Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Model Memorization Thresholds in Citation Generation: Evidence from Field-Level Cloze on Bibliographic Records

Junichiro Niimi

jniimi@meijo-u.ac.jp

Meijo University

Article

Keywords:

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9733633/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. J.N. serves as a technical advisor to a company in the manufacturing sector; this role is unrelated to the models, data, and analysis reported in this study. J.N. declares no other competing interests.

Cross-Model Memorization Thresholds in Citation Generation: Evidence from Field-Level Cloze on Bibliographic Records

Junichiro Niimi[1,*]

[1]Faculty of Business Management, Meijo University, Nagoya, Aichi 468-8502, Japan
[*]jniimi@meijo-u.ac.jp

ABSTRACT

Large language models (LLMs) reproduce some bibliographic records verbatim while fabricating others, with reproduction probability scaling in pretraining frequency and citation count as the standard proxy. A prior study reports GPT-4.1's first-author memorization begins near 90 citations on a model-chosen test set, leaving generalizability open. We sample 1,118 computer science and 1,079 oncology papers from OpenAlex, stratified by citation count, and run a field-level cloze task on eight bibliographic fields with strict binary scoring. Seven conditions are evaluated: three frontier flash-class models (Gemini 3 Flash Preview, GPT-5.4-mini, DeepSeek V4-Flash), the full-tier GPT-5.4, and their reasoning-enabled variants. The field hierarchy replicates across conditions. The first-author inflection spans 87–502 citations in computer science across Gemini, DeepSeek, and full-tier GPT-5.4, with GPT-5.4-mini a further decade higher at ∼1,500, a one-to-two-order-of-magnitude range. The outlier is the *mini* tier, not the GPT family or open/closed status. Reasoning is neutral for Gemini, adverse for DeepSeek, but substantially helpful for GPT-5.4-mini, shifting its curve left by nearly a decade: reaching records direct decoding cannot, not adding knowledge. The same hierarchy, mini outlier, and order-of-magnitude spread recur on oncology at uniformly lower recovery. Memorization thresholds are model-, mode-, and domain-specific, not universal.

Introduction

The fabrication of academic citations by large language models (LLMs) is one of the more visible reliability failures of modern generative models. Empirical estimates put the rate of entirely fabricated references at roughly 55% for GPT-3.5 and 18% for GPT-4[1]; medical-literature audits report ∼47% in short ChatGPT-generated reviews[2]; hallucinated citations now appear in hundreds of accepted papers across major NLP and security venues[3,4]; and even purpose-built legal retrieval-augmented systems hallucinate 17–33% of the time[5,6]. The phenomenon follows from the autoregressive training objective combined with RLHF, which penalizes "I don't know" and rewards confidently plausible output[7–9]. A calibration argument strengthens this: a statistically calibrated generative model must hallucinate "arbitrary" facts at roughly the rate at which they appear once or never in training, independent of architecture or data quality[10], and computability arguments suggest fabrication cannot be eliminated for general LLMs[11].

A theoretical handle on the problem is that the probability with which an LLM reproduces training data scales with that data's frequency in the pretraining corpus[12–14]: extraction probability grows log-linearly with duplication count[13], and downstream factual accuracy scales with the number of relevant pretraining documents[14]. A parallel effect appears in question answering via entity popularity[15]. Within a record, training-dynamics analyses on open models show that nouns, proper names, and numbers are memorized first, hypothesized to act as identifiers indexing the rest of the example[16].

The closest prior methodological work probes an LLM's bibliographic knowledge directly, detecting hallucinations from self-inconsistent answers across independent sessions[17]. We complement that line with a different design—a fixed paper sample stratified by citation count—allowing us to characterize the conditions under which a record is recovered, not only flag those that are not. This raises the question of *how often* a record needs to appear before a model reproduces it from memory rather than reconstructing it from parts. A prior study[18] addressed this in a *closed-system* setting—pretraining-only recall, with retrieval and external tools disabled—using citation count as a proxy for record redundancy across pretraining sources. Three findings emerged, on a single model (GPT-4.1) and a single domain (computer science):

- **Log-linear scaling.** Cosine similarity between generated and authentic bibliographic records scaled approximately log-linearly with citation count ($r = 0.75$ across 100 verified records).
- **Two threshold estimates.** A logistic fit on that single model and domain gave a preliminary inflection point near 90

citations (transition from probabilistic synthesis to memorization-based recall) and a saturation point near 1,200 citations (above which records were reproduced essentially verbatim), both from a model-chosen test set.

- **Hierarchical memorization.** Within a single record, metadata fields crossed the memorization boundary at different citation levels: titles and first-author names earliest, co-authors and venues later, volume and pages later still, and the publication year essentially not at all.

Within the same protocol, a separate line of work[19] showed that, under explicit format constraints (e.g. "recommend only peer-reviewed journal articles"), reasoning-enabled LLMs do not violate the constraint when they lack matching knowledge—they instead *distort* a known record (e.g. converting a NeurIPS paper into a fabricated IEEE Transactions entry) so that the output looks compliant. This is a more dangerous failure mode than wholesale fabrication: the distorted record retains correct authors and topic and only misleads on venue, so it passes the cheap checks (author search, topical plausibility) that readers apply first, and catching it requires field-by-field cross-referencing.

Together these results establish that citation hallucination is shaped by a redundancy–memorization mechanism whose parameters can be estimated. Two questions inherited from that prior work motivate the present study. First, are these threshold estimates universal or model-specific? The two figures came from one model on one domain; if a single citation threshold of that order generalizes, downstream tools (confidence calibrators, hallucination flaggers) can use it as a fixed constant, otherwise every deployment needs its own calibration. Second, how robust are the field hierarchy and threshold estimates to test-set construction? The original thresholds came from open-ended recommendation prompts ("suggest papers about *X*"), so the model itself chose what to surface—a strong selection bias that makes the test set non-comparable across models.

This study addresses both through a change of methodology. Papers are sampled directly from OpenAlex with reproducible random sampling stratified into half-decade citation bands, so the test set is fixed rather than model-chosen. Within each paper, one bibliographic field is masked at a time and the model recovers it given the rest of the record plus the abstract, isolating per-field memorization while holding paper content constant. The same 1,118-paper sample (200 per stratum; the top is pool-limited to 118) is run through seven conditions: three frontier flash-class models from three providers (Gemini 3 Flash Preview, GPT-5.4-mini, DeepSeek V4-Flash), the full-tier GPT-5.4 as a size-tier control, and reasoning-enabled variants of Gemini, DeepSeek, and GPT-5.4-mini. The hierarchical ordering of which fields memorize first replicates across all conditions, but the citation level at which each field crosses 50% recall differs by one to two orders of magnitude. Across Gemini, DeepSeek, and full-tier GPT-5.4 the first-author inflection ranges from 87 to 502 citations—a factor-of-six spread—while GPT-5.4-mini's lies a further decade higher at 1,546. The outlier is the *mini* size tier specifically: full-tier GPT-5.4 on the identical sample falls among Gemini and DeepSeek, so neither open-weight status nor provider explains the gap. Reasoning's effect is heterogeneous: nearly null for Gemini, mostly adverse for DeepSeek, but strongly helpful for GPT-5.4-mini (dropping the `title` inflection from 1,588 to 484 citations and shifting `co_authors`/`pages` by over a decade). The reading is that reasoning restructures how a model resolves an answer rather than expanding what it knows. The practical implication is that a single fixed citation threshold for trusting LLM-generated bibliographic records will be miscalibrated for at least one widely deployed frontier model, and depends on whether thinking is enabled; the per-condition tables we provide can serve as a starting point for model- and mode-specific calibration.

Results

Each condition produced 7,367 cloze trials over the 1,118-paper sample (200 papers per half-decade citation stratum; the top stratum, $\geq$ 3,163 citations, was pool-limited to 118). Of those, 1,118 trials each cover `first_author` and `year` (populated for every paper) and 1,116 cover `title` (two records have an empty title in OpenAlex); the remaining fields have fewer, ranging from 581 (`issue`) to 1,031 (`co_authors`), under the variable mask (see Methods). API failures were rare across all seven conditions—zero to six errored rows per condition, never more than $< 0.1\%$—and errored rows are scored 0 and excluded from the inference statistics below. The Gemini reasoning condition was run via the provider's Batch API rather than synchronously (the synchronous endpoint's daily quota is too small for a 7,367-call sweep at the low tier; see Methods); all 7,367 responses returned with no missing items. The GPT-5.4-mini reasoning condition was split into two non-overlapping halves of the deterministic sample and merged.

Citation gradient and field hierarchy replicate across models

Table 1 reports the per-(tier, field) accuracy for the four non-reasoning conditions, with the six half-decade strata collapsed into three coarse tiers (low: 10–100 citations; mid: 101–1,000; high: $>$ 1,000) for compactness; the full-resolution view is in Figs. 1–2. Three patterns hold across all four conditions. First, accuracy increases monotonically with citation tier for the information-rich fields (`title`, `first_author`, `co_authors`, `pages`). Second, the low-cardinality fields (`year`, `venue`, `volume`, `issue`) plateau by the mid tier; their mid→high transitions are mostly not significant under Bonferroni-corrected pairwise proportion tests, while every information-rich field's transition is. Third, within each tier the field-level

ranking respects the "title and first-author memorize first; co-authors and pages later" order originally reported for GPT-4.1[18]—consistent with the training-dynamics observation that LMs memorize nouns, proper names, and numbers earliest as unique identifiers[16], and with independent audits of ChatGPT-generated medical citations that find volume, pages, and year to be the least-reliable bibliographic fields[2]. The hierarchical structure of memorization is therefore robust across model families and size tiers.

Where the conditions disagree is in absolute level, and the split is not by provider. At the high tier, Gemini 3 Flash Preview, DeepSeek V4-Flash, and the full-tier GPT-5.4 all recover the first author in roughly four-fifths to nine-tenths of cases (92% / 89% / 79%), while GPT-5.4-mini does so in fewer than six in ten (59%). The same pattern holds for `title` (83% / 83% / 86% versus 57%) and is even more pronounced for `co_authors` (74% / 66% / 63% versus 17%) and `pages` (71% / 49% / 70% versus 24%). GPT-5.4-mini is the consistent low outlier; the full-tier GPT-5.4 sits among the other frontier flash-class models.

	Gemini 3 Flash			GPT-5.4-mini			GPT-5.4 (full)			DeepSeek V4-Flash		
field	low	mid	high	low	mid	high	low	mid	high	low	mid	high
`title`	18	57	**83**	4	23	57	15	61	**86**	21	50	**83**
`first_author`	30	70	**92**	6	21	59	9	37	**79**	18	66	**89**
`co_authors`	23	48	**74**	2	7	17	7	24	**63**	12	33	**66**
`year`	77	86	86	53	68	81	67	83	84	73	84	84
`venue`	59	75	76	27	43	55	43	65	75	58	75	75
`volume`	74	90	96	46	55	65	64	76	83	63	76	86
`issue`	87	94	98	33	35	47	51	62	80	76	78	82
`pages`	26	61	**71**	2	3	24	6	35	**70**	10	23	**49**

Table 1. Per-field cloze recovery accuracy (%) for the four non-reasoning conditions, with the six half-decade citation strata collapsed into three coarse tiers (low: 10–100; mid: 101–1,000; high: $>$ 1,000). Cells are computed only over rows where the corresponding field is populated in the OpenAlex record (variable mask). Boldface marks the high-tier values for the four "information-rich" fields whose accuracy increases significantly across all three transitions (low→mid→high). Reasoning-enabled results are discussed in the text and shown in Fig. 2: Gemini's reasoning column is within a few points of the value shown; DeepSeek's is similar except that high-tier `first_author` falls (89 → 77) and `pages` rises (49 → 66); GPT-5.4-mini's is far higher than the value shown—high-tier `title` 57 → 79, `co_authors` 17 → 68, `pages` 24 → 70.

Inflection points span two orders of magnitude

To compare directly with the ∼90-citation inflection reported for GPT-4.1[18], we fit, for each condition and each field, a logistic regression of binary recovery on $\log_{10}(\text{cit})$ and read off the citation count at which the fitted curve crosses $P = 0.5$. Table 2 lists the inflections for the four information-rich fields. For the low-cardinality fields (`year`, `venue`, `volume`, `issue`) the fitted curve is already above 0.5 at the bottom of the citation range in most conditions, so an inflection in raw citations is not meaningfully defined and they are omitted; we also do not pool the four information-rich fields into a single summary inflection, since their citation–recovery curves saturate at different levels and a pooled fit would average over quantities whose interpretations differ.

Three patterns are visible. (i) For first-author memorization the inflection is 87 citations on Gemini, 145 on DeepSeek, and 502 on the full-tier GPT-5.4—the three values span a factor of ∼6 in raw citation count—while the GPT-5.4-mini estimate of 1,546 sits roughly 17× above Gemini's. The earlier single-model GPT-4.1 figure of ∼90 citations[18] falls within this range but does not anchor it: the spread is the substantive result. (ii) The same ordering holds for `title`, `co_authors`, and `pages`: Gemini, DeepSeek, and full-tier GPT-5.4 cluster within a few hundred to a few thousand citations of one another (the full-tier GPT-5.4 ties Gemini for the lowest `title` inflection, ∼200), whereas GPT-5.4-mini's range from ∼8× to more than 40× higher. The outlier is clearly the mini size tier, not the GPT family. (iii) Enabling reasoning does not move the inflections in a single direction. It is essentially flat for Gemini (first-author 87 → 71, `title` 214 → 226); mostly adverse for DeepSeek (first-author 145 → 458, `title` 242 → 379; only `co_authors` and `pages` improve); but strongly beneficial for GPT-5.4-mini, where every field's inflection drops substantially—`title` from 1,588 to 484, `pages` from $\sim 1.3\times 10^4$ to 863, and `co_authors` from $\sim 5.7\times 10^4$ (an extrapolation, see footnote of Table 2) into the low thousands. This heterogeneity is taken up in detail below; the directional claims are formally tested in the next subsection.

Uncertainty intervals and paired contrasts

The intervals reported in Table 2 confirm that the inflection differences emphasized above are not artifacts of point estimation. GPT-5.4-mini's first-author 95% interval ([1,219, 1,987]) is disjoint from those of every other non-reasoning condition (Gemini

	Gemini 3 Flash		GPT-5.4-mini		GPT-5.4	DeepSeek V4-Flash	
field	off	on	off	on	(full)	off	on
`first_author`	87 [72, 105]	71 [58, 86]	1,546 [1,219, 1,987]	1,071 [889, 1,317]	502 [424, 607]	145 [121, 173]	458 [382, 558]
`title`	214 [176, 260]	226 [187, 270]	1,588 [1,254, 2,086]	484 [406, 589]	198 [166, 235]	242 [197, 296]	379 [318, 461]
`co_authors`	310 [238, 404]	195 [151, 248]	$5.7\times10^{4\dagger}$ $[2.1, 29]\times10^{4}$	1,007 [813, 1,322]	1,205 [955, 1,574]	839 [649, 1,144]	584 [459, 758]
`pages`	201 [144, 271]	133 [100, 174]	$1.3\times10^{4\dagger}$ $[0.75, 2.5]\times10^{4}$	863 [699, 1,081]	721 [572, 899]	2,689 [1,877, 4,182]	881 [687, 1,146]

Table 2. Citation count at which the logistic recovery probability crosses 0.5, per condition and field, with 95% confidence intervals (in brackets, smaller font) from a $B = 1{,}000$ paper-cluster bootstrap (see Methods). "off"/"on" columns are the same model with reasoning disabled/enabled. †The 50% crossing for GPT-5.4-mini's `co_authors` and `pages` lies above the maximum citation count in the sample, so these two values—and the corresponding CIs, which extend to $\sim 10^5$ at the upper end—are extrapolations of the fitted curve, not interpolations. The earlier single-model GPT-4.1 first-author estimate from a model-chosen test set was ∼90 citations[18]; it falls within the present spread but does not anchor it.

[72, 105]; DeepSeek [121, 173]; full GPT-5.4 [424, 607]). To make magnitudes directly comparable to the log-citation x-axis of Figs. 1–2, we report paper-cluster bootstrap contrasts as $\log_{10}$ *inflection differences* in decades (equivalently, $\log_{10}$ of the inflection ratio). On first-author, GPT-5.4-mini's inflection sits $+1.25$ decades $[+1.12, +1.38]$ to the right of Gemini's, $+1.03$ $[+0.91, +1.16]$ to the right of DeepSeek's, and $+0.49$ $[+0.39, +0.60]$ to the right of full-tier GPT-5.4's; all three contrasts have $p < 0.001$.

The reasoning contrasts cleanly separate the three families. For Gemini, reasoning shifts the first-author inflection by -0.09 decades $[-0.15, -0.03]$ ($p = 0.006$)—statistically distinguishable from zero but practically small (a ∼20% citation difference at a base of ∼90). For DeepSeek the first-author shift is $+0.50$ decades $[+0.41, +0.59]$ ($p < 0.001$)—a clear adverse half-decade, driven by the intercept drop on `title` and `first_author`. For GPT-5.4-mini the first-author shift is -0.16 decades $[-0.27, -0.05]$ ($p = 0.004$); the much larger mini reasoning rescue lives on the other information-rich fields, with `title` moving by -0.52 decades $[-0.63, -0.42]$, `co_authors` by -1.75 $[-2.46, -1.35]$ (the wide upper bound reflects the extrapolation), and `pages` by -1.17 $[-1.46, -0.96]$, all with $p < 0.001$. In other words, mini's reasoning rescue is concentrated on the fields where its non-reasoning curve sits farthest to the right.

As a parametric complement, we also fit cluster-robust generalized estimating equations (GEE) per field with all seven conditions pooled and clustering on `openalex_id`, which separate intercept shifts from log-linear slope shifts that the bootstrap (working on the inflection directly) does not decompose. The GEE results agree with the bootstrap contrasts above in direction and significance for every effect the paper hangs on; we relegate the full coefficient tables and the intercept-vs-slope decomposition to the Supplementary Information.

Continuous citation curves

Figure 1 plots the same data as a continuous view: per-model binned mean accuracy with 95% confidence intervals and the fitted logistic curve, overlaid for the four non-reasoning conditions on each of the eight fields. The visual pattern matches the inflection table: Gemini, DeepSeek, and full-tier GPT-5.4 produce S-curves that lie within roughly half a decade of each other on every panel, while GPT-5.4-mini's curve is shifted to the right by roughly one full decade on the citation axis. For `co_authors` and `pages`, the GPT-5.4-mini curve does not reach 50% within the citation range covered by the data, so its inflection there is an extrapolation outside the test set.

Size tier, not family: the full-tier GPT-5.4 recovers normally

The GPT-5.4-mini result raises the question of whether the OpenAI family as a whole is conservative about memorizing bibliographic records, or whether the *mini* size tier specifically is. To test this we ran the full-tier GPT-5.4 on the identical 1,118-paper sample. The full model is dramatically stronger than the mini at every tier and field: at the high tier its recovery is 86% on `title`, 79% on `first_author`, 63% on `co_authors`, and 70% on `pages`, versus 57% / 59% / 17% / 24% respectively for the mini. Its inflection points fall in the same range as Gemini and DeepSeek—`title` at 198 citations (essentially tied with Gemini's 214 for the lowest of all non-reasoning conditions), `first_author` at 502, `co_authors` at 1,205, `pages` at 721—all between roughly 3× and 50× lower than the mini's. The "GPT family suppresses memorization" framing is therefore wrong: it is the mini-tier model that is the outlier, while the full-tier GPT-5.4 sits squarely among the other frontier flash-class models.

Reasoning's effect on recovery is model-dependent

Enabling chain-of-thought does not move recovery in one consistent direction; the sign and size of the effect depend on the model. For Gemini 3 Flash Preview the reasoning and non-reasoning curves are nearly co-located on every field — the first-author inflection shifts only from 87 to 71 citations, `title` from 214 to 226, and the largest single change is `pages` from 201 to 133 (Fig. 2). For DeepSeek V4-Flash the effect is larger and mostly adverse: the first-author inflection rises from 145 to 458 citations and `title` from 242 to 379, with only `pages` (inflection 2,689 → 881) and `co_authors` (839 → 584) improving. For these two models the pattern is consistent with prior closed-system findings that reasoning restructures how a model resolves an answer rather than giving it access to knowledge it does not have: when the answer is already memorized, extra deliberation can introduce second-guessing that displaces a correct verbatim recall; when the answer is partially reconstructed from structure (as for `pages`, where issue/volume/first-page co-occur in predictable ways), the deliberation can help.

GPT-5.4-mini, however, is the striking exception. With reasoning disabled it is the conservative outlier described above (first-author inflection 1,546 citations, with `co_authors` and `pages` crossings off the chart); enabling reasoning shifts its S-curves left across the board, dropping `title` from 1,588 to 484 citations (-0.52 decades) and `co_authors` and `pages` by more than a decade into the low thousands —back into the range occupied by the other frontier models. The first-author inflection moves more modestly, from 1,546 to 1,071, since that field's gap to the other models is smaller to begin with. In other words, reasoning does not give GPT-5.4-mini new knowledge so much as it lets the model *reach* bibliographic records its direct-decoding pass leaves inaccessible, partially closing the size-tier gap to GPT-5.4 full. The broader point is that whether "turn on thinking" helps a model reproduce a citation is not a property of reasoning in general but of the particular model: neutral for Gemini, mildly harmful for DeepSeek, substantially helpful for the weakest-at-direct-recall model in our set.

Cross-domain replication on Oncology

To bound the concern that the seven inflections above might be properties of a single domain (computer-science AI) or a single decoding sample, we repeated the entire protocol—master sampling, the seven conditions, paper-cluster bootstrap, and cluster-robust GEE—on OpenAlex's Oncology subfield (`primary_topic.subfield.id = subfields/2730`, under field `Medicine`), with the same filters and the same six half-decade citation bands. The Oncology master contains 1,079 papers (200 in each of the lower five bands; 79 in the top, where the underlying pool is sparser). Figure 3 plots the seven inflection points and their 95% paper-cluster bootstrap intervals side-by-side for the two domains; the corresponding Oncology per-tier accuracy and per-condition inflection tables (direct counterparts of Tables 1–2) are reported in the Supplementary Information.

The four structural claims replicate. Recovery scales log-linearly with citation count on every condition; the title/first-author-first hierarchy is preserved; GPT-5.4-mini is again the outlier, with its first-author inflection $+1.06$ decades $[+0.94, +1.20]$ to the right of Gemini's, $+1.22$ $[+1.08, +1.37]$ of DeepSeek's, and $+0.63$ $[+0.51, +0.77]$ of full-tier GPT-5.4's; and the per-field seven-condition GEE confirms that the first-author curves differ in intercept only on Oncology too (slope interaction $\chi^2(6) = 6.9$, $p = 0.33$, indistinguishable from the $\chi^2(6) = 8.2$, $p = 0.22$ on CS-AI). The shape of the citation–memorization curve is therefore stable to a change of research domain.

The absolute inflection levels are uniformly higher on Oncology by roughly a factor of 1.5–3: Gemini's first-author inflection moves from 87 to 218, full GPT-5.4's from 502 to 597, GPT-5.4-mini's from 1,546 to 2,519. This pattern is consistent with weaker pretraining coverage of clinical bibliographic records relative to CS-AI—fewer arXiv mirrors, fewer code-released artefacts, fewer scraped web copies per paper for the same citation count. The hierarchical and ordinal claims survive intact; what does not is the numeric value of the trust threshold. A practitioner using a fixed T across domains would systematically over-trust Oncology output.

The reasoning effect is partly preserved and partly domain-shifted. Gemini's effective neutrality replicates almost exactly ($\Delta = -0.01$ decades on first author $[-0.07, +0.06]$, $p = 0.83$, against the CS-AI value of -0.09). DeepSeek's adverse first-author effect *intensifies*: $+1.05$ decades on Oncology $[+0.93, +1.18]$ versus $+0.50$ on CS-AI, with the per-family GEE intercept dropping by $\beta = -3.07$ ($p = 4 \times 10^{-9}$); the model that already lost ground with thinking enabled loses substantially more in a domain where the identifier was less likely to be memorized to begin with. The most striking shift is on GPT-5.4-mini: the across-the-board reasoning rescue on CS-AI becomes *partial* on Oncology. On `title` reasoning still helps (-0.38 decades, $p < 0.001$); on `pages` the help is even larger (-2.19, $p < 0.001$); but on `first_author` the sign reverses ($+0.40$ decades $[+0.24, +0.57]$, $p < 0.001$). The reading that fits both domains is that reasoning lets the mini *reach* records on fields where partial structure is available (a title is recognisable from abstract context; pages follow journal-encoded numeric patterns) but cannot synthesize an identifier the model has not encoded—and in Oncology that identifier class is on average less memorized than in CS-AI. The practical takeaway is that whether reasoning helps depends both on the model and on *which* bibliographic slot is queried.

Adjusted accuracy isolates the memorization signal

For low-cardinality fields, a fraction of the raw recovery comes from priors rather than record-level memorization. We adjust by subtracting, per (tier, field) cell, the modal-value frequency on the master pool—the upper bound on accuracy attainable by always emitting the most common value (Methods).

The adjustment matters most for the saturated fields. The chance baseline reaches ∼33% for `issue` in the high tier (because the modal issue number across high-cited AI papers is small and sharply peaked), and rises from ∼10% to ∼14% for `year` from the low to the high tier (publication years are few and recent years dominate). After adjustment, `year`'s gain over chance is still ∼72% on Gemini, ∼70% on DeepSeek, and ∼67% on GPT-5.4-mini at the high tier; the year is recovered through genuine record-level memorization in addition to a strong prior. The adjustment leaves the cross-model ordering on the four information-rich fields qualitatively unchanged because their chance baselines are at most a few percent.

Discussion

The central finding is that the citation–memorization curve proposed in earlier single-model work generalizes in shape but not in scale. Every condition shows the same log-linear scaling of bibliographic recovery with citation count, the same field-level ordering of which metadata gets memorized first, and the same flattening once a field has saturated. What differs by one to two orders of magnitude is *where on the citation axis* the inflection sits, and the variation does not follow provider or open/closed boundaries: Gemini 3 Flash Preview (proprietary), DeepSeek V4-Flash (open-weight), and full-tier GPT-5.4 (proprietary) place the first-author inflection between 87 and 502 citations—a factor-of-six spread on its own—and GPT-5.4-mini sits a further decade higher still. The operative axis is the size tier, not the model family.

The practical implication is direct. A tool that uses a single citation threshold T for trusting an LLM-generated bibliographic record will be badly miscalibrated for at least one widely deployed frontier model: the first-author crossing already spans 87–502 citations across the three flash-class/full conditions and is roughly ∼1,500 for GPT-5.4-mini, so no single T ranks all four correctly. T also depends on *how* the model is run: enabling reasoning shifts GPT-5.4-mini's curve left by close to a decade, so the same model needs a roughly tenfold lower T in thinking mode. Calibration must be per-deployment and per-mode—and given that "mini" and "flash-lite" tiers serve much high-volume traffic, the within-family size-tier gap is not a corner case. This parallels question-answering findings that LM recall scales with entity popularity and retrieval is needed on the long tail[14,15]: our threshold is the bibliographic analogue of the popularity-based abstention threshold proposed there. A complication is that LLMs also *emit* head-of-distribution papers more often, reproducing an amplified Matthew effect[20]; a fixed citation threshold therefore risks compounding that bias and should be paired with retrieval-based abstention for long-tail records.

Why the mini tier is the outlier we cannot fully resolve here. The candidate mechanisms—fewer effective active parameters (parameters set the upper bound on stored knowledge: ∼2 bits per parameter on controlled synthetic data[21], on top of the generic capability-vs-size laws[22]), more aggressive training-data filtering or deduplication (deduplication alone reduces verbatim regurgitation by an order of magnitude[23]), or RLHF reward shaping that penalizes verbatim reproduction—are not separable with API-only access. One observation does narrow the field: enabling reasoning largely closes the mini's gap, which implies that the relevant records are *present* in the model and that the shortfall is in the path by which a single direct-decoding pass surfaces them, rather than in whether the records were learned at all. That points toward a decoding-policy or RLHF-shaping explanation over an outright data-filtering one, though it does not settle it. What the data do rule out is a provider-level or open/closed-level explanation: the open-weight DeepSeek and the proprietary Gemini cluster together, and the two GPT models are on opposite ends of the spread.

The reasoning condition further constrains the story. Chain-of-thought does not move recovery uniformly: Gemini's reasoning and non-reasoning curves are co-located; DeepSeek's reasoning curve sits *below* its non-reasoning one on first-author and title (89% → 77% at the high tier); GPT-5.4-mini's sits *far above*, with the `title` inflection falling from 1,588 to 484 and `co_authors`/`pages` both shifting by more than a decade. The unifying reading is that reasoning restructures how a model resolves an answer rather than expanding what it knows: when the answer is a memorized verbatim string, deliberation can displace it with a plausible reconstruction (the DeepSeek case, consistent with prior closed-system results[19] and with the meta-analytic finding that chain-of-thought[24] gains are concentrated in math/symbolic tasks and largely vanish on knowledge-heavy ones[25]); when direct decoding would otherwise confabulate, deliberation can route the model to the stored record (GPT-5.4-mini on CS-AI). Oncology sharpens this: the mini's reasoning rescue persists on `title` and `pages` but reverses on `first_author`—reasoning helps when partial structure can bootstrap recovery, but not when the identifier itself must be synthesized, and the balance is domain-dependent. The net effect of enabling reasoning therefore depends on model, field, and domain; practitioners cannot assume "turn on thinking" will help or hurt for bibliographic tasks without checking the specific combination.

On the strict scoring: a paraphrased title that captures the topic but is not the original is scored 0. The GPT-5.4-mini gap would shrink under a softer metric, but we retain strict scoring because the practical question is whether a user finds the *same*

paper, not one on the same topic—a topic-matched but non-identical record is precisely the misleading failure mode.

For the lowest-cardinality fields (`year`, `venue`, `volume`, `issue`) the adjusted-accuracy analysis showed that prior frequency contributes a non-trivial chunk of the raw score, but record-level memorization still accounts for the majority of recovery in the high tier. This nuances, but does not overturn, the earlier single-model observation that the year is essentially unlearned on GPT-4.1[18]: on every condition we tested the year is recovered well above chance even at low citation counts, but its citation gradient is much weaker than that of title or first author.

The study has several limitations. First, the master pool covers two OpenAlex subfields (computer-science AI and Oncology under Medicine) and only English-language articles indexed since 2010; broader cross-domain (e.g. humanities) and cross-language generalization remain untested. Second, the cross-vendor comparison is matched on each provider's stated capability/cost tier ("flash" / "mini"), which is a positioning claim rather than a controlled variable: parameter counts are not all disclosed, and a vendor's "flash" model may not be the analogue of another's "mini". We therefore treat the flash/mini triple as the nominal same-tier comparison and use GPT-5.4-full only as a within-vendor scale control; the mini-versus-full contrast inside the GPT family is suggestive but not a controlled scaling experiment. Third, the reasoning knob is not standardized across vendors—Gemini's dynamic thinking budget, DeepSeek's `high` effort level, and OpenAI's `medium` effort level are different operating points—so the reasoning comparison conflates "reasoning on vs. off" with "different reasoning intensities"; the qualitative direction of each within-model effect is robust to this, but the magnitudes are not directly comparable across models. Fourth, the cloze probe forces a single answer and does not surface uncertainty; pairing it with self-reported confidence or an explicit abstention option would distinguish "GPT-5.4-mini does not have the record memorized" from "GPT-5.4-mini has it but declines to commit to it" (the reasoning result suggests the latter is at least part of the story). Fifth, all conditions use each provider's default decoding temperature (1.0) and each cell is queried once, reflecting deployed behavior; the bootstrap captures between-paper but not per-trial sampling variability. The binned recovery curves are stable to per-trial noise at much higher precision than the inflection-CI width, so the headline comparisons are unlikely to flip under multi-trial decoding, but a fully decoupled estimate would need multi-trial decoding at every cell—earmarked for follow-up work. Finally, fabricated records that happen to look like the target paper are still scored 0—the desired behavior for our research question, but characterizing *which* substitute paper a model offers when it fails to recover the original is a separate analysis we do not attempt here.

Methods

Data collection

We retrieved the master pool from the OpenAlex Works API. We restricted to records whose primary topical classification is the Artificial Intelligence subfield (`primary_topic.subfield.id = subfields/1702`), were of `type:article`, were not retracted or paratext, were in English, and had a non-empty `abstract_inverted_index` so that the abstract could be reconstructed and used as cloze context. We required `publication_year` ≥ 2010 to balance coverage of the post-arXiv era against the citation pool size in the high tier.

Within these filters, papers were stratified into six half-decade citation bands so that the citation axis is sampled roughly uniformly in $\log_{10}$: $[10,32)$, $[32,100)$, $[100,316)$, $[316,1{,}000)$, $[1{,}000,3{,}162)$, and $[3{,}162,\infty)$ (the band edges are $10^{1.0}, 10^{1.5}, \ldots, 10^{3.5}$). This replaces the coarse low/mid/high tiering of an earlier single-model study[18] (whose boundaries were anchored on its own $\sim$90 and $\sim$1,200 estimates) with a design that does not pre-commit to where the inflection lies and yields smooth continuous recovery curves.

OpenAlex's `sample` parameter is hard-capped at 200 records per call and does not support cursor pagination. To obtain an unbiased random sample within each band we issued one or more `sample=200` calls with rolling seeds and de-duplicated by `id`. We verified empirically that this preserves the within-band citation median (near the geometric centre of each band), whereas cursor pagination *without* `sample` inherits OpenAlex's default sort by citation count and biases the sample toward the upper edge of a band.

We drew 200 papers per band, except the top band ($\geq 3{,}162$ citations), for which only 118 records in the AI subfield satisfied the filters; the CS-AI master sample therefore contains 1,118 papers (200 in each of the lower five bands, 118 in the top). For the cross-domain replication we sampled the same six bands from OpenAlex's Oncology subfield (`subfields/2730`, under field `Medicine`) with the identical filter set; the top band there yielded 79 papers, so the Oncology master contains 1,079 papers (200 in each of the lower five bands, 79 in the top). For within-master reproducibility, each band's papers are sorted by `openalex_id` (a stable order) and shuffled deterministically with seed 42; all conditions in a given domain are evaluated on exactly the same sample, so within-domain cross-condition differences cannot be attributed to sample composition. (For two conditions — GPT-5.4-mini with reasoning on each domain — the run was split into two non-overlapping halves of the deterministic order and merged — see *Runtime infrastructure*; on Oncology the second half draws zero papers from the top band, since with `-start 100` the band of size 79 is exhausted by the first half.)

Cloze task design

The cloze design adapts the LAMA fill-in-the-blank knowledge probe[26] from (subject, relation, object) knowledge-base triples to bibliographic records: for each paper we mask one bibliographic field at a time and ask the model to recover it conditional on the rest of the record plus the abstract. Eight fields are maskable: `title`, `first_author`, `co_authors`, `year`, `venue`, `volume`, `issue`, and `pages`. Numeric fields (`volume`, `issue`, `pages`) are frequently empty for conference papers and preprints; rather than restrict the analysis to fully populated records (which would bias toward older traditional journal articles) we use a *variable mask*: for each paper we mask only the fields that are non-empty in the OpenAlex record. The reported n per (tier, field) cell therefore varies and is shown alongside every accuracy figure.

The abstract is included as part of the unmasked context for two reasons. First, recovering the title from authors and venue alone is not a memorization probe; it is essentially blind guessing over a target space whose cardinality is the entire AI literature. The abstract makes the title-recovery task tractable as a measurement of whether the specific paper is known to the model, while leaving the other fields' tasks largely unaffected. Second, providing the abstract operationalizes the realistic user query of the form "is there a paper that does X?", where X is a description of the content. Our cloze condition is thus a controlled instance of the deployment scenario in which fabricated citations are most damaging.

The model is asked to return a JSON object with a single key `answer` whose value is an object mapping each masked field to its predicted value. Output is constrained to the JSON-mode of each API where available, with markdown-fence fallback parsing for less constrained backends. The full prompt template is provided in the project repository.

Scoring

Each masked field is scored as 0 or 1 by a deterministic field-specific match rule applied to normalized strings. The normalization rules are: case-folding, Unicode NFKC normalization, and field-specific whitespace and punctuation handling (`title`); surname and first-initial agreement across both display-style "Geoffrey E. Hinton" and APA-style "Hinton, G. E." (`first_author`); ordered list match on the co-author sequence with the same name comparator (`co_authors`); extraction of the first four-digit number (`year`); a curated alias table for major AI venues that maps abbreviations to their canonical full names (e.g. "NeurIPS" $\leftrightarrow$ "Advances in Neural Information Processing Systems"), with normalized exact match otherwise (`venue`); stripping of "Vol."/"No."/"Issue" prefixes followed by alphanumeric exact match (`volume`, `issue`); and pp. stripping plus dash-variant unification (`pages`).

We deliberately do not use token-similarity, embedding-similarity, or other graded fidelity metrics for the title or other long-string fields. The research target of this study is not to measure how *semantically close* a model's output is to the true bibliographic record, but to characterize the conditions under which the model returns a record that a downstream user would correctly interpret as the true paper. Under that lens a paraphrased or topic-similar but non-identical title is precisely the failure mode of interest—it is a record that may not refer to any real paper at all, or refers to a different real paper, while presenting itself as the original. Graded metrics inflate accuracy on exactly these cases. The strict field-level binary score is therefore the appropriate operationalization of "correctly recovered the bibliographic record".

Statistical analysis

For each field and each of three coarse tiers (low: 10–100; mid: 101–1,000; high: $>$ 1,000; obtained by collapsing the six sampling bands in pairs) we report the raw accuracy together with a *chance baseline*, defined as the modal-value frequency of that field within that tier in the master sample. This is the maximum accuracy attainable by a model that always emits the modal value, and bounds above the contribution of prior-frequency to the observed accuracy. The adjusted accuracy is the raw accuracy minus the chance baseline. This adjustment is most consequential for low-cardinality fields: `year` and `issue` have chance baselines of 10–33%, so a non-trivial portion of their raw recovery would be obtainable without any record-level memorization.

We characterize the citation–memorization curve per field by maximum likelihood logistic regression of the binary score on $\log_{10}(\text{cit})$, $P(\text{score}=1) = \sigma(b_0 + b_1 \log_{10}(\text{cit}))$, with standard errors derived from the inverse Hessian. The inflection point where $P = 0.5$ is reported in raw citations as $10^{-b_0/b_1}$, allowing direct comparison with the $\sim$90-citation estimate of prior work[18]. Where the fitted slope is shallow and the crossing falls outside the sampled citation range (notably `co_authors` and `pages` for GPT-5.4-mini without reasoning), the reported value is an extrapolation and is flagged as such.

For tier-level inference we report a χ^2 test on the (3 coarse tier $\times$ correct/incorrect) contingency table per field, plus pairwise proportion z-tests for the low$\rightarrow$mid, mid$\rightarrow$high, and low$\rightarrow$high transitions. Significance is Bonferroni-corrected over the 24 pairwise comparisons (8 fields $\times$ 3 pairs).

The per-field logistic design is paper-clustered in a benign way: by construction each paper contributes at most one trial per field, so within-field observations are independent at the paper level and the inverse-Hessian standard errors for b_0 and b_1 are valid. Cross-field summaries and any condition–condition contrast that uses the same paper twice do not enjoy this property. To bound the inference correctly in all cases we attach a paper-cluster bootstrap to every inflection in Table 2: with $B = 1{,}000$

iterations, each iteration resamples 1,118 paper ids with replacement from the master sample, gathers all trials belonging to those papers (so a resampled paper contributes its full set of fields together), re-fits the per-(condition, field) logistic, and stores the resulting inflection $10^{-b_0/b_1}$. The 95% intervals shown in the table are the 2.5th and 97.5th percentiles of these bootstrap distributions. For paired contrasts (Δinflection between two conditions, e.g. GPT-5.4-mini vs. Gemini), the two conditions use the *same* bootstrap resample on each iteration; the two-sided p-value is $2\min(\Pr[\Delta < 0], \Pr[\Delta > 0])$ over the iterations. Bootstrap inflections that are non-finite (degenerate fit on a small resample) are dropped before percentile computation; the proportion is always under 5% even for the extrapolated GPT-5.4-mini `co_authors/pages` cells.

Parametric cluster-robust GEE complements to the bootstrap analysis (per-field omnibus Wald tests and per-family reasoning interactions), together with the full coefficient tables, are reported in the Supplementary Information.

For the continuous view we report mean accuracy with Wald 95% confidence intervals in log-spaced bins of width 0.3 on $\log_{10}(\text{cit})$ ($\approx 2\times$ in raw citations per bin). The fitted logistic curve is overlaid on the binned points in the per-model and cross-model figures.

Models and tier matching

We evaluated three frontier "flash"-class models from three different providers: Gemini 3 Flash Preview (Google AI Studio), GPT-5.4-mini (OpenAI Chat Completions), and DeepSeek V4-Flash (DeepSeek API). These are matched on each provider's stated fast/low-cost tier rather than on disclosed parameter counts (which are not all public); the tier labels are positioning claims, not a controlled variable, so the matching is approximate—a vendor's "flash" model is not guaranteed to be the analogue of another's "mini". To probe whether the GPT family's apparent memorization-suppression is a property of the family or of the "mini" size tier specifically, we additionally ran the full-size GPT-5.4 on the identical sample as a within-vendor scale control. Counting the three reasoning-enabled variants (below), seven conditions in total were run on the 1,118-paper sample: Gemini 3 Flash Preview, GPT-5.4-mini, DeepSeek V4-Flash, GPT-5.4 (full), and the reasoning-enabled variants of Gemini, DeepSeek, and GPT-5.4-mini. The GPT-5.4-mini reasoning condition was included specifically so that the reasoning comparison has a representative from each provider at the flash/mini tier, rather than only Gemini and DeepSeek.

For reproducibility, the exact API model identifier strings queried were `gemini-3-flash-preview` (Google AI Studio), `gpt-5.4-mini` and `gpt-5.4` (OpenAI Chat Completions), and `deepseek-v4-flash` (DeepSeek API). All runs were executed in May 2026; the API providers may revise the underlying snapshots without changing the user-facing identifier, so the per-condition thresholds reported here are valid for the snapshots served during that period.

Reasoning condition

Each model family exposes its own knob for hybrid thinking-mode inference, with no shared API. We mapped a single boolean `reasoning` flag onto each provider as follows:

- **Gemini 3 Flash Preview.** `generationConfig.thinkingConfig.thinkingBudget` controls the budget. `0` disables thinking, `-1` enables a dynamic budget where the model decides how much to think per request. We use `0` for the non-reasoning condition and `-1` for the reasoning condition. The reasoning trace is not surfaced as visible output (Gemini does not expose thinking parts to API callers in non-streaming mode), but the thinking tokens are billed at the output-token rate and counted in the response's `usageMetadata.thoughtsTokenCount`.

- **OpenAI GPT-5.4 / GPT-5.4-mini.** `reasoning_effort` controls the budget on Chat Completions, with levels `none`, `low`, `medium`, `high`, and `xhigh`. The two non-reasoning OpenAI conditions (GPT-5.4-mini, GPT-5.4 full) use `none`. The reasoning-enabled condition is GPT-5.4-mini at `medium`: the mini is the model whose direct-decoding curve is the outlier, so whether thinking moves it is the question of interest, and running the mini (rather than the full model) with reasoning keeps the reasoning comparison at the same tier as the Gemini and DeepSeek reasoning conditions. The reasoning trace is not surfaced in the Chat Completions response, but the thinking tokens are billed and counted in `usage.completion_tokens_details.reasoning_tokens`.

- **DeepSeek V4-Flash.** The body field `thinking` takes `{"type": "enabled"}` or `{"type": "disabled"}`. When enabled, `reasoning_effort` provides finer control with documented levels `low`, `medium`, `high`, and `max`. We used `high` for the reasoning condition. The reasoning trace is returned as a separate `reasoning_content` field on the message object (sibling of `content`); thinking tokens are billed at the output-token rate and counted in `usage.completion_tokens_details.reasoning_tokens`.

For every reasoning-enabled condition the maximum output length was raised from 256 to 4,096 tokens to give the thinking trace room to complete; on sanity checks the observed thinking length per cloze item was a few hundred tokens (e.g. ∼130 for DeepSeek on `first_author`, ∼350–540 for Gemini), and all three reasoning runs completed comfortably within the 4,096 budget with no length-truncated responses. Outside the reasoning conditions the cap remains 256 tokens.

Setting `response_format` to JSON-object (or `responseMimeType` on Gemini) was retained for both conditions. Temperature defaulted to the provider's default (`1.0` on all three) so as not to introduce a confound between the conditions purely from sampling settings; this is the deployed-behavior setting, and we discuss its consequences for the sampling variability of the inflection estimates in the Limitations.

Runtime infrastructure

The runner was implemented in Python with a unified `complete()` interface (taking the prompt, a boolean reasoning toggle, a JSON-mode flag, temperature, and a maximum token budget) across all providers, with provider-specific retry-with-backoff on transient errors. API errors were recorded as a non-null `error` column and the corresponding row scored 0; the run continued through them. Errors were rare across all seven conditions ($< 0.1\%$ each): one for Gemini, zero for Gemini with reasoning, zero for GPT-5.4-mini, six for GPT-5.4-mini with reasoning, one for GPT-5.4 full, zero for DeepSeek, and five for DeepSeek with reasoning (mostly read timeouts on long author lists or sporadic HTTP 400s).

Two conditions used a non-synchronous execution path. (i) Gemini's synchronous `generateContent` endpoint has a daily request quota too small for a 7,367-call sweep at the tier available to us, so the Gemini-with-reasoning condition was submitted through Google's Batch API: the request set was written as JSONL, uploaded via the Files API, and submitted as batch jobs; because a single batch job above a few thousand requests is rejected, the set was split into four chunks of $\leq 1{,}850$ requests, each chunk's output JSONL was downloaded and the results merged. All 7,367 responses returned, with no missing items, and the batch completed in minutes; Batch API pricing is half the synchronous rate. (ii) The GPT-5.4-mini-with-reasoning run is slow per call ($\sim$8–13 s, dominated by the reasoning pass), so it was split into two non-overlapping halves of the deterministic sample order and run in parallel ($\sim$9 h wall-clock), then merged.

Each non-reasoning run produced 7,367 cloze trials (1,118 papers $\times$ variable mask, average ~ 6.6 fields per paper) in roughly 2–4 hours wall-clock ($\sim$1–3 s per call median) and cost on the order of USD 2–5 per model on paid tiers. The DeepSeek reasoning run, executed synchronously, took $\sim$13 h; the Gemini reasoning batch returned in minutes. The runner stores each cloze outcome as a row with provenance fields (`openalex_id`, `master_parquet`, `seed`, `start`, `model`, `reasoning`) so that runs from different models, modes, or sample slices can be cross-referenced and merged.

Reproducibility

The master and run data files (including the batch-job state files for the Gemini reasoning condition and the two half-runs for GPT-5.4-mini with reasoning), normalization code, and analysis scripts are released at the project repository. Each run file is a self-contained artefact: the same combination of `master_parquet` file, `seed`, `start`, and per-band sample size deterministically reproduces the exact set of 1,118 papers used. Strict field-level scoring and the chance-baseline computation are likewise deterministic functions of the master and run files.

References

1. Walters, W. H. & Wilder, E. I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Reports* **13**, 14045, DOI: 10.1038/s41598-023-41032-5 (2023).

2. Bhattacharyya, M., Miller, V. M., Bhattacharyya, D. & Miller, L. E. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* **15**, e39238, DOI: 10.7759/cureus.39238 (2023).

3. Sakai, Y., Kamigaito, H. & Watanabe, T. Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in acl conferences. *arXiv preprint arXiv:2601.18724* DOI: 10.48550/arxiv.2601.18724 (2026).

4. Xu, Z. *et al.* Ghostcite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718* (2026).

5. Dahl, M., Magesh, V., Suzgun, M. & Ho, D. E. Large legal fictions: Profiling legal hallucinations in large language models. *J. Leg. Analysis* **16**, 64–93, DOI: 10.1093/jla/laae003 (2024).

6. Magesh, V. *et al.* Hallucination-free? assessing the reliability of leading ai legal research tools. *J. empirical legal studies* **22**, 216–242, DOI: 10.48550/arxiv.2405.20362 (2025).

7. Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. Why language models hallucinate. *arXiv preprint arXiv:2509.04664* (2025).

8. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Inf. Syst.* **43**, 1–55 (2025).

9. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM computing surveys* **55**, 1–38, DOI: 10.1145/3571730 (2023).

10. Kalai, A. T. & Vempala, S. S. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, DOI: 10.1145/3618260.3649777 (2024).

11. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* DOI: 10.48550/arxiv.2401.11817 (2024).

12. Carlini, N. *et al.* Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650, DOI: 10.48550/arxiv.2012.07805 (2021).

13. Carlini, N. *et al.* Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, DOI: 10.48550/arxiv.2202.07646 (2022).

14. Kandpal, N., Deng, H., Roberts, A., Wallace, E. & Raffel, C. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, DOI: 10.48550/arxiv.2211.08411 (2023).

15. Mallen, A. *et al.* When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, DOI: 10.18653/v1/2023.acl-long.546 (2023).

16. Tirumala, K., Markosyan, A. H., Zettlemoyer, L. & Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, DOI: 10.48550/arxiv.2205.10770 (2022).

17. Agrawal, A., Suzgun, M., Mackey, L. & Kalai, A. T. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, DOI: 10.48550/arxiv.2305.18248 (2024).

18. Niimi, J. Hierarchical memorization in large language models: Evidence from citation generation. In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART)*, DOI: 10.48550/arxiv.2511.08877 (Springer LNCS, 2026). Accepted in post-proceedings.

19. Niimi, J. Distortion instead of hallucination: The effect of reasoning under strict constraints (2026). DOI: 10.48550/arxiv.2601.01490. 2601.01490.

20. Algaba, A. *et al.* Large language models reflect human citation patterns with a heightened citation bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 6844–6879, DOI: 10.18653/v1/2025.findings-naacl.381 (2025).

21. Allen-Zhu, Z. & Li, Y. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405* (2024).

22. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* DOI: 10.48550/arXiv.2001.08361 (2020).

23. Lee, K. *et al.* Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8424–8445, DOI: 10.18653/v1/2022.acl-long.577 (2022).

24. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. neural information processing systems* **35**, 24824–24837 (2022).

25. Sprague, Z. R. *et al.* To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, DOI: 10.48550/arxiv.2409.12183 (2025).

26. Petroni, F. *et al.* Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, DOI: 10.18653/v1/d19-1250 (2019).

Author contributions

J.N. designed the study, developed the data and analysis pipeline, ran the experiments, and wrote the manuscript.

Funding

This work was supported by JSPS KAKENHI Grant Number JP24K16472.

Data availability

OpenAlex (https://openalex.org) is a public, openly licensed bibliographic database. The master samples, the combined cloze-trial records, the analysis outputs, and the analysis code for this study are archived at Zenodo (https://doi.org/10.5281/zenodo.20237161).

Additional information

Competing interests. J.N. serves as a technical advisor to a company in the manufacturing sector; this role is unrelated to the models, data, and analysis reported in this study. J.N. declares no other competing interests.

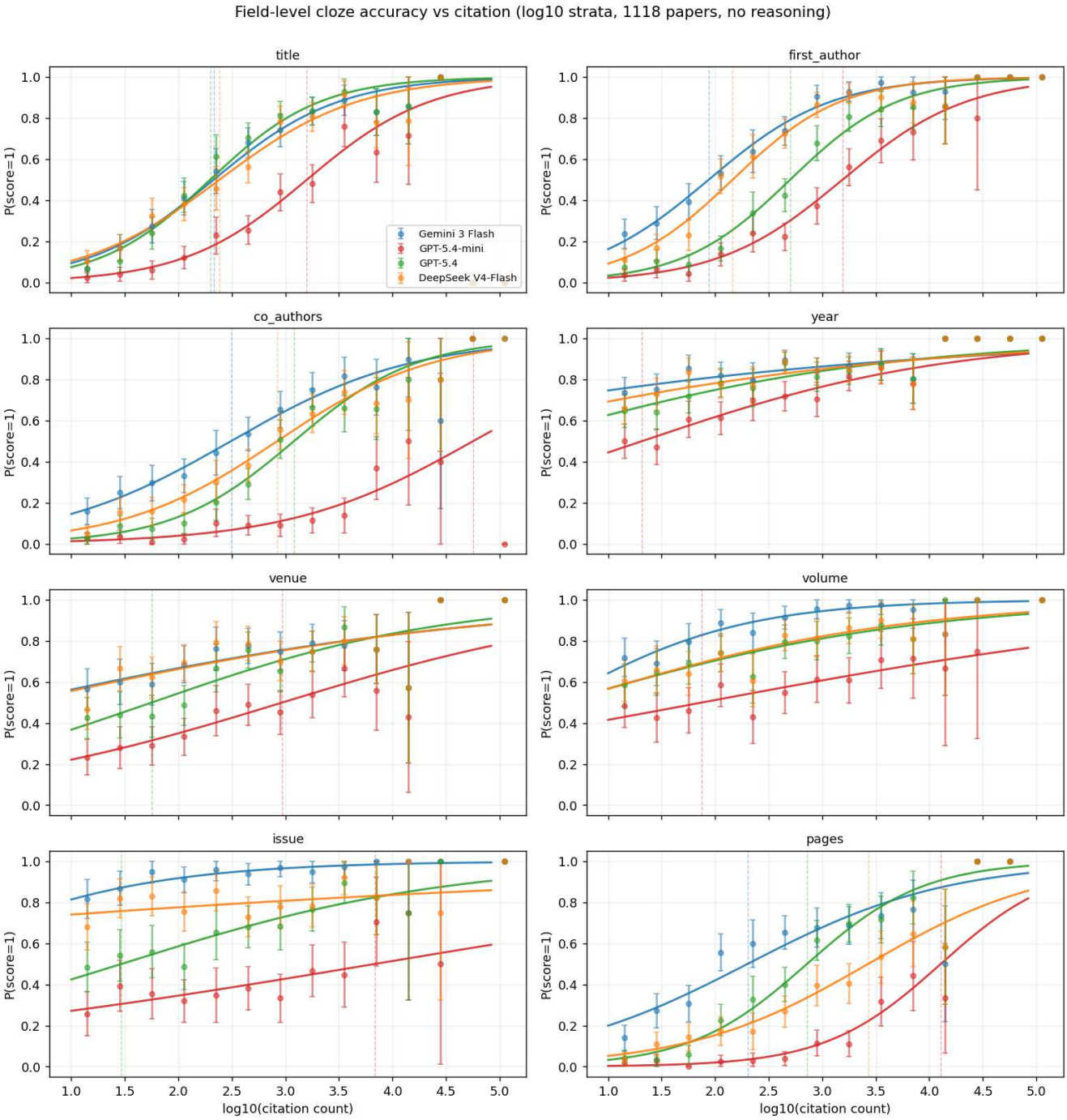


Figure 1. Per-field cloze recovery accuracy as a function of $\log_{10}(\text{citation count})$ for the four non-reasoning conditions. Dots mark binned mean accuracy with Wald 95% CIs in log-spaced bins of width 0.3; solid lines are the per-condition logistic fits; vertical dashed lines mark the $P = 0.5$ inflection. Gemini 3 Flash Preview (blue), GPT-5.4 full (green), and DeepSeek V4-Flash (orange) track within roughly half to three-quarters of a decade of one another on every field; GPT-5.4-mini (red) is shifted right by roughly one decade or more, and on `co_authors` and `pages` its curve does not reach 50% within the sampled citation range.

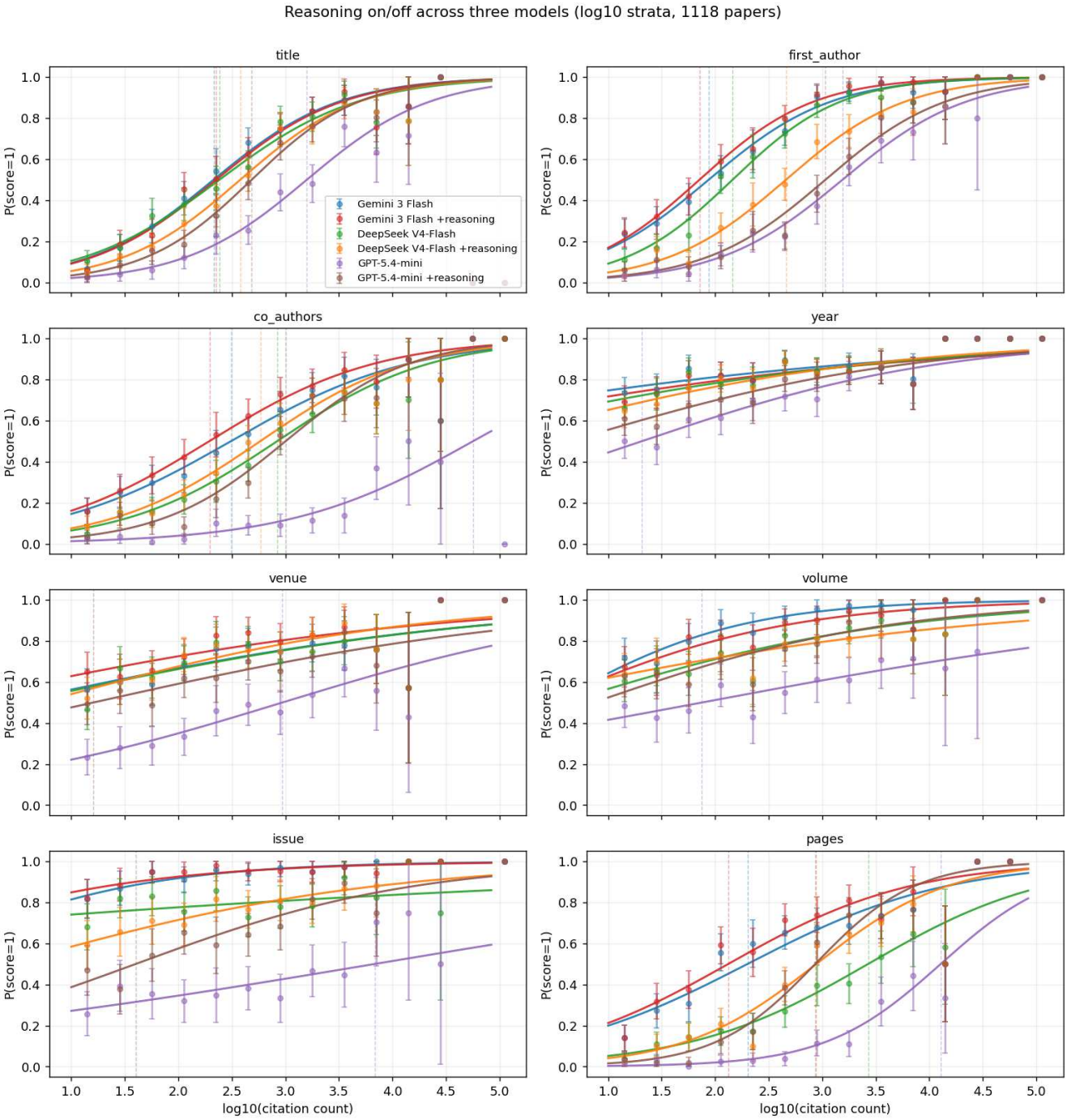


Figure 2. Effect of enabling chain-of-thought on cloze recovery, for three models, each plotted with reasoning off and on. Same axes and binning as Fig. 1. The effect is heterogeneous. Gemini 3 Flash Preview's reasoning (red) and non-reasoning (blue) curves are nearly co-located. DeepSeek V4-Flash's reasoning curve (orange) is mostly below its non-reasoning curve (green) on the identifier fields. But GPT-5.4-mini's reasoning curve (brown) lies *far above* its non-reasoning curve (purple): enabling reasoning shifts the mini's S-curves left by roughly a decade on the citation axis, recovering most of its gap to the other models.

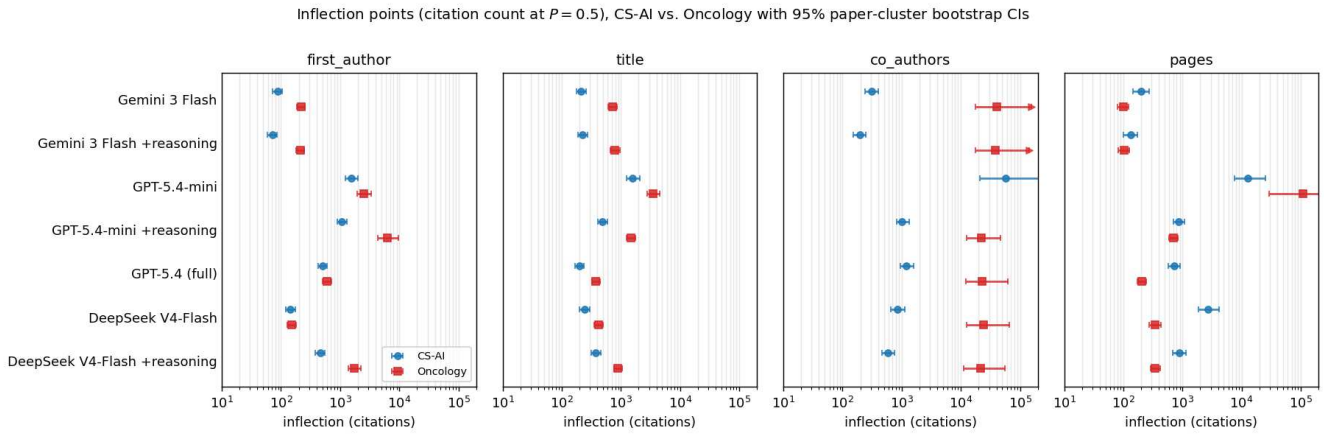


Figure 3. Per-field inflection points and 95% paper-cluster bootstrap intervals for the seven conditions, in CS-AI (blue circles) and Oncology (red squares). *x*-axis is citation count on a log scale. For each condition the Oncology marker is offset slightly below the CS-AI marker on the same row. Faint right-arrows on a CI bar indicate that the upper bound was clipped at 10^5 citations (it extends further, reflecting an extrapolation beyond the sampled range). The four structural claims of the CS-AI analysis—log-linear scaling, the field hierarchy, GPT-5.4-mini as the cross-vendor outlier, and shared first-author slope across models—replicate qualitatively in Oncology; absolute inflection levels shift higher by a factor of 1.5–3, and the GPT-5.4-mini reasoning rescue seen on CS-AI becomes partial on Oncology (helpful on `title` and `pages`, reversed on `first_author`).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf