14.96% to 31.15% and 37.98%. In absolute terms, the number of Scientific Use papers increases from 1,649 to 7,276 while the archival denominator increases from 11,023 to 19,156. The rise is therefore not an artifact of proceedings growth alone. It reflects a larger fraction of papers incorporating covered closed models into their scientific workflow.

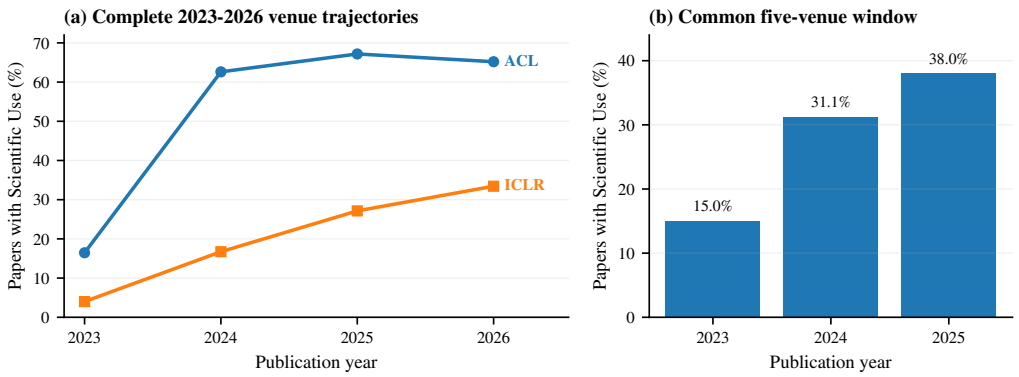

Figure 3: Temporal prevalence using complete within-venue coverage and the common five-venue window.

The models carrying that dependence also turn over quickly. Figure 4 shows absolute paper counts for selected high-frequency generational families in the same 2023–2025 window. GPT-4-family dependence grows from 485 papers in 2023 to 3,152 in 2024 and 6,395 in 2025. GPT-3-family dependence instead peaks at 2,618 papers in 2024 before falling to 1,748 in 2025. Over just 2024–2025, Claude 3 grows from 315 to 1,623 papers, Gemini 1 from 420 to 1,069, and Gemini 2 from 6 to 866. These counts are non-exclusive across families, but they show that the literature is not accumulating dependence on one stable reference system: new generations become scientifically embedded while older generations remain present in historical comparisons and prior experiments.

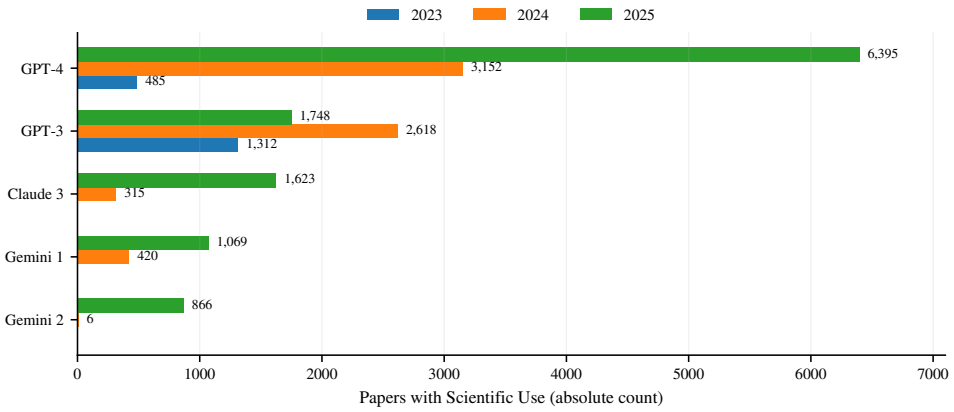

Figure 4: Absolute Scientific Use paper counts for selected model families in the common 2023–2025 window. Families are non-exclusive within papers.

4.3 Where closed-model dependence enters the research workflow

The role distribution shows that growth is not confined to comparisons against commercial systems. In the common 2023–2025 window, Baseline prevalence rises from 10.21% to 24.60%, Data Generation from 7.57% to 19.62%, Direct from 8.07% to 15.29%, and Judge from 1.47% to 11.50% (Figure 5a). Judge dependence is the fastest-moving role: its corpus prevalence grows nearly eightfold, and among Scientific Use papers its share rises from 9.8% in 2023 to 30.3% in 2025. This shift matters because closed models are increasingly used not only as objects of comparison but as part of the procedure that produces the comparison.

Role overlap makes that measurement dependence more consequential. Across the complete corpus, 2,739 Judge papers (53.8% of all Judge papers) also contain closed-model Data Generation, and 3,155 (62.0%) also contain a closed-model Baseline. These intersections do not imply that the same model

served both roles. They do show that model-mediated evaluation is often embedded in experiments that already contain another closed-model dependency, so preserving the judge alone would not necessarily preserve the full historical workflow.

A fixed-year comparison within NeurIPS isolates a different source of heterogeneity (Figure 5b). In 2025, 63.98% of Datasets & Benchmarks papers contain Scientific Use compared with 21.32% of Main-track papers—a 42.7 percentage-point difference within the same venue and year. The gap appears in every measured role: BASELINE (50.3% vs. 12.6%), DATA GENERATION (35.8% vs. 10.1%), JUDGE (19.1% vs. 6.3%), and DIRECT (21.9% vs. 8.4%). Track assignment is not randomized, so this is descriptive rather than causal. Scientifically, however, it identifies a particularly important location of dependence: work whose outputs are intended to become datasets, benchmarks, annotations, and evaluation infrastructure reused by later papers. Our primary estimator counts only dependencies evidenced within each paper and deliberately does not propagate upstream dependencies through reused datasets, benchmarks, or prior work to every downstream user.

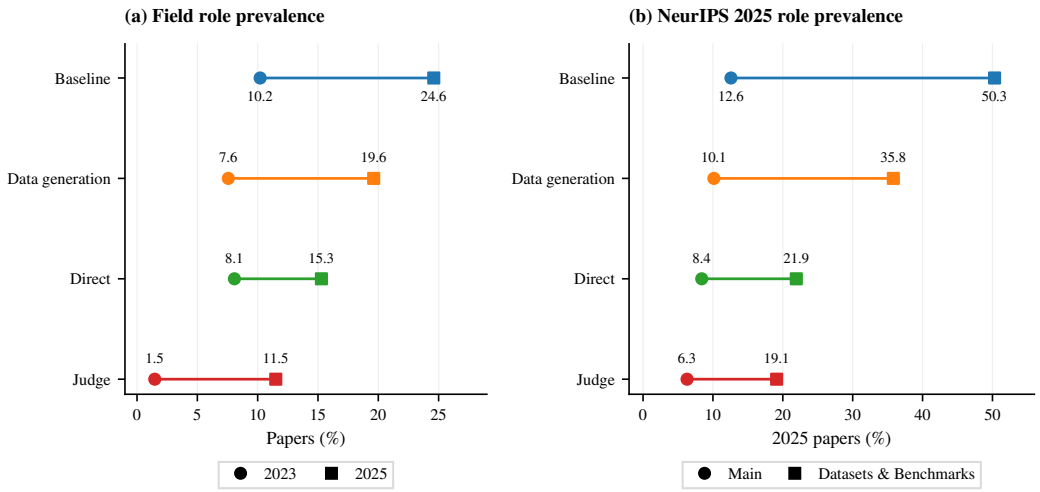

Figure 5: Scientific-role prevalence: (a) field-wide change from 2023 to 2025; (b) NeurIPS 2025 Main versus Datasets & Benchmarks.

Dependence also becomes more composite across organizations. Among Scientific Use papers, the share involving at least two covered providers rises from 7.3% in 2023 to 20.3% in 2024 and 33.6% in 2025; the mean provider count rises from 1.08 to 1.45. The full model/family/provider multiplicity distributions are reported in Appendix G. Multi-provider use does not remove lifecycle dependence; it can place several independently governed service lifecycles inside one experiment.

4.4 WHICH MODELS CARRY THE DEPENDENCE, AND DO THEY PERSIST?

Scientific dependence is concentrated in model lineages that are already experiencing different lifecycle outcomes. Figure 6 ranks the most frequent normalized models and annotates first-party state at the September 24, 2026 cutoff. GPT-4o appears scientifically in 7,340 papers (41.1% of Scientific Use papers) and GPT-4 in 6,383 (35.8%); these percentages are non-exclusive because papers can use both. Seven models in the top 20 are already retired from their first-party APIs: Claude 3.5 Sonnet (1,377 papers), GPT-4 Turbo Preview (1,258), Gemini 1.5 Pro (982), Gemini 2.0 Flash (807), Claude 3.7 Sonnet (725), `text-davinci-003` (619), and Gemini 1.0 Pro (590) (Anthropic, 2026a; OpenAI, 2026a; Google, 2026b;a). Several other high-frequency models are deprecated or retain an active lineage while historically relevant snapshots are on explicit shutdown paths.

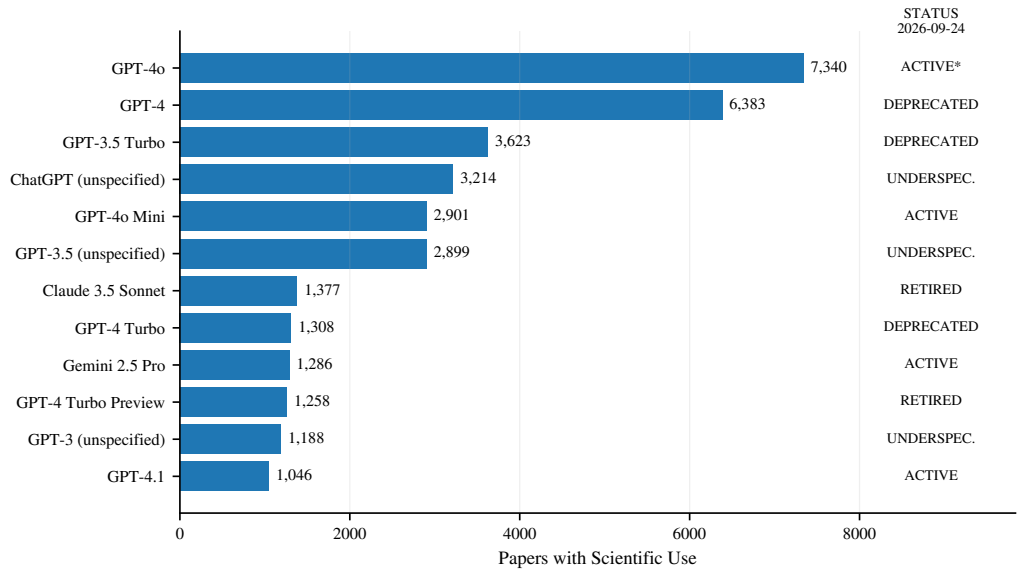

Figure 6: Scientific footprint and lifecycle state of the most common normalized closed models.

A normalized model name is not necessarily a preserved experimental artifact. At the cutoff, OpenAI's catalog classifies GPT-4 as deprecated; `gpt-4-0314` has already shut down and `gpt-4-0613` is scheduled for shutdown on October 23, 2026 (OpenAI, 2026a;c). GPT-4o remains available at the cutoff, but `gpt-4o-2024-05-13` is deprecated and scheduled for shutdown on the same date (OpenAI, 2026a;b). The literature-level label "GPT-4o" therefore spans served artifacts with different persistence states; preservation analysis ultimately requires the reported identifier when one exists.

The same problem is visible among evaluation models. GPT-4o is used as a JUDGE in 1,673 papers and GPT-4 in 1,453. Among the ten most frequent normalized judges, only three are cleanly active and specifically identifiable at the cutoff; the remainder are retired, deprecated/partially deprecated, or underspecified (Appendix F). When a baseline disappears, its historical score may remain while the reference condition becomes non-re-executable; when a judge disappears, the published measurement procedure can likewise become non-re-executable for later systems.

The mismatch between publication and service lifecycles is already observable. Claude 3.5 Sonnet, retired from Anthropic's API on October 28, 2025, appears with Scientific Use in 349 papers in the 2026 archival corpus; Gemini 1.5 Pro, shut down September 29, 2025, appears in 185; Gemini 1.0 Pro, no longer supported from February 18, 2025, appears in 78; and `text-davinci-003`, shut down January 4, 2024, appears in 11 (Anthropic, 2026a; Google, 2026a; OpenAI, 2026a). Publication year is not invocation date, so these counts do not imply calls to retired endpoints. They show something narrower: papers can enter the durable archival record after first-party access to a named dependency has already ended.

Finally, lifecycle loss is distinct from identity loss. Frequent reported strings include broad labels such as `GPT-4o` (6,923 papers), `GPT-4` (5,930), and `ChatGPT` (3,183), alongside pinned forms such as `gpt-4o-2024-08-06` (640) and `gpt-4-0613` (480). Because a paper can contain several forms, these are non-exclusive per-identifier paper counts rather than a paper-level reporting-quality rate. Even continued access to a model lineage cannot reconstruct the original experiment if the paper never fixed which served artifact produced the reported result.

5 DISCUSSION

The results separate three preservation risks in the research evidence base. Closed-model dependence is already common enough to be a property of the research record rather than a small set of case studies; it increasingly enters data construction and evaluation rather than only comparison tables; and a large scientific footprint does not imply persistence. Several of the most frequently used model

lineages have already lost first-party availability, while others are deprecated or identified only at a coarse product/family level. Table 3 maps those observations to concrete intervention points.

Table 3: Preservation interventions suggested by the observed dependency structure.

Intervention	Proposed norm	Empirical motivation
Authors / artifacts	Prefer open-weight or locally retainable models when they can serve the same role. If closed access is necessary, report exact model/snapshot, access surface/date, material settings, and archivable inputs/outputs.	Broad product/family identifiers coexist with pinned snapshots; identity loss and access loss are separate failure modes.
Peer review	Prefer preservable references for acceptance-critical evidence; require closed-model comparisons only when the claim depends on that model or capability regime.	Baselines are the most common role, but 27.90% of the corpus has an active non-baseline dependency; review can otherwise add non-preservable dependencies for generic completeness.
Providers / archives	Preserve high-footprint models via low-throughput research access, trusted escrow, retained execution traces, or sunset weight release.	Seven of the top 20 normalized models are retired; model and Judge prevalence identify artifacts with unusually large scientific footprints.

When an open-weight or locally retainable model can serve the same role, it should be preferred: preservation then comes by construction. Closed access remains appropriate when the scientific question requires a specific proprietary system or capability regime. For closed judges and data generators, archiving material inputs and outputs can retain audit evidence when re-execution becomes impossible, subject to privacy, copyright, and contractual constraints (Smith et al., 2016; Barker et al., 2022).

Scientific preservation need not require indefinite production serving. Johnson & Obradovich (2024) discuss continued research access and archival mechanisms; Anthropic has retained weights and demonstrated restricted post-retirement access for Claude Opus 3 (Anthropic, 2025; 2026b). Mechanisms can vary with safety, privacy, licensing, and legal constraints. Paper count and scientific role provide practical signals for prioritizing artifacts with broad footprints in the literature.

6 Limitations and Scope

Measurement at this scale requires explicit boundaries. We fix venues, model registry, and scientific roles, count only dependencies evidenced within each paper, and do not recursively propagate dependencies through reused datasets, benchmarks, or prior work. This makes a diffuse, fast-moving infrastructure dependency countable and auditable. The percentages are best read for scale, temporal direction, role/track structure, and lifecycle exposure–not as universal constants or a paper-level "irreproducibility" rate; lifecycle labels are cutoff- and access-surface-specific.

That pragmatism is deliberate: across 53,538 papers, finite coverage and within-paper attribution provide ecosystem-level visibility, while blinded validation checks the instrument (Scientific Use F1=0.967; role micro-F1=0.888). The $n = 200$ validation sample is not a per-model or per-stratum error guarantee, and the estimates do not imply that dependencies outside the registry or inherited indirectly through reused artifacts are absent. These boundaries define the estimates; the aim is to make a deeply embedded, fast-changing, otherwise intangible dependency visible to the broader AI research ecosystem.

7 Conclusion

Across 53,538 archival papers, 33.32% contain measured scientific dependence on a covered closed model, and that dependence increasingly enters data and evaluation infrastructure while outliving model service lifetimes. A scholarly record that preserves scores after their computational anchors disappear is incomplete: epistemic durability requires preserving the identity and continued interrogability of those anchors.

AI USE STATEMENT

OpenAI GPT-5.6 Sol, configured at High reasoning effort, was used under author direction as an agentic coding and research assistant during development of the measurement system. It assisted with implementing components of the large end-to-end pipeline, diagnosing failures and inconsistencies, and exploratory analysis and interpretation used to set up and finalize the pipeline. The authors specified the research questions and methodological choices, reviewed proposed changes, and initiated and ran the experiments and full-corpus pipeline.

A distinct use of generative AI is part of the scientific method itself: OpenAI `gpt-oss-120b` is the production adjudicator described in section 3. It was served locally with reasoning effort set to `low` and run on bounded evidence for unresolved model entities under the adjudication and reconciliation constraints reported in the paper; it could not overwrite deterministic decisions. Human reference labels used for validation were assigned by the authors.

GPT-5.6 Sol, again at High reasoning effort, was also used agentically to assist with polishing author-written manuscript drafts, identifying related work, checking citations against primary sources, and preparing and refining the LaTeX manuscript. The authors verified the final manuscript against the frozen data and analysis outputs, original author drafts, primary sources, code and tests, and human validation records. The authors take full responsibility for the methodological choices, experiments, claims, code, citations, and final manuscript content.

REPRODUCIBILITY STATEMENT

Code, processed study data, frozen analysis outputs, validation labels, stage launchers, and exact environment specifications are provided with the submission. Together these artifacts are sufficient to reproduce all reported statistics, tables, figures, and validation results from the provided processed data, while also documenting the complete measurement pipeline and its stage contracts. Appendix A gives the artifact organization, execution order, host and environment details, fixed seeds, and stage-specific serving/configuration information. The archival PDFs that make up the study corpus are not redistributed; corpus membership definitions and acquisition code are provided, and rerunning acquisition or model-serving stages is not required to regenerate the reported results.

REFERENCES

Anthropic. Commitments on model deprecation and preservation. `https://www.anthropic.com/research/deprecation-commitments`, November 2025. Accessed 2026-09-24.

Anthropic. Model deprecations - Claude Platform Docs. `https://platform.claude.com/docs/en/about-claude/model-deprecations`, 2026a. Accessed 2026-09-24.

Anthropic. An update on our model deprecation commitments for Claude Opus 3. `https://www.anthropic.com/research/deprecation-updates-opus-3`, February 2026b. Accessed 2026-09-24.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.5. URL `https://aclanthology.org/2024.eacl-long.5/`.

Michelle Barker, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, and Tom Honeyman. Introducing the FAIR principles for research software. *Scientific Data*, 9:622, 2022. doi: 10.1038/s41597-022-01710-x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019.

Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Roberto Di Cosmo and Stefano Zacchiroli. Software heritage: Why and how to preserve software source code. In *iPRES 2017: 14th International Conference on Digital Preservation*, pp. 1–10, Kyoto, Japan, September 2017. URL https://www.digipres.org/publications/ipres/ipres-2017/papers/software-heritage-why-and-how-to-preserve-software-source-code/.

Google. Release notes | Gemini API | Google AI for Developers. https://ai.google.dev/gemini-api/docs/changelog, 2026a. Accessed 2026-09-24.

Google. Gemini deprecations | Gemini API | Google AI for Developers. https://ai.google.dev/gemini-api/docs/deprecations, 2026b. Accessed 2026-09-24.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1644–1651, 2018. doi: 10.1609/aaai.v32i1.11503.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.306. URL https://aclanthology.org/2025.findings-acl.306/.

Tim Johnson and Nick Obradovich. New methods for deprecating artificial intelligence systems will preserve history and facilitate research. *Nature Communications*, 15:10254, 2024. doi: 10.1038/s41467-024-54758-1.

Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover. Scholarly context adrift: Three out of four URI references lead to changed content. *PLOS ONE*, 11(12):e0167475, 2016. doi: 10.1371/journal.pone.0167475.

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. Scholarly context not found: One in five articles suffers from reference rot. *PLOS ONE*, 9(12):e115253, 2014. doi: 10.1371/journal.pone.0115253.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. doi: 10.48550/arXiv.1907.11692. URL https://arxiv.org/abs/1907.11692.

Erlis Lushtaku, Bora Kargi, Ali Elganzory, Fabio Ferreira, Alejandro R. Salamanca, Julia Kreutzer, and David Salinas. JudgeArena: A unified framework for reproducible LLM-judge evaluation. *arXiv preprint arXiv:2608.02620*, 2026. doi: 10.48550/arXiv.2608.02620. URL https://arxiv.org/abs/2608.02620.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. Deprecations | OpenAI API. https://developers.openai.com/api/docs/deprecations, 2026a. Accessed 2026-09-24.

OpenAI. GPT-4o model | OpenAI API. https://developers.openai.com/api/docs/models/gpt-4o, 2026b. Accessed 2026-09-24.

OpenAI. All models | OpenAI API. https://developers.openai.com/api/docs/models/all, 2026c. Accessed 2026-09-24.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. Software citation principles. *PeerJ Computer Science*, 2:e86, 2016. doi: 10.7717/peerj-cs.86.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Reproducibility Details

A.1 Provided research artifacts

All reported statistics, tables, and figures can be reproduced from the processed study data and analysis code provided with the submission. The accompanying artifacts also provide the complete measurement pipeline code, frozen analysis outputs and selected pipeline diagnostics, human validation reference labels, stage-specific requirements, and launchers for each pipeline stage.

The end-to-end pipeline runs in the documented order: corpus acquisition, PDF/text processing, candidate retrieval, deterministic evidence classification, LLM adjudication of unresolved entities, deterministic reconciliation, validation, and read-only analysis. The launchers preserve stage boundaries and required inputs; the validation sample uses the fixed seed `20260829`, and adjudication follows the request-level seed and retry contract reported in Appendix D. The archival PDFs that make up the study corpus are not redistributed; corpus membership definitions and acquisition code are provided, while the provided processed data and analysis outputs reproduce the reported results without rerunning acquisition or model-serving stages.

A.2 Execution environment

The recorded execution host was Microsoft Windows 11 Pro 25H2 (build 26200) with two NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs. GROBID and the local vLLM services used for OCR and `gpt-oss-120b` adjudication ran as containerized services with fixed versions. The provided launchers retain the corresponding service versions, serving arguments, and stage commands; the scientific serving configuration is described in Appendix D.

The common Python environment used across the pipeline had the package versions in Table 4. Containerized services and any stage-specific requirements are specified separately with the corresponding pipeline modules.

Table 4: Python packages in the common execution environment.

Package	Version	Package	Version
certifi	2026.7.22	charset-normalizer	3.5.1
colorama	0.4.6	contourpy	1.3.3
cycler	0.12.1	fonttools	4.63.0
idna	3.19	iniconfig	2.3.0
kiwisolver	1.5.1	lxml	6.1.2
matplotlib	3.11.1	numpy	2.5.2
packaging	26.3	pillow	12.3.0
pip	24.3.1	pluggy	1.6.0
Pygments	2.21.0	pymupdf	1.28.2
pyparsing	3.3.2	pytest	8.4.2
python-dateutil	2.9.0.post0	requests	2.34.2
six	1.17.0	tqdm	4.70.0
urllib3	2.7.0		

B Corpus and Archival Scope

B.1 Population definition

The study population is a frozen census of final archival proceedings membership, not a sample discovered by keyword search. The scope reference date is August 22, 2026. We include ACL 2023–2026 (Main Long, Main Short, Findings), EMNLP 2023–2025 (Main, Findings), ICLR 2023–2026, ICML 2023–2025 (Main plus archival Position papers where present), and NeurIPS 2023–2025 (Main, Datasets & Benchmarks, plus the 2025 Position track). This yields 53,538 papers in 34 prespecified venue/year/track strata. The exact frozen membership is reported in Table 5.

Table 5: Frozen archival corpus membership (34 venue/year/track strata; total 53,538 papers).

Venue	Year	Track	Papers
ACL	2023	Findings	901
ACL	2023	Main Long	910
ACL	2023	Main Short	164
ACL	2024	Findings	975
ACL	2024	Main Long	864
ACL	2024	Main Short	76
ACL	2025	Findings	1,387
ACL	2025	Main Long	1,602
ACL	2025	Main Short	97
ACL	2026	Findings	2,163
ACL	2026	Main Long	2,222
ACL	2026	Main Short	74
EMNLP	2023	Findings	1,059
EMNLP	2023	Main	1,047
EMNLP	2024	Findings	1,003
EMNLP	2024	Main	1,268
EMNLP	2025	Findings	1,405
EMNLP	2025	Main	1,809
ICLR	2023	Conference	1,574
ICLR	2024	Conference	2,260
ICLR	2025	Conference	3,703
ICLR	2026	Conference	5,351
ICML	2023	Conference	1,828
ICML	2024	Conference	2,535
ICML	2024	Position	75
ICML	2025	Conference	3,257
ICML	2025	Position	73
NeurIPS	2023	Datasets & Benchmarks	322
NeurIPS	2023	Main	3,218
NeurIPS	2024	Datasets & Benchmarks	459
NeurIPS	2024	Main	4,034
NeurIPS	2025	Datasets & Benchmarks	497
NeurIPS	2025	Main	5,286
NeurIPS	2025	Position	40

B.2 Denominator reconciliation

Our scientific denominator is the final archival record. Notification-time acceptance counts, conference-program pages, and final proceedings can differ because of withdrawals, non-proceedings program items, late camera-ready changes, or inconsistent first-party reporting. We therefore retain the archival count and audit the discrepancy separately rather than forcing the archive to match a headline acceptance number. Table 6 lists the cases requiring explicit reconciliation.

Table 6: Selected archival-versus-program/acceptance reconciliations. A causal explanation is assigned only when a first-party record supports it numerically.

Venue-year	Archive	Comparator	Δ	Audit state	Reconciliation note
ICLR 2023	1,574	1,574	0	Match	1,584 virtual-program posters include 10 Journal-to-Conference showcases outside the proceedings scope.
EMNLP 2024	2,271	2,300	-29	Explained	Official proceedings report 3 Main and 26 Findings withdrawals after acceptance.
ICLR 2025	3,703	3,704	-1	Unresolved	Final proceedings contain one fewer paper; no exact first-party cause was established.
ICLR 2026	5,351	5,355 / 5,357	-4 / -6	Source conflict	Two first-party ICLR sources report different accepted totals; neither is promoted to ground truth.
ICML 2025	3,330	3,333	-3	Unresolved	The delta is entirely in Main; the final PMLR count is used.
NeurIPS 2024	4,493	4,497	-4	Unresolved	Final proceedings are lower by 3 Main and 1 Datasets & Benchmarks paper.

The downloader freezes source provenance and expected membership before PDF processing. Each paper receives a stable venue/year/track/paper identifier, and downstream stages reconstruct the denominator from that manifest rather than from the candidate set. Acquisition checks found all 34/34 strata complete, all 53,538/53,538 expected PDFs present, and zero hard cross-stratum identity-collision groups. Thus later screening can reduce the number of papers requiring semantic processing without changing the study population.

C Covered Models and Identity Resolution

C.1 Registry scope

The registry is a prespecified, high-recall operational coverage set for major proprietary closed-model families rather than an attempt to define every closed model in existence. It contains 21 retrieval entities mapping to 193 finite normalized model keys across OpenAI, Anthropic, Google, Amazon, Microsoft, and AI21 families. Broad retrieval families make lexical search robust to aliases and formatting variants, while the finite normalized vocabulary prevents the analysis from inventing new semantic model identities from arbitrary strings encountered in papers.

Table 7: Retrieval families in the closed-model registry. The 21 retrieval entities map to 193 finite normalized model keys.

Provider	Retrieval family	Registry ID	Model keys
OpenAI	ChatGPT	`openai_chatgpt`	1
OpenAI	GPT-3 family	`openai_gpt_3_family`	24
OpenAI	GPT-4 family	`openai_gpt_4_family`	20
OpenAI	GPT-5 family	`openai_gpt_5_family`	30
OpenAI	OpenAI o-series	`openai_o_series`	11
Anthropic	Claude 1 family	`anthropic_claude_1_family`	10
Anthropic	Claude 2 family	`anthropic_claude_2_family`	4
Anthropic	Claude 3 family	`anthropic_claude_3_family`	8
Anthropic	Claude 4 family	`anthropic_claude_4_family`	13
Anthropic	Claude 5 family	`anthropic_claude_5_family`	6
Anthropic	Claude (unspecified generation)	`anthropic_claude_generic`	1
Google	Gemini 1 family	`google_gemini_1_family`	10
Google	Gemini 2 family	`google_gemini_2_family`	14
Google	Gemini 3 family	`google_gemini_3_family`	16
Google	Gemini (unspecified generation)	`google_gemini_generic`	1
Google	PaLM 2	`google_palm_2`	8
Google	PaLM	`google_palm_generic`	1
Google	Bard	`google_bard`	1
Amazon	Amazon Titan Text	`amazon_titan_text`	5

Provider	Retrieval family	Registry ID	Model keys
Microsoft	Bing Chat / Microsoft Copilot / GitHub Copilot	`microsoft_bing_chat`	4
AI21 Labs	Jurassic / Jurassic-2	`ai21_jurassic`	5

The primary registry is frozen for the study. A separate diagnostic watch surface can flag model-like terms outside the measured registry for manual inspection, but those terms do not create candidates or labels and cannot silently expand the estimand during processing.

C.2 Three identity levels

Every retained occurrence is represented at three levels:

1. **Retrieval family**: the broad lexical envelope used to retrieve evidence (for example, the GPT-4 generation family).
2. **Normalized model**: the finite semantic key used for aggregation (for example, `gpt-4o`).
3. **Reported identifier**: the exact identifier written by the paper when recoverable (for example, `gpt-4o-2024-08-06`).

Unknown, malformed, or coarse forms are never converted into newly invented model keys; where appropriate they map to a family-specific unspecified key. Reported identifiers are retained even when several identifiers map to the same normalized model because snapshots, revisions, and mutable aliases can have different lifecycle properties. This separation is also what permits the lifecycle analysis to distinguish model availability from artifact identifiability.

D Measurement Pipeline

The measurement system is staged so that high-recall retrieval is separated from role classification, deterministic decisions are not overwritten by later semantic inference, and the final prevalence denominator remains the frozen proceedings census. Table 8 summarizes the realized pipeline counts; the following subsections describe the scientific contract of each stage and its output.

Table 8: Realized pipeline counts by stage. Counts at different entity levels are not additive: one paper can contain multiple retrieval families and both deterministically resolved and LLM-adjudicated entities.

Stage	Component	Realized output
1	Corpus acquisition	34/34 strata frozen; 53,538/53,538 PDFs present; 0 hard cross-stratum identity-collision groups.
2	PDF/text processing	53,538 final successful parses: 53,528 standard structured parses and 10 OCR recoveries; 0 final parse failures.
3	Candidate retrieval	21,587 candidate papers (40.32%); 31,951 screen-negative papers retained in the denominator.
4	Deterministic triage	12,992 reconciled paper-family rows retained deterministic decisions across 9,809 papers.
5	LLM adjudication	35,948 paper-family rows across 17,930 ambiguous papers; accepted attempts 17,230/688/12 on attempts 1/2/3; 0 final run errors.
6	Reconciliation	48,940 paper-family rows and 71,413 paper-model rows frozen; 0 unresolved rows under –require-complete.
7	Validation sampling	Simple random sample n=200 from all 53,538 papers, seed 20260829; 200/200 human labels resolved.
8	Analysis reconstruction	53,538 population rows rebuilt from the manifest; 17,841 Scientific Use papers and 2,486 Supporting Use papers in the frozen analysis.

D.1 Stage 1: corpus acquisition and manifest freeze

Stage 1 instantiates the population defined in Appendix B; it does not discover papers through model-related search terms. Venue-specific downloaders enumerate the prespecified proceedings strata, record source evidence, assign stable paper identifiers, and verify the acquired PDFs against the frozen membership ledger. This separation is important because every later prevalence statistic is defined over the manifest, including papers that never become lexical candidates.

Realized output. Acquisition checks verified 34/34 primary strata, 53,538/53,538 expected final PDFs, and zero hard cross-stratum identity-collision groups. The signed acceptance/program discrepancies remain an audit layer (Table 6) and do not alter the proceedings denominator.

D.2 Stage 2: PDF processing and text recovery

Each paper is processed through two independent text channels. The structured channel uses GROBID TEI converted to canonical Markdown, with citation/header consolidation disabled. The raw channel uses independent PDF text extraction and is retained as a recall surface for parser-loss detection. Canonical Markdown shorter than 500 characters is treated as a failed structured parse. Reference regions are represented as spans rather than by truncating everything after the first "References" heading, allowing appendices or supplementary sections after a bibliography to re-enter admissible body text.

Failed structured parses are retried before OCR recovery. OCR-recovered papers are explicitly marked and do not receive fabricated TEI. The recorded run used GROBID 0.9.1-crf for the primary path; the recovery path used Unlimited-OCR through a local OpenAI-compatible server, in four-page requests with an 8,192-token output ceiling. These serving details affect extraction and are therefore part of the processor configuration.

Realized output. All 53,538 papers have final successful processor records. Of these, 53,528 completed through the standard structured path and 10 required Unlimited-OCR recovery; no parse failures remain in the frozen manifest. OCR recovery therefore affects a small but explicit subset rather than being silently mixed with ordinary parses.

D.3 Stage 3: high-recall candidate retrieval

Candidate marking asks only whether a paper should proceed to closed-model use classification. Text is normalized using Unicode NFKC normalization, dash and whitespace normalization, soft-hyphen and invisible-format removal, and separator-tolerant model matching to recover common PDF artifacts such as line-broken or character-spaced names. Generation-family overlap suppression prevents a more generic alias from double-promoting the same occurrence when a more specific family match is present. Release-plausibility diagnostics, including apparent model mentions predating public release, were used during retrieval-rule hardening to locate lexical failure modes such as table-column seams, adjacent version numbers, and short-token collisions; publication year is not used as a retrieval or classification feature.

Bibliography-only structured matches do not promote a paper. Raw text is not an unrestricted second search surface: a raw occurrence can rescue candidacy only under the parser-loss policy, such as when the entity is absent from canonical Markdown or appears there only inside a detected reference span while raw text contains an admissible pre-reference occurrence. Every hit retains the retrieval family, normalized model key, matched variant, reported identifier where recoverable, source channel, and whether that hit was candidate-promoting.

Realized output. The screen promoted 21,587 of 53,538 papers (40.32%) and left 31,951 as explicit screen negatives. Candidate status is only a resource-allocation variable: all 53,538 population rows return at analysis time. The completed human validation later measured candidate-screen recall directly (Appendix E).

D.4 Stage 4: deterministic evidence classification

Candidate evidence is evaluated at paper $\times$ retrieval-family level while retaining normalized-model identity on every evidence item. Evidence windows record source channel, section, matched variant, reported identifier when available, and separate spans for broad use scoring and tighter role attribution. The classifier returns `USE`, `NO_USE`, or `AMBIGUOUS`; these states allocate downstream adjudication and are not the paper's scientific ontology.

The classifier combines section priors with interpretable pattern classes for current-author operational language, API/configuration terms, prior work, generic mentions, inherited artifacts, and the five role

families (four Scientific Use roles plus Supporting Use). The principal score contributions are shown in Table 9. Markdown evidence requires a score of at least 5; raw-only evidence requires 7.

Table 9: Principal deterministic evidence contributions. Exact regular expressions, ordering, and rule contracts are preserved in the frozen classifier source.

Evidence type	Score	Evidence type	Score
Current-author operational use	+5	Imported baseline construction	+9
Judge / DataGen / Direct / Supporting cue	+5	Baseline cue	+4
Operational configuration	+2	High-use section	+1
Prior-work / inherited-artifact cue	-4	Generic-mention cue	-2
Low-use section	-2	Raw-only evidence penalty	-1

A deterministic `USE` additionally requires locally compatible role evidence. Prior-work/background role vocabulary cannot donate a role to current-paper use, and inherited data does not become current-paper Data Generation without an explicit current generation action. Ambiguity is intentional when use is apparent but no role is safely established, when multiple normalized models in one operational family could inherit the same nearby role evidence, or when Supporting Use and a scientific role cannot be separated cleanly. `NO_USE` requires sufficiently strong prior/background/inherited evidence without current operational evidence.

Realized output. In the final reconciled family-level dataset, 12,992 paper-family rows retained deterministic decisions across 9,809 papers; 10,229 of these rows were USE decisions, spanning 7,980 papers with at least one deterministically resolved used family. These paper counts overlap with Stage 5 because one paper can contain both resolved and ambiguous families.

D.5 Stage 5: evidence-constrained LLM adjudication

Only unresolved entities enter semantic adjudication. All ambiguous entities from one paper are packed into one initial request, so the invocation unit is an ambiguous paper rather than an occurrence. The local `gpt-oss-120b` adjudicator was served through vLLM with reasoning effort set to `low`; initial requests used temperature 0, seed 42, structured JSON-schema output, and a 4,096-token output ceiling. Technical recovery was capped at 16,384 output tokens and three semantic attempts; malformed or schema-invalid responses use explicitly seeded small temperature perturbations, while transport or length failures preserve the preceding decoding configuration (with a larger output budget after truncation). The prompt ceiling was 45,000 characters, and the adjudication client used 64 concurrent workers.

The adjudicator never receives the full paper, deterministic score, or provisional classifier answer. It receives retained evidence under two distinct channels:

- **LOCAL**: tight model-local evidence. Only LOCAL evidence can authorize a scientific or supporting role.
- **CONTEXT**: broader adjacent evidence that can help distinguish operational use from surrounding discourse but cannot donate a role.

Evidence IDs are bound to exact normalized model keys in the output schema, preventing evidence retained for one model from authorizing a role for another model in the same retrieval family. The classifier supplies up to 10,000 evidence characters per paper packet; LOCAL evidence has priority over CONTEXT. If compression is required, CONTEXT is removed first. An irreducibly oversized LOCAL packet fails rather than silently dropping role-authorizing evidence.

A valid first generation is accepted and is not resampled for consensus. Retry attempts are technical recovery for request errors, truncation, invalid JSON, or schema/contract failure; they are not independent semantic votes. Exhaustion produces an error and causes the require-complete downstream build to fail. The deployed GPT-OSS MXFP4/MARLIN path did not support batch-invariant execution, so the fixed seed is a request-level reproducibility setting rather than a claim of scheduling-independent bitwise identical generations. Raw responses and request/retry metadata are retained for the adjudicated entities.

Realized output. Stage 5 adjudicated 17,930 ambiguous papers containing 35,948 paper-family rows. Of those papers, 17,230 resolved on the first accepted semantic attempt, 688 on the second, and 12 on the third; all 17,930 ultimately resolved and no final LLM run errors remained. The reconciled LLM-resolved subset contains 28,739 used family rows across 15,496 papers with at least one LLM-resolved used family.

D.6 Stage 6: deterministic reconciliation and label freeze

Reconciliation is asymmetric by construction. Deterministically resolved entities remain fixed; LLM output is accepted only for entities that were explicitly queued as ambiguous. No human correction file was supplied to reconciliation. A `--require-complete` condition prevents final label output if any requested entity remains unresolved.

The reconciler produces paper-family labels, normalized-model labels, and paper-level role labels. Family roles are derived from model-level evidence-bound decisions rather than guessed independently. The candidate file is not treated as the corpus: screen-negative papers are reintroduced from the population manifest in Stage 8.

Realized output. Reconciliation froze 48,940 paper-family rows (12,992 deterministic and 35,948 LLM-resolved) and 71,413 paper-model rows, with zero unresolved rows. No manual correction layer contributed to the final label dataset.

D.7 Stage 7: blinded validation sampling

Validation sampling is performed after label freeze but samples from the complete processor population rather than from candidates or predicted positives. The sampler draws a simple random sample without replacement, hides system predictions from the human-facing annotation sheets, and retains candidate/pipeline diagnostics separately from the annotation sheets. This makes the validation end to end: extraction, candidate screening, and downstream classification are evaluated jointly rather than conditionally.

Realized output. The prespecified sample contains 200 papers selected from all 53,538 population papers with seed 20260829. All 200 sampled paper keys and sample IDs were unique. The completed human scoring outcome is reported separately in Appendix E.

D.8 Stage 8: full-population reconstruction and analysis

The analysis stage is read-only with respect to upstream labels. It reconstructs all 53,538 papers from the frozen processor manifest and left-joins reconciled family/model/paper labels. Papers that never passed candidate retrieval therefore remain explicit zero-labeled population rows rather than disappearing from the denominator. Scientific Use and Supporting Use are analyzed independently; provider/model/model-family percentages are non-exclusive because one paper can use multiple entities. Exact reported identifiers remain separate from normalized semantic models and are not converted into lifecycle claims by string heuristics.

Realized output. Analysis reconstructed 53,538 population rows and produced 17,841 Scientific Use papers and 2,486 Supporting Use papers; 886 papers had Supporting Use without Scientific Use. A combined-use diagnostic contains 18,727 papers, but the primary analysis keeps Scientific Use and Supporting Use separate. The analysis also produces venue/year/track, role, model, family, provider, multiplicity, and reported-identifier tables.

Implementation checks. The implementation passes 302 unit tests (116 in corpus acquisition and 186 across the remaining modules). These tests check implementation contracts and edge cases; they are not a substitute for human label validation, which is reported next.

E HUMAN VALIDATION

E.1 ANNOTATION PROTOCOL

The validation sample is the fixed Stage-7 simple random sample of $n = 200$ papers from the full population. Sampling does not condition on candidate status, deterministic/LLM decision source, OCR status, or system prediction. A single human annotator reviewed blinded sheets without system predictions and assigned all four scientific roles and Supporting Use; the primary Scientific Use endpoint is their union.

For the completed human reference set we report

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2PR}{P + R}.$$

The corpus prevalence values are descriptive quantities for the frozen archival census; uncertainty intervals are used for validation proportions rather than attached to the fact that a fixed number of rows in the census received each label.

E.2 COMPLETED VALIDATION RUN

All 200 sampled papers received complete final annotations, with zero unresolved rows and zero annotation-problem rows. The final human reference set contains 76 Scientific Use positives and 124 negatives. The end-to-end confusion matrix is 74 true positives, 121 true negatives, 3 false positives, and 2 false negatives.

Table 10: End-to-end validation of the primary paper-level Scientific Use label. Intervals are 95% Wilson intervals for binomial proportions; F1 is reported as a point estimate.

Metric	Estimate	95% CI
Precision	0.961	[0.892, 0.987]
Recall	0.974	[0.909, 0.993]
F1	0.967	–
Accuracy	0.975	[0.943, 0.989]
Specificity	0.976	[0.931, 0.992]

E.3 ERROR LOCALIZATION AND VALIDATION SCOPE

Candidate screening recovered all 76 reference-positive papers, for recall 1.000 [0.952, 1.000]. Both end-to-end false negatives therefore occur downstream of retrieval; all five binary disagreements are among papers that entered the candidate set. The validation sample contains 76/200 reference positives (38.0%; 95% Wilson interval 31.6–44.9%). This interval contains the measured corpus prevalence of 33.32%, but the sample is used to validate the instrument rather than to recalibrate the archival census.

The final annotations also validate the four non-exclusive role labels directly. Multilabel exact-match accuracy is 0.890, with micro-F1 0.888 and macro-F1 0.884. Per-role performance is shown in Table 11.

Table 11: Role-level validation on the same 200-paper sample.

Role	Gold +	Precision	Recall	F1
Direct	32	0.906	0.906	0.906
Baseline	48	0.855	0.979	0.913
Judge	29	0.923	0.828	0.873
Data Generation	27	0.730	1.000	0.844

The sample now supports both the primary paper-level endpoint and direct role-level checks, but $n = 200$ still limits precise guarantees for individual venue/year strata or model identities. Machine-

readable final labels, per-role metrics, and disagreement records accompany the provided research artifacts.

F LIFECYCLE CURATION

F.1 EVIDENCE SOURCE AND STATUS TAXONOMY

Lifecycle status is curated independently of the usage classifier. The classifier asks what scientific role a paper assigns to a model; lifecycle curation asks whether the referenced model remains identifiable and accessible at a specified later date. Neither source is allowed to infer the other.

For the model-level analysis, lifecycle state was manually curated from first-party provider documentation with a cutoff of September 24, 2026. We use five display states:

- **A (Active):** the normalized model is specifically identifiable and the provider lists its stable API form without an identified general shutdown at the cutoff.
- **A* (Active with historical snapshot loss):** the normalized lineage remains available, but one or more historically relevant snapshots are deprecated, retired, or on an announced shutdown path. This does not imply that every paper using the normalized model is affected.
- **D (Deprecated):** the provider lists the normalized model/lineage as deprecated or on a shutdown path.
- **R (Retired):** the first-party endpoint/model has been shut down or retired.
- **U (Underspecified):** the literature label is too coarse to identify one served artifact.

OpenAI distinguishes deprecation from shutdown and publishes snapshot-level schedules (OpenAI, 2026a); Google similarly defines shutdown as the endpoint being completely turned off (Google, 2026b); Anthropic distinguishes Active, Legacy, Deprecated, and Retired, with requests to retired models failing (Anthropic, 2026a). Surface matters: retirement from a first-party API does not prove absence from every partner or restricted-access platform. Exact paper-level survivability therefore requires joining provider, normalized model, reported identifier, and access surface to lifecycle records rather than propagating normalized-model status naively.

F.2 LIFECYCLE STATE OF THE MOST COMMON MODELS

Table 12: Top 20 normalized models by number of papers with Scientific Use, with lifecycle classification at the 2026-09-24 cutoff. Counts are non-exclusive across models. "Active*" means the normalized lineage remains available while at least one historically relevant snapshot is deprecated or scheduled for shutdown.

Rank	Model	Status	Papers	Baseline	Judge	Lifecycle note
1	gpt-4o	A*	7,340	4,268	1,673	GPT-4o remains listed as available, but gpt-4o-2024-05-13 is deprecated and scheduled for shutdown.
2	gpt-4	D	6,383	3,344	1,453	OpenAI model catalog marks GPT-4 deprecated; gpt-4-0314 has shut down and gpt-4-0613 is scheduled for shutdown.
3	gpt-3.5-turbo	D	3,623	1,942	289	OpenAI model catalog marks GPT-3.5 Turbo deprecated; historical snapshots have already shut down and 0125 is scheduled for shutdown.
4	chatgpt-unspecified	U	3,214	1,639	227	ChatGPT is a product-level label and does not uniquely identify a fixed model snapshot.
5	gpt-4o-mini	A	2,901	1,568	584	GPT-4o Mini remains listed as available; no general text-model shutdown was identified at the cutoff.

Rank	Model	Status	Papers	Baseline	Judge	Lifecycle note
6	gpt-3.5	U	2,899	1,740	207	GPT-3.5 is a family/generation label rather than a unique served artifact.
7	claude-3.5-sonnet	R	1,377	953	134	Both Claude 3.5 Sonnet API snapshots were retired.
8	gpt-4-turbo	D	1,308	785	306	OpenAI model catalog marks GPT-4 Turbo deprecated; shutdown is scheduled.
9	gemini-2.5-pro	A	1,286	917	167	Google lists the stable Gemini 2.5 Pro model with no shutdown date announced.
10	gpt-4-turbo-preview	R	1,258	717	223	The gpt-4-turbo-preview alias pointed to gpt-4-0125-preview, which was shut down.
11	gpt-3	U	1,188	683	64	GPT-3 is a generation-level label; original GPT-3 completion/base endpoints were retired, but the label does not uniquely identify an artifact.
12	gpt-4.1	A	1,046	599	251	GPT-4.1 remains listed as an available API model at the cutoff.
13	gpt-5	A*	1,011	602	155	GPT-5 remains listed, but the original gpt-5-2025-08-07 snapshot is deprecated and scheduled for shutdown.
14	gemini-1.5-pro	R	982	757	77	Google release notes report that the Gemini 1.5 Pro API model was shut down on September 29, 2025.
15	gemini-2.5-flash	A	966	662	132	Google lists stable Gemini 2.5 Flash with no shutdown date announced.
16	gemini-2.0-flash	R	807	583	73	Google shut down Gemini 2.0 Flash and Gemini 2.0 Flash 001.
17	claude-3.7-sonnet	R	725	490	73	Claude 3.7 Sonnet was retired from the Claude API.
18	text-davinci-003	R	619	339	36	OpenAI shut down text-davinci-003.
19	gemini-unspecified	U	610	274	72	An unversioned Gemini family/product label does not uniquely identify a served artifact.
20	gemini-1.0-pro	R	590	421	42	Google release notes state that Gemini 1.0 Pro is no longer supported.

F.3 Realized retirement among high-footprint models

For the timing observation in subsection 4.4, Table 13 reports models retired before 2026 that still appear in 2026 archival papers. Publication year is not API invocation date; the table demonstrates a mismatch between publication timing and provider lifecycle rather than proving that the API call occurred after retirement.

Table 13: Models retired before 2026 that still appear in 2026 archival papers.

Model (retired before 2026)	2026 papers	Direct	Judge	Data gen.	Baseline
Claude 3.5 Sonnet	349	86	39	64	241
Gemini 1.5 Pro	185	31	8	25	147
Gemini 1.0 Pro	78	17	7	11	51
text-davinci-003	11	0	0	3	9

Table 14: Selected models already retired from their first-party APIs by the lifecycle cutoff. Role counts are non-exclusive paper counts.

Model	Papers	Base.	Judge	Data gen.	First-party status
Claude 3.5 Sonnet	1,377	953	134	248	Retired 2025-10-28
GPT-4 Turbo Preview	1,258	717	223	300	Retired 2026-03-26
Gemini 1.5 Pro	982	757	77	142	Retired 2025-09-29
Gemini 2.0 Flash	807	583	73	133	Retired 2026-06-01
Claude 3.7 Sonnet	725	490	73	134	Retired 2026-02-19
text-davinci-003	619	339	36	176	Retired 2024-01-04
Gemini 1.0 Pro	590	421	42	88	Retired 2025-02-18

F.4 Lifecycle state of common closed-model judges

Table 15: The ten normalized models most frequently used as closed-model judges. Active* indicates an active lineage with relevant historical snapshot loss.

Rank	Judge model	Judge papers	Status
1	GPT-4o	1,673	A*
2	GPT-4	1,453	D
3	GPT-4o Mini	584	A
4	GPT-4 Turbo	306	D
5	GPT-3.5 Turbo	289	D
6	GPT-4.1	251	A
7	ChatGPT (unspecified)	227	U
8	GPT-4 Turbo Preview	223	R
9	GPT-3.5 (unspecified)	207	U
10	Gemini 2.5 Pro	167	A

G Supplementary Empirical Results

G.1 Scientific-role combinations

Table 16: Most frequent exact scientific-role combinations among Scientific Use papers. Shares use all Scientific Use papers as the denominator.

Role combination	Papers	Share of Scientific Use
Baseline	2,905	16.3%
Direct + Baseline	2,086	11.7%
Data Generation	1,980	11.1%
Baseline + Data Generation	1,855	10.4%
Direct + Baseline + Data Generation	1,754	9.8%
Direct	1,210	6.8%
Baseline + Judge + Data Generation	1,069	6.0%
Direct + Data Generation	964	5.4%
Direct + Baseline + Judge + Data Generation	775	4.3%
Judge	763	4.3%
Baseline + Judge	753	4.2%
Judge + Data Generation	577	3.2%

G.2 Dependency multiplicity

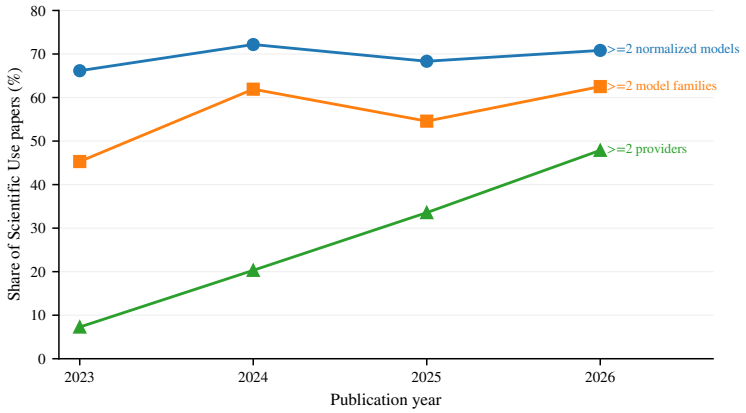

Figure 7: Dependency multiplicity by publication year; 2026 contains ACL and ICLR only.

Across the complete corpus, 69.70% of Scientific Use papers involve at least two normalized covered models, 57.55% involve at least two retrieval families, and 31.76% involve at least two providers. These quantities describe dependency density, not independent failure probability: several normalized models can share infrastructure or lifecycle policies.

G.3 NeurIPS Main and Datasets & Benchmarks

Table 17: NeurIPS Scientific Use, Baseline, and Judge prevalence by track and publication year.

Year	Track	Papers	Scientific use	Baseline	Judge
2023	Datasets & Benchmarks	322	21.4%	13.7%	2.8%
2023	Main	3,218	6.6%	3.6%	0.7%
2024	Datasets & Benchmarks	459	49.5%	35.9%	15.5%
2024	Main	4,034	15.4%	8.0%	3.7%
2025	Datasets & Benchmarks	497	64.0%	50.3%	19.1%
2025	Main	5,286	21.3%	12.6%	6.3%

G.4 Persistence of retired models in the publication stream

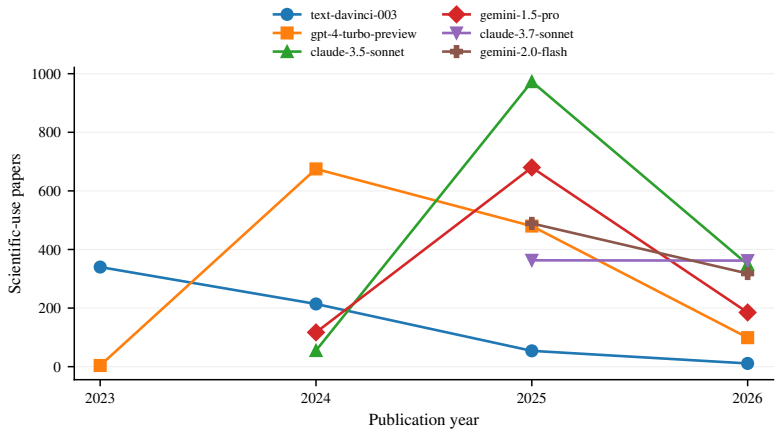

Figure 8: Publication-year counts for selected models retired by the lifecycle cutoff.