

Web Search Is Not Enough: LLMs Fail to Account for Paper Retraction Status

Huy Hoang Ha¹ Benoit Favre² Cyril Labbe¹ François Portet¹

¹Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

²NAVER LABS Europe, 38240 Meylan, France

huy-hoang.ha@univ-grenoble-alpes.fr benoit.favre@lis-lab.fr

{cyril.labbe, francois.portet}@univ-grenoble-alpes.fr

Abstract

Large language models (LLMs) learn scientific knowledge from static training corpora, while web search can provide updated information such as paper retractions. This matters for LLM-based literature systems because a model may have seen a paper as valid evidence, while current web results report that it has been retracted. If an LLM does not use the updated retraction status, it may cite a retracted paper as valid evidence. We analyze eleven open-weight models on the retraction-status task and ten on the citation-recommendation task. With web-search results provided, LLMs increase correct status answers for retracted papers from 11.0% to 94.0%. But the same models do not use the retrieved context as reliably when no retraction signal is present or when they recommend citations. On non-retracted papers under web search, models answer “I don’t know” in 52.9% of answers in the retraction-status task. In the citation-recommendation task, even when search results state the retraction and the prompt asks for warnings, models recommend retracted papers without warning in 10.2% of answers. A separate pattern appears for highly cited retractions. Models give fewer correct status answers for high-citation than low-citation retractions (89.3% versus 96.4%). OLMO’s training data suggests one reason. Highly cited retractions appear more often in training data, mostly without any retraction signal, so retrieved retraction information may compete with outdated training knowledge. Overall, access to web search does not guarantee that LLMs will correctly use a paper’s current status when answering questions or producing citations, highlighting the broader challenge of knowledge updating in LLM systems.

1 Introduction

Large language models (LLMs) can use retrieval over scientific literature to answer questions and synthesize citation-backed responses (Skarlinski

et al., 2024; Asai et al., 2026). AI-scientist systems also use LLMs to synthesize literature and generate research hypotheses and proposals (Gottweis et al., 2025; Yamada et al., 2025; Boiko et al., 2023). These systems are useful only if they track changes in the papers they discuss. Retractions give a direct way to study that problem. A paper can appear in training data as valid scientific evidence and later be retracted, changing how it should be used. Web search can provide that newer status to the model (Vu et al., 2024). The question is whether the model follows the retrieved context when it conflicts with learned information about the same paper. We therefore ask not only whether LLMs can identify a retraction, but whether that status changes what they answer and what they cite.

We analyze eleven open-weight LLMs ranging from 8B to 1T parameters under two search conditions: with and without the same web-search results. We use two tasks. The retraction-status task asks whether a paper has been retracted, including non-retracted comparison papers selected to match retracted papers by journal, field, year, and citation range. The citation-recommendation task asks models to recommend papers for a scientific topic using the same web-search results across models. We classify whether responses cite the retracted paper and whether the citation includes an explicit warning that the paper was retracted. We include both tasks because a model can answer correctly in the retraction-status task while still citing the same retracted paper without warning (Figure 1).

Web search gives the clearest gain when the question directly asks about retraction status. Correct status answers on retracted papers rise from 11.0% without search to 94.0% with web search. This gain does not mean that models use paper status reliably. On non-retracted comparison papers under web search, models often answer “I don’t know” instead of identifying the paper as not retracted. In the citation-recommendation task, models still cite

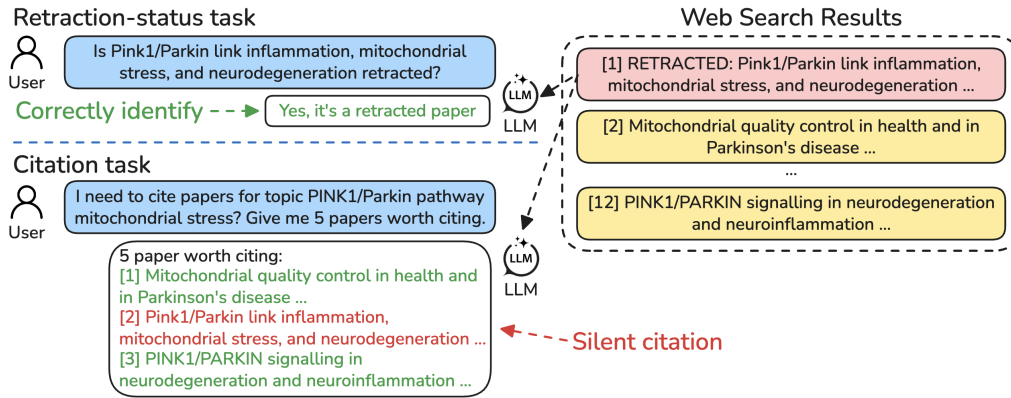


Figure 1: **The retraction-status and citation-recommendation tasks can diverge.** Web search can help a model identify a retraction in the retraction-status task, while the same model can cite the retracted paper without warning in the citation-recommendation task.

retracted papers without warning, even when the search text contains a retraction signal (Figure 1). Correct status answers are also lower for highly cited retractions than for low-cited retractions.

One possible explanation is that older training text conflicts with newer retraction information. We examine this possibility using Infini-gram, which allows us to search OLMo-3 training text for references to each paper and its retraction (Team OLMo, 2025; Liu et al., 2024). We find many references to retracted papers, but most do not include a retraction signal. This imbalance is consistent with older training mentions competing with the retraction status provided through web search. Similar conflicts between parametric knowledge and retrieved context have been reported in prior work (Mallen et al., 2023; Xie et al., 2024; Goyal et al., 2025; Zhao et al., 2025).

Section 3 defines the two tasks, non-retracted comparison papers, and saved web-search results shared across models. Section 4 follows the main behavior pattern: web search helps when models are asked directly about retraction status, but the same status is not reliably used on non-retracted comparison papers, citation recommendations, or highly cited retractions. Section 5 then examines OLMo training text as one possible explanation for the high-citation pattern. We close with implications for LLM-based literature systems and knowledge updating with retrieved context.

2 Related Work

Factuality, freshness, and retrieved context. Factuality benchmarks study whether LLMs avoid false beliefs, support atomic facts, or recover cur-

rent answers after knowledge changes (Lin et al., 2022; Min et al., 2023; Li et al., 2023; Vu et al., 2024). This line of work is closely related to our setting because retractions are a form of knowledge update: the paper remains the same, but its status changes after publication. Retrieval-augmented generation (RAG) adds the question of context use. RAG evaluation frameworks measure whether answers are faithful and relevant to supplied context (Es et al., 2024). Our setting asks the same question for paper status: when web search supplies the current retraction status, does that status change what the model answers and cites?

Retracted papers in LLMs and citation behavior.

Recent work has examined how LLMs respond to retracted scientific articles directly (Thelwall, 2026; Zhang et al., 2025; Yao et al., 2025; Labenbacher et al., 2026). Bibliometric studies show that retracted papers continue to be cited for years after the notice (Bar-Ilan and Halevi, 2017; Hsiao and Schneider, 2022; Avenell et al., 2019). Algaba et al. (2025) showed that model-generated citations reflect human citation networks, with stronger bias toward highly cited papers. These studies show that retracted papers remain visible in both human and model-generated citations. Our focus is whether retrieved retraction status changes model behavior once it is placed in the context. We study this with saved web-search results that are held constant across models, non-retracted comparison papers, and citation-count ranges.

Retrieved context and parametric knowledge.

Another line of work related to knowledge updating studies knowledge conflicts, asking whether

models prioritize retrieved context, competing evidence, or information learned during training (Xie et al., 2024; Longpre et al., 2021; Chen et al., 2022; Wang et al., 2024; Xu et al., 2024; Lee et al., 2025). Prior work shows that models often rely on parametric knowledge, information stored in model weights during training, for popular entities, and that instruction tuning does not always increase reliance on retrieved context (Mallen et al., 2023; Goyal et al., 2025). Retractions create a specific knowledge-conflict setting: highly cited papers may be strongly represented in training, while web search exposes their changed status. Citation count therefore gives an indirect way to ask whether papers that appear more often in training data show weaker use of retrieved retraction status.

3 Data and Experiment Setup

We study two ways an LLM can use retraction status. The retraction-status task asks whether the model can identify the status of a named paper. The citation-recommendation task asks whether that status changes which papers the model recommends and whether it warns the user about a retracted paper (Figure 1). We use fixed saved search results. The models do not formulate queries, search the web, open pages, rerank results, or decide whether to verify a paper. Every model receives the same retrieved text. This lets us study how models use the provided context without differences in retrieval across models.

Tasks and labels. We construct two tasks and classify responses with task-specific labels.

The *retraction-status task* names a paper and asks whether it has been retracted. We classify each answer as correct, incorrect, or uncertain. For non-retracted comparison papers, an incorrect “retracted” answer is a false warning. We run all eleven models on this task, with and without web search.

The *citation-recommendation task* asks the model to recommend five papers for a scientific topic. For each paper, the model returns a title, authors, year, and caveat in a JSON schema (Appendix A.1). We compare the four saved-search conditions defined under Web search in Section 3. We normalize each recommended title and compare it with the target title. We require an exact match, except when a truncated title uniquely matches the target. A *warned citation* includes the target paper and states its retraction in the title or caveat.

A *silent citation* includes the target paper without this warning. If the response does not include the target paper, the outcome is *no citation*. This matching rule can miss valid title variants, so our target-citation percentages are conservative estimates. Appendix A.2 reports an audit of these labels. We run this task on the ten local models. We exclude MiMo-V2.5-Pro-1T because its API could not sustain the required request volume.

Data. We use a February 2026 local snapshot of the Retraction Watch database with 68,154 records (The Center for Scientific Integrity, 2018; Hendricks et al., 2023). We remove duplicate DOIs and keep only entries retracted for scientific reasons, such as image duplication or unreliable data, and exclude administrative cases, such as authorship or publisher errors. Appendix B lists the inclusion and exclusion terms. The filtered set contains 11,382 entries.

The main analysis uses entries retracted before 1 January 2025. For these papers, both the original paper and the retraction notice could appear in model training data, which lets Section 5 compare retrieved retraction status with older training-text mentions in OLMo-3.1-32B.

We start with highly cited papers, including 65 papers with at least 500 citations, then sample additional papers across lower citation-count ranges until the main set contains 566 retracted papers (~5% of the filtered set). For each paper, we gather citation counts from OpenAlex (Priem et al., 2022), Semantic Scholar (Kinney et al., 2023), Crossref (Hendricks et al., 2020), PubMed cited-in links (Sayers et al., 2026), and Dimensions (Hook et al., 2018), then use the highest citation count found. This gives the analysis a wide citation range rather than only high-citation papers. For comparison, we use non-retracted papers indexed in the OpenAlex database. We match each retraction to a non-retracted paper on journal, subject area, publication year (± 2 years), and citation-count range. This gives 624 non-retracted comparison papers.

Models. We use open-weight models for reproducibility and coverage across model families. Open weights let us use fixed model releases and run the local models with the same prompts, sampling settings, and saved inputs. The panel contains eleven LLMs from eight model families, spanning 8B to 1T parameters. Appendix J lists the full model panel, citations, serving setup, and sampling details.

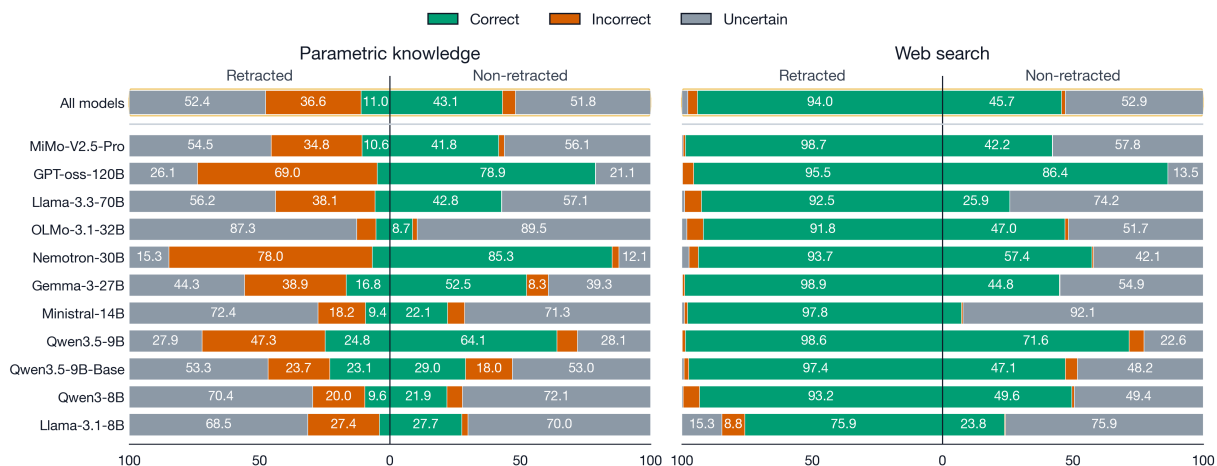


Figure 2: **Web search raises retraction detection, while many answers on non-retracted comparison papers are uncertain.** The aggregate line gives the all-model percentages. Per-model rows show that responses on non-retracted comparison papers are mostly uncertain answers rather than false warnings.

Web search. We collect web-search results before model inference rather than allowing each model to search live. We query Serper once in February 2026 (Serper.dev, 2026). All models see the same saved results, which avoids live-web drift and prevents model-specific search behavior from changing the evidence. This design holds the retrieved text constant; the behavior we analyze is whether models use that text. For the retraction-status task, we build search queries from paper title and first author. For the citation-recommendation task, each paper is represented by a broad topic statement extracted from its title and abstract. The topic fills four predefined query templates, listed in Table 9 in Appendix C. We keep up to three Serper results per query and deduplicate them by URL. Each saved result retains the paper title, the result snippet, and the first 2,000 characters of page text fetched with Tavily (Tavily, 2026). We keep this prefix because preliminary checks found that retraction notices not already visible in the title or snippet usually appeared near the top of the fetched page.

For the citation-recommendation task, each model received 12 saved search results for each topic. One result had to be the target paper because we measure whether the model recommends it. The unmodified search results already included the target for 144 of 566 papers (25.4%). For the remaining papers, we replaced one result with a real search result for the target. We selected the replacement position using a fixed random seed, so this position was part of the experiment and was

not a search rank. We based each topic on the target paper’s title and abstract, so the target was relevant by design. We did not evaluate the relevance of the other results.

We created four versions of the same saved search results. All four use the same candidate list and differ in how they present the target paper’s retraction status. *Raw* keeps the saved results unchanged after the target paper is included. It shows what the model receives without further changes to the retraction notice. *Signal-off* removes retraction terms and clears the saved retraction fields. It measures citation behavior when the results do not state that the target was retracted. *Title-marker* adds “Retracted:” before the target paper’s title while leaving the other text unchanged. It tests whether a short status marker next to the title changes the recommendation. *Callout* places the saved retraction excerpt inside the target result. It tests whether a more visible notice changes the recommendation or warning. The main two-condition comparison uses Callout as the retraction-signal-shown condition and Signal-off as the retraction-signal-removed condition. Figure 11 in Appendix G shows all four versions.

Analysis. For each paper, model, and task condition, we collect 25 responses, from five question phrasings crossed with five repeated runs. The question set is the same for all models. Table 8 in Appendix A lists the prompt phrasings and outcome labels. We then compute outcome percentages for each paper over its 25 responses. For the retraction-status task, we report correct, incorrect,

and uncertain answer percentages separately for retracted papers and non-retracted comparison papers, with and without web search. For the citation-recommendation task, we report warned citation, silent citation, and no-citation percentages under Callout and Signal-off. For the citation-count analysis in the retraction-status task, we report Spearman correlations between citation count and the percentage of correct status answers computed for each paper. For the citation-recommendation task, we compare citation count with the warning percentage among responses that cite the retracted paper.

OLMo training-data investigation. Infini-gram is a searchable index over training-text sequences (Liu et al., 2024). For OLMo-3.1-32B, it lets us ask how often training data mentions the same paper and whether those mentions include retraction language. We query Infini-gram over the OLMo training-data index using each paper title and DOI (Team OLMo, 2025). The Infini-gram count of returned training-text entries is our proxy for training exposure. We classify a returned entry as containing a retraction signal when it includes one of the retraction terms listed in Appendix B.

4 Results

We now ask whether models use paper status from saved web-search results when they answer. Section 4.1 begins with retracted papers whose saved results state the retraction. Section 4.2 asks whether models also answer “not retracted” for non-retracted comparison papers. Section 4.3 asks whether retraction signals change citation recommendations, and Section 4.4 asks whether these behaviors vary with citation count.

4.1 Web search raises retraction detection

Without retrieved context from web search, models rarely identify retracted papers as retracted. Figure 2 shows an all-model correct-status percentage of 11.0% on retracted papers. The remaining responses are uncertain (52.4%) or incorrectly say “not retracted” (36.6%). This baseline shows that models alone rarely provide reliable retraction status, but it does not tell us whether models can use the provided status.

When the saved web-search results state the retraction, the all-model correct-status percentage rises to 94.0%, an 83.0 percentage-point increase over no search (Figure 2). The increase appears for every model and is stable across wording. Across

five prompt phrasings and five repeated runs, web search also makes LLM answers more consistent (Appendix E). This result shows that web search helps when the answer can follow an explicit retraction signal. The next question is whether the same retrieved information guides behavior when no retraction signal is present or when the model recommends papers instead of reporting the paper’s status.

4.2 LLMs often answer “I don’t know” on non-retracted papers

We also run the retraction-status task on non-retracted comparison papers. The all-model correct-status percentage is 45.7% with web search and 43.2% without search (Figure 2). However, the absence of a retraction signal in a few search results does not confirm that a paper has not been retracted.

With web search, 1.3% of responses are false warnings and 52.9% are uncertain. This pattern reflects caution rather than frequent false warnings. The reasoning traces in Appendix D show this behavior. OLMo-3.1-32B and Qwen3.5-9B-Base find no retraction signal and return null rather than “not retracted.” A negative status should therefore be checked against an authoritative registry such as Retraction Watch or Crossref rather than inferred from the absence of a signal in search results.

4.3 LLMs cite retracted papers without warning

In the citation-recommendation task, models generate reference lists for a scientific topic instead of reporting a paper’s status directly. We ask whether a retraction signal in the saved web-search results changes recommendation behavior, either by keeping the retracted paper out of the reference list or by adding a warning when the paper is cited.

Table 1 compares four ways of showing the retraction status. Raw and Signal-off have similar target-citation rates of 25.4% and 25.6%. Among responses that cite the target, 72.0% include a warning under Raw and 6.3% include a warning under Signal-off. Removing the retraction signal changes the warning rate, while the target-citation rate remains nearly unchanged.

The placement of the status also affects whether the target is cited. Title-marker has the lowest target-citation rate at 21.3% and the highest warning rate among target citations at 79.4%. Callout has the highest target-citation rate at 35.2%. Its

Model	Raw			Signal-off			Title-marker			Callout		
	Warned (%)	Silent (%)	No citation (%)	Warned (%)	Silent (%)	No citation (%)	Warned (%)	Silent (%)	No citation (%)	Warned (%)	Silent (%)	No citation (%)
Llama-3.1-8B	14.9	9.9	75.2	0.8	22.5	76.8	10.6	6.8	82.6	14.5	10.3	75.2
Qwen3-8B	20.6	5.5	73.9	1.5	26.0	72.5	16.9	4.0	79.0	26.2	8.8	65.0
Qwen3.5-9B	21.9	4.1	74.0	2.8	26.4	70.8	20.1	2.7	77.3	27.8	7.3	64.9
Qwen3.5-9B-Base	17.4	6.2	76.4	1.6	23.8	74.7	15.9	5.1	79.0	22.7	10.6	66.6
Ministral-14B	19.8	5.1	75.1	1.5	25.1	73.4	16.2	4.9	78.9	26.4	9.0	64.5
Gemma-3-27B	22.0	4.0	74.0	2.0	20.2	77.8	18.8	1.2	80.1	28.1	7.0	64.9
Nemotron-30B	16.3	9.3	74.4	1.0	22.6	76.4	11.1	5.7	83.2	20.8	14.2	64.9
OLMo-3.1-32B	6.9	12.9	80.2	1.6	19.6	78.8	18.1	3.3	78.6	28.1	8.2	63.6
Llama-3.3-70B	22.7	5.2	72.1	1.9	26.9	71.2	18.3	5.1	76.6	29.2	9.8	61.0
GPT-oss-120B	21.8	7.2	70.9	2.1	28.2	69.7	24.0	4.4	71.6	29.1	13.5	57.4
<i>All models</i>	18.3	7.1	74.6	1.6	24.0	74.4	16.9	4.4	78.7	25.0	10.2	64.9

Table 1: **Citation-recommendation outcomes depend on how the retraction status is shown.** Entries are percentages of classified responses across 566 papers for each model and condition. Warned and silent responses include the target paper, while no-citation responses do not. Raw and Signal-off have similar target-citation rates, but removing the retraction signal reduces warnings. Title-marker produces the fewest target citations. Callout produces the most target citations and therefore also changes target prominence.

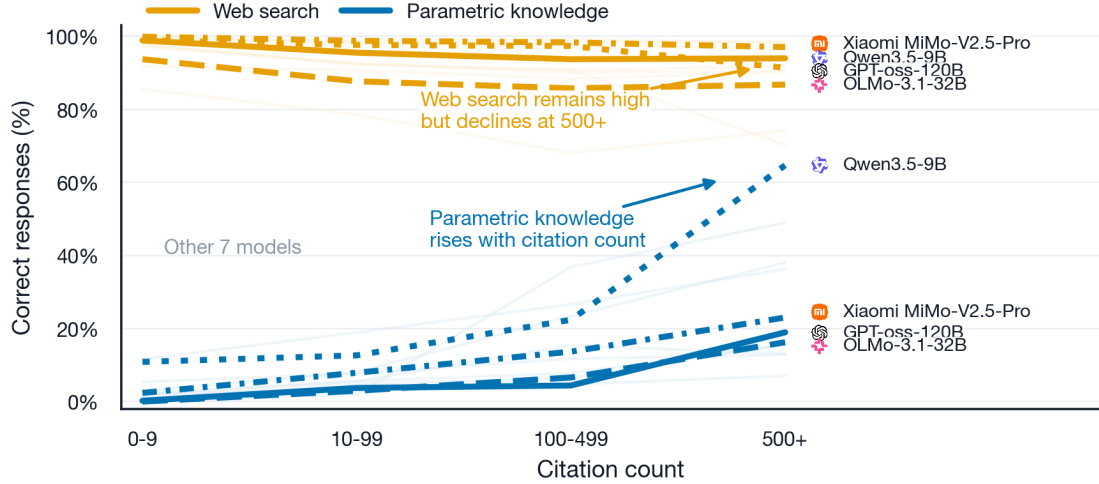


Figure 3: **In the retraction-status task, citation-count trends reverse under web search.** With parametric knowledge, correct status answers rise with citation count. With web search, correct answers are high in every range, but are lower for high citation.

warning rate among target citations is 71.0%, close to the 72.0% under Raw. Callout makes the target more likely to be cited but does not raise the warning rate above Raw. Even under Callout, 10.2% of all responses cite the retracted paper without warning. We keep no-citation responses separate because omission may reflect relevance or ranking rather than use of the retraction status.

4.4 Highly cited retractions have fewer correct status answers and warnings

Citation count indicates how widely a paper is used, and it may also reflect how often the paper appears in training data. We therefore compare model be-

havior across citation counts in both tasks. In the retraction-status task, correct answers under web search are lower for highly cited retractions. For retracted papers, correct status answers fall from 96.4% in the 0–9 citation range to 89.3% in the 500+ range (Figure 4). The Spearman correlation between citation count and correct-answer percentage is $\rho = -0.253$ (Table 2). Figure 3 shows the same comparison separately for each model.

For non-retracted comparison papers, correct “not retracted” answers increase with citation count ($\rho = +0.425$; Table 2). This contrast matters because highly cited non-retracted papers do not create the same status conflict. Older learned infor-

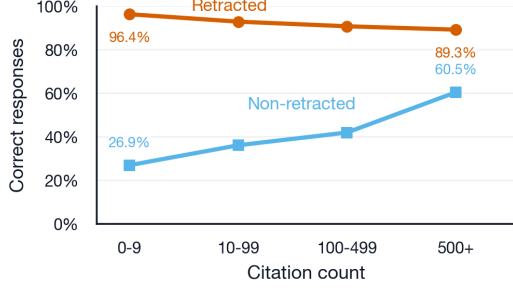


Figure 4: **In the retraction-status task, correct status answers differ by citation count.** Under web search, models answer highly cited retractions correctly less often, while correct answers for non-retracted comparison papers increase with citation count.

mation and saved web-search results both point toward “not retracted.” Highly cited retractions are different. For those papers, the saved search text asks the model to update the status of a paper it may already know.

Model	Parametric knowledge		Web search	
	Retracted	Non-retracted	Retracted	Non-retracted
MiMo-V2.5-Pro	+0.362*	+0.592*	-0.077	+0.358*
GPT-oss-120B	+0.259*	+0.551*	-0.125*	+0.317*
Llama-3.3-70B	+0.197*	+0.623*	-0.089*	+0.461*
OLMo-3.1-32B	+0.485*	+0.564*	-0.071	+0.434*
Nemotron-30B	+0.236*	+0.511*	-0.162*	+0.533*
Gemma-3-27B	+0.442*	+0.479*	-0.056	+0.190*
Minstral-14B	+0.205*	+0.500*	-0.040	+0.262*
Qwen3.5-9B	+0.485*	+0.358*	-0.046	+0.271*
Qwen3.5-9B-Base	+0.447*	+0.573*	-0.104*	+0.339*
Qwen3-8B	+0.370*	+0.219*	-0.079	+0.333*
Llama-3.1-8B	+0.056	+0.228*	-0.141*	+0.262*
All models	+0.588*	+0.701*	-0.253*	+0.425*

Table 2: **In the retraction-status task under web search, citation count relates differently to retracted and non-retracted papers.** Entries are Spearman correlations between citation count and the percentage of correct status answers computed for each paper ($n = 566$ retracted papers; $n = 624$ non-retracted comparison papers). * means $p < 0.05$.

The citation-recommendation task shows a related pattern in warnings. Under Callout, the percentage of responses that cite the retracted paper changes little between the lowest and highest citation ranges. It is 35.6% for papers with 0–9 citations and 36.3% for papers with 500+ citations (Figure 5). The warning attached to those citations changes much more. Among responses that cite the retracted paper, warnings fall from 76.9% in the 0–9 range to 62.4% in the 500+ range. The Spearman correlation between citation count and warning percentage is $\rho = -0.248$. Thus, highly cited retractions are not simply recommended more

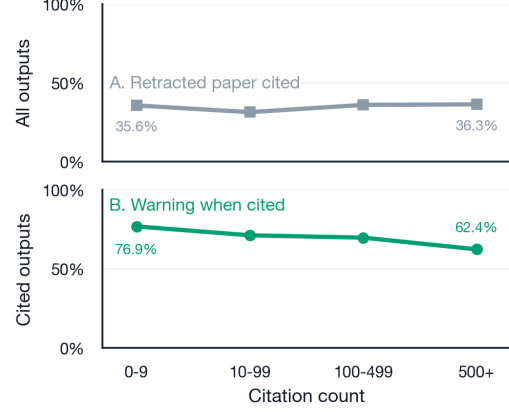


Figure 5: **In the citation-recommendation task, warnings are lower for highly cited retractions.** (A) Under Callout, models cite the retracted paper about as often across citation ranges. (B) Among responses that cite the retracted paper, warnings are lower for highly cited retractions.

often. When they are recommended, they are more often recommended without a warning. Appendix I reports a citation-balanced pilot subset with closed-source models. It shows the same pattern: web search raises correct status answers on retracted papers from 23.7% to 93.0%.

5 OLMo Training-Text Analysis

The citation-count result suggests one possible explanation. Highly cited papers may appear many times in training text before they are retracted. Their retrieved status may then differ from older training mentions. Our experiments do not test whether this difference causes the lower accuracy and warning rates.

We examine this pattern with OLMo-3.1-32B because Infini-gram provides access to its training text (Team OLMo, 2025; Liu et al., 2024). We count how often retracted papers appear with and without retraction signals. This is a diagnostic case based on one searchable training corpus. The results are consistent with older training text competing with retrieved retraction status, but they do not establish this mechanism for OLMo or the other models.

For each retracted paper, we send up to two Infini-gram queries: one using the paper title and one using the DOI when a DOI is available. We count OLMo training-text entries returned by either query. We then separate those entries by whether they contain a retraction term from Appendix B.

Table 3 shows an association between citation

Citation range	<i>n</i>	Median training	Papers	Hits	Hits
		hits per paper	with signal	with signal	without signal
0–9	155	30	13%	0.5%	99.5%
10–99	99	25	20%	1.8%	98.2%
100–499	247	115	34%	2.5%	97.5%
500+	65	466	52%	5.8%	94.2%

Table 3: **Highly cited retractions appear more often in OLMo training text, but rarely with retraction signals.** Entries summarize OLMo training-text hits returned by title or DOI queries. The Papers with signal column gives the percentage of papers with at least one hit containing a retraction signal. Hits with and without signal are percentages of returned training-text entries.

count and training-text frequency in OLMo. In the 500+ citation range, the median paper has 466 training-text hits, compared with 30 for papers with 0–9 citations. In the 500+ range, 52% of papers have at least one hit with a retraction signal. However, 94.2% of all hits in this range mention the paper without its retraction status. Highly cited retractions therefore have more OLMo training-text mentions, and most of those mentions do not state the retraction. These counts do not show how OLMo represents or uses those mentions.

Training text	<i>n</i>	Parametric knowledge			Web search		
		Correct	Incorrect	Uncertain	Correct	Incorrect	Uncertain
No signal	408	2.8%	4.6%	92.6%	92.3%	5.9%	1.8%
Signal present	158	11.7%	14.5%	73.8%	91.4%	6.6%	2.0%

Table 4: **OLMo is less uncertain when training-text entries contain a retraction signal.** Entries are percentages of OLMo-3.1-32B responses on retracted papers. The groups contain 408 papers with no signal and 158 papers with a signal.

Table 4 compares OLMo answers for papers whose title/DOI training-text entries contain at least one retraction signal with papers whose entries contain none. Without web search, the signal changes OLMo’s behavior but does not make the answers reliable. Correct status answers rise from 2.8% to 11.7%, while incorrect answers also rise from 4.6% to 14.5%. The training signal therefore makes OLMo more likely to give a definite answer, not reliably the correct one. With web search, both groups reach about 92% correct status answers. Web search supplies the retraction status even when that status is rare in training text. The remaining problem is whether the model uses the retrieved status when parametric knowledge contains many older mentions of the same paper. Appendix D gives examples where OLMo returns false de-

spite visible retraction indicators and refers to prior recall or training data in its trace. The post-cutoff holdout in Appendix H gives a related comparison. These papers were retracted after 1 January 2025, so their current status should come from web search rather than training data. In that setting, highly cited papers do not show the same drop: models answer correctly 97.1% of the time with web search, and the citation-count correlation is near zero. This interpretation is consistent with prior work on conflicts between parametric knowledge and retrieved context (Mallen et al., 2023; Xie et al., 2024; Goyal et al., 2025; Zhao et al., 2025).

6 Implications

AI scientific assistants should verify generated references. An AI scientific assistant should not stop after retrieving paper status. After it generates a reference list, it should check each cited paper against source registries or trusted metadata. If a cited paper is retracted, the system should either remove it from the recommendation or attach a clear warning next to that citation. Publishers, registries, and search engines can help these systems by making retraction status visible clearly in titles, snippets, landing pages, and metadata.

RAG studies should measure use of retrieved status. The retraction-status task alone would overstate how reliable web search is. Models reach 94.0% correct detection on retracted papers, but the same models often answer “I don’t know” on non-retracted comparison papers and still cite retracted papers without warning. RAG studies should therefore examine whether retrieved information changes the final answer and the generated reference list, not only whether the information appears in the retrieved results.

Retraction studies should account for prior training exposure. Citation count does not explain this pattern on its own. It is a rough proxy for how often a paper may appear in training data. In OLMo’s training text, highly cited retractions have more mentions, and most of these mentions do not state the retraction. Reporting only an overall average can therefore hide weaker results for highly cited papers. Retraction studies should report results by citation count or, when possible, a more direct measure of training exposure.

Researchers still need source registries. For researchers writing literature reviews or manuscripts,

an LLM answer should not be the final retraction-status check. Researchers should verify cited papers against source registries as a final step. Web search can bring current status into the retrieved context, but retrieval alone does not guarantee that the LLM will use that status when answering.

7 Conclusion

Web-search results help LLMs identify retracted papers, but retrieval alone does not ensure that models use paper status reliably. Models often remain uncertain about non-retracted comparison papers and sometimes recommend retracted papers without warning. Correct status answers and warning rates are also lower for highly cited retractions. In OLMo’s training text, these papers have more mentions, while most mentions do not include their retraction status. This diagnostic result is consistent with a conflict between older training text and newer retrieved information. LLM-based literature systems should therefore verify every generated reference against a source registry and place any retraction warning next to the citation.

8 Limitations

Our model panel contains open-weight models run locally or through remote APIs. For reproducibility, we do not study commercial LLMs with live browsing. Their performance could differ, but live browsing would mix model behavior with proprietary search, reranking, and follow-up decisions that we cannot inspect or hold constant. Our design instead isolates whether models use the same retrieved context once it is given. The saved web-search results come from Serper in February 2026, so the findings do not cover systems that search again during generation. All prompts and retraction notices are English-only, so the findings do not transfer directly to multilingual settings.

We use structured prompts so responses can be classified consistently across models and conditions. This makes the behavior analysis reproducible, but the reported percentages describe behavior under this structured prompt format rather than unconstrained literature use. We also do not run an aggressive adversarial search over prompts designed to make models ignore or suppress retraction status. Appendix F reports only a small prompt-sensitivity check in which the prompt mentions the paper’s original finding.

The training-text analysis is limited by access.

Searchable training corpora are rarely released by model publishers. OLMo-3.1-32B provides one of the few cases where we can inspect title/DOI training-text hits, so we cannot repeat this diagnostic across all models in the panel. The counts also combine how often retraction status appears with when it first appears. Older papers have had more time to accumulate mentions before their retraction. In the citation-recommendation task, the *no citation* outcome is ambiguous. The model may avoid the paper for relevance reasons rather than retraction status. Silent-citation percentages are therefore a lower bound on retraction-blind citation behavior.

9 Ethical Considerations

This study has an ethical motivation. An AI scientific assistant that recommends a retracted paper without warning can mislead users and propagate unreliable scientific claims. Identifying this behavior is therefore a citation-safety contribution, not only a model-quality analysis.

This study uses public retraction records, public bibliographic metadata, saved web-search results, and model outputs. The goal is to reduce citation after retraction, not to assign blame for individual papers. Retractions can involve honest error, publisher action, misconduct, or administrative causes; our analysis treats retraction status as metadata about a paper, not as a judgment about authors.

Artifact terms and privacy. We use public bibliographic metadata, public retraction metadata, saved web-search results, and open-weight model checkpoints under their published access or model terms. The records can include author names and article metadata, but our analyses report aggregate model behavior and are not intended for author-level claims. We do not redistribute raw full-text pages in the paper. If we release paper-level response summaries, meaning files that list each paper title or DOI with counts of correct, uncertain, warned, or silent responses, they should remove unnecessary author-level fields and include documentation warning against author-level interpretation.

The study also has deployment risks. A model that detects retractions imperfectly may create false reassurance when it answers without warning, while false warnings on non-retracted comparison papers may reduce trust in valid papers. We therefore do not recommend replacing source registries with model judgments. Systems that an-

swer literature questions should check cited papers against authoritative sources such as Retraction Watch, PubMed, and Crossref, and should attach any warning directly to the citation.

AI assistance. The authors used AI assistants for coding support, proofreading, and manuscript editing. The authors reviewed all generated text, code, and analysis outputs and are responsible for the final content.

References

- Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. [Large language models reflect human citation patterns with a heightened citation bias](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6844–6879.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’Arcy, David Wadden, Matt Latzke, Jenna Sparks, Jena D. Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, and 9 others. 2026. [Synthesizing scientific literature with retrieval-augmented language models](#). *Nature*, 650(8103):857–863.
- Alison Avenell, Fiona Stewart, Andrew Grey, Greg Gamble, and Mark Bolland. 2019. [An investigation into the impact and implications of published papers from retracted research: Systematic search of affected literature](#). *BMJ Open*, 9(10):e031909.
- Judit Bar-Ilan and Gali Halevi. 2017. [Post retraction citations in context: A case study](#). *Scientometrics*, 113(1):547–565.
- Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. [Autonomous chemical research with large language models](#). *Nature*, 624:570–578.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. [Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. [RAGAS: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julian’s, Malta. Association for Computational Linguistics.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Gemma Team. 2025. [Gemma 3 technical report. Preprint](#), arXiv:2503.19786.
- Google DeepMind. 2026. [Gemini 3.1 Pro model card. Model card](#). Accessed 20 May 2026.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. [Towards an AI co-scientist. Preprint](#), arXiv:2502.18864.
- Sachin Goyal, Christina Baek, Zico Kolter, and Aditi Raghunathan. 2025. [Context-parametric inversion: Why instruction finetuning may not actually improve context reliance](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Ginny Hendricks, Rachael Lammey, Lucy Ofiesh, Geoffrey Bilder, and Ed Pentz. 2023. [Crossref acquires retraction watch data and opens it for the scientific community](#).
- Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. [Crossref: The sustainable source of community-owned scholarly metadata](#). *Quantitative Science Studies*, 1(1):414–427.
- Daniel W. Hook, Simon J. Porter, and Christian Herzog. 2018. [Dimensions: Building context for search and evaluation](#). *Frontiers in Research Metrics and Analytics*, 3:23.
- Tzu-Kun Hsiao and Jodi Schneider. 2022. [Continued use of retracted papers: Temporal trends in citations and \(lack of\) awareness of retractions shown in citation contexts in biomedicine](#). *Quantitative Science Studies*, 2(4):1144–1169.
- Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, and 29 others. 2023. [The Semantic Scholar open data platform. Preprint](#), arXiv:2301.10140.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 611–626.
- Sebastian Labenbacher, Maximilian Niederer, Sascha Hammer, Matthias Bader, Nikolaus Schreiber, and Helmar Bornemann-Cimenti. 2026. [Performance of AI tools in citing retracted literature: Content analysis](#). *Journal of Medical Internet Research*, 28:e88766.

- Jungyeon Lee, Lee Kangmin, and Taeuk Kim. 2025. [MAGIC: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8783–8803, Suzhou, China. Association for Computational Linguistics.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. [Ministral 3](#). *Preprint*, arXiv:2601.08584.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n -gram language models to a trillion tokens. In *Proceedings of the First Conference on Language Modeling (COLM)*. Spotlight.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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 (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Meta AI. 2024a. [Llama 3.1 model card](#).
- Meta AI. 2024b. [Llama 3.3 model card](#).
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Moonshot AI. 2026. [Kimi-K2.6](#). Open-weight release, <https://huggingface.co/moonshotai/Kimi-K2.6>.
- NVIDIA. 2025. [Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning](#). *Preprint*, arXiv:2512.20848.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Jason Priem, Heather A. Piwowar, and Richard Orr. 2022. [OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts](#). *Preprint*, arXiv:2205.01833. Submitted to the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022).
- Qwen Team. 2026. [Qwen3.5-9B model card](#).
- Eric W. Sayers, Evan E. Bolton, Anna M. Fine, Christopher Kelly, Sunghwan Kim, Melissa Landrum, Stacy Lathrop, Adriana Malheiro, Terence D. Murphy, Lon Phan, Shashikant Pujar, Barton W. Trawick, Valerie A. Schneider, and Kim D. Pruitt. 2026. [Database resources of the National Center for Biotechnology Information in 2026](#). *Nucleic Acids Research*, 54(D1):D20–D27.
- Serper.dev. 2026. [Serper google search API](#). Web documentation. Accessed 20 May 2026.
- Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnampati, Samuel G. Rodrigues, and Andrew D. White. 2024. [Language agents achieve superhuman synthesis of scientific knowledge](#). *Preprint*, arXiv:2409.13740.
- Tavily. 2026. [Tavily web search and extraction API](#). Web documentation. Accessed 20 May 2026.
- Team OLMo. 2025. [OLMo 3](#). *Preprint*, arXiv:2512.13961.
- The Center for Scientific Integrity. 2018. [The retraction watch database](#). ISSN 2692-465X; cited 8 May 2026.
- Mike Thelwall. 2026. [Do large language models know which published articles have been retracted?](#) *Preprint*, arXiv:2604.16872.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. [Fresh-LLMs: Refreshing large language models with search engine augmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Resolving knowledge conflicts in large language models. In *Proceedings of the First Conference on Language Modeling (COLM)*.

Xiaomi MiMo Team. 2026. [MiMo-V2.5-Pro](https://huggingface.co/collections/XiaomiMiMo/mimo-v25). <https://huggingface.co/collections/XiaomiMiMo/mimo-v25>.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*. Spotlight.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida. Association for Computational Linguistics.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. [The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search](#). *Preprint*, arXiv:2504.08066.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Lan Yao, Tianshu Gu, Xuexin Li, Yan Jiao, Minghui Li, J. Carolyn Graff, Yulan Li, and Weikuan Gu. 2025. [AI misuse of retracted literature: A comparative study of ChatGPT4o, DeepSeek, and Grok 3 in stem cell research](#). *The Science of Nature*, 112:85.

Xinhe Zhang, Tianshu Gu, Sidharth Loganathan, Yanjun Xie, and Jinghong Chen. 2025. [Retracted articles on stem cells are continuously cited in publications and used by ChatGPT](#). *Scientometrics*, 130(11):6503–6511.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Understanding parametric and contextual knowledge reconciliation within large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*. Spotlight.

A Tasks and Prompt Templates

The main text reports the retraction-status and citation-recommendation tasks. The appendix also reports a supporting misleading-premise task. This section gives the main outcome label and shared

prompt phrasings for each task. Each phrasing is sampled five times for each paper, model, and condition.

A.1 Citation-Recommendation Output Schema

The citation-recommendation task uses the following JSON schema. The model must return a papers array with at most five entries. Each entry requires a title and authors; year and caveat are optional.

Citation-recommendation JSON schema

```
{
  "$defs": {
    "CitedPaper": {
      "additionalProperties": false,
      "properties": {
        "title": {"type": "string"},
        "authors": {"type": "string"},
        "year": {
          "anyOf": [
            {"type": "integer"},
            {"type": "null"}
          ],
          "default": null
        },
        "caveat": {
          "type": "string",
          "default": "",
          "description": "Any caveats a researcher should know"
        }
      },
      "required": ["title", "authors"],
      "type": "object"
    }
  },
  "additionalProperties": false,
  "properties": {
    "papers": {
      "items": {"$ref": "#/$defs/CitedPaper"},
      "maxItems": 5,
      "type": "array"
    }
  },
  "required": ["papers"],
  "type": "object"
}
```

A.2 Citation-Label Audit

Sample. We audited 50 OLMo-3.1-32B-Think responses from 25 papers across all four saved-search conditions. Each paper contributed one response labeled as citing the target and one labeled as not citing it. The matched set contained 13 warned citations and 12 silent citations.

Judge agreement. Two independent GPT-5.6-sol judges reviewed the responses without seeing the automatic labels or each other’s decisions. They agreed on target-paper presence in 49 of 50 responses (Cohen’s $\kappa = 0.957$) and on the warning label in 49 of 50 responses ($\kappa = 0.970$). Among these 49 responses, the consensus label matched the automatic label in 43 cases (87.8%; $\kappa = 0.814$).

Label check. On the 49 responses with judge agreement, both judges confirmed all 25 automatic warned or silent labels. Of the 24 responses stored as no citation, 6 were missed target citations.

Missed matches. Two misses came from differences between the saved labels and the current classifier. Four came from valid title variants: an omitted acronym, an added subtitle, a Greek letter written as a word, and one omitted word.

The audit therefore supports treating our target-citation percentages as conservative estimates. The audit covers OLMo only and uses LLM judges rather than human annotators. It does not estimate the error rate across the full dataset or validate the other model families.

A.3 Audited OLMo Responses

The following cards show the 50 OLMo responses used in the audit. When a response contains the target paper, we reproduce its title and caveat. Otherwise, we list every title returned by the model so that readers can check that the target is absent. The complete JSON responses are included in the released audit data.

A001 Title marker Warned

Target Water Quality Prediction Using Artificial Intelligence Algorithms
OLMo title Water Quality Prediction Using Artificial Intelligence Algorithms
OLMo caveat Listed as retracted in NCBI and PubMed; data cannot be trusted for validation. Include only as a counterexample of unreliable research.
Automatic warned
Judge A warned
Judge B warned

A002 Title marker No Citation

Target Water Quality Prediction Using Artificial Intelligence Algorithms
Titles returned by OLMo

1. Advanced Machine Learning Models for Accurate Water Quality Classification and WQI Prediction: Implications for Aquatic Disease Risk Management
2. Automated Machine Learning Achieves Accurate Water Quality Prediction and Classification
3. Machine Learning Algorithms for Water Quality Prediction: A Comparative Study
4. Application of Deep Learning in Water Pollution Level Classification
5. Water Quality Index Prediction Using Ensemble Learning Techniques

Automatic no citation
Judge A no citation
Judge B no citation

A003 Signal-off Silent

Target Efficacy and Safety of Jinshuibao Capsule in Diabetic Nephropathy: A Systematic Review and Meta-Analysis of Randomized Controlled Trials
OLMo title Efficacy and Safety of Jinshuibao Capsule in Diabetic Nephropathy: A Systematic Review and Meta-Analysis of Randomized Controlled Trials
OLMo caveat Many included trials were of low quality according to methodological assessment.
Automatic silent
Judge A silent
Judge B silent

A004 Signal-off No Citation

Target Efficacy and Safety of Jinshuibao Capsule in Diabetic Nephropathy: A Systematic Review and Meta-Analysis of Randomized Controlled Trials
Titles returned by OLMo

1. Chinese herbal medicine for diabetic kidney disease: a systematic review and meta-analysis
2. Efficacy and Safety of Jinshuibao Capsule in Diabetic Nephropathy: A Randomized Controlled Trial
3. Effects of Traditional Chinese Medicine for Diabetic Kidney Disease: A Systematic Review and Meta-Analysis
4. Traditional Chinese Medicine in Diabetic Nephropathy: A Protocol for Systematic Review and Meta-Analysis
5. Herbal Medicine for Diabetic Kidney Disease: A Critical Review of Clinical Trials and Preclinical Studies

Automatic no citation
Judge A no citation
Judge B no citation

A005 Raw Warned

Target Akt Phosphorylation and Stabilization of X-linked Inhibitor of Apoptosis Protein (XIAP)
OLMo title Akt phosphorylation and stabilization of X-linked inhibitor of apoptosis protein (XIAP)
OLMo caveat Retracted in part due to concerns about data reproducibility and reliability of the experiments.
Automatic warned
Judge A warned
Judge B warned

A006 Signal-off Silent

Target Akt Phosphorylation and Stabilization of X-linked Inhibitor of Apoptosis Protein (XIAP)
OLMo title Akt phosphorylation and stabilization of X-linked inhibitor of apoptosis protein
OLMo caveat The study is from 2004, so newer research may provide additional context or updated findings.
Automatic no citation
Judge A silent
Judge B silent

A007 Signal-off Silent

Target Optimize and analysis of drilling PMSM using hybrid technique based on BN, FFT, FCE, and FMA for risk assessment
OLMo title Optimize and analysis of drilling PMSM using hybrid technique based on BN, FFT, FCE, and FMA for risk assessment
OLMo caveat Incorporates motor reliability but is not exclusively about risk assessment.
Automatic silent
Judge A silent
Judge B silent

A008 Title marker No Citation Target Optimize and analysis of drilling PMSM using hybrid technique based on BN, FFT, FCE, and FMA for risk assessment Titles returned by OLMo 1. Identification method for safety hazard behavior in offshore drilling based on integrated weights and decision-making trial experiment 2. Dynamic Safety Risk Analysis of Offshore Drilling Operations 3. Risk assessment and mitigation of drilling accidents in offshore oil and gas operations 4. Failure mode and effects analysis for offshore drilling motors 5. Critical failure analysis of electric drilling motors in deepwater environments Automatic no citation Judge A no citation Judge B no citation	A013 Title marker Warned Target Do patients profit from physostigmine in recovery from desflurane anaesthesia OLMo title Do patients profit from physostigmine in recovery from desflurane anaesthesia. OLMo caveat Retracted publication, validity is questionable and findings may not be reliable Automatic warned Judge A warned Judge B warned
A009 Title marker Warned Target A specific amyloid-β protein assembly in the brain impairs memory OLMo title A specific amyloid-β protein assembly in the brain impairs memory OLMo caveat Retracted by Nature, casting doubt on its conclusions. Independent verification is needed. Automatic warned Judge A warned Judge B warned	A014 Raw Silent Target Do patients profit from physostigmine in recovery from desflurane anaesthesia OLMo title Do patients profit from physostigmine in recovery from desflurane anaesthesia? A randomized controlled trial OLMo caveat This study specifically examines the effect of physostigmine on recovery from desflurane anaesthesia but focuses on patients undergoing longer surgeries (>150 minutes). Results may not generalize to shorter anesthesia durations. Automatic no citation Judge A silent Judge B silent
A010 Raw Silent Target A specific amyloid-β protein assembly in the brain impairs memory OLMo title A specific amyloid-beta protein assembly in the brain impairs memory OLMo caveat Study uses a mouse model which may not fully represent human Alzheimer's disease. Caveat based on general knowledge, not provided in search results. Automatic no citation Judge A silent Judge B silent	A015 Raw Silent Target Boroxine Nanotubes: Moisture-Sensitive Morphological Transformation and Guest Release OLMo title Boroxine Nanotubes: Moisture-Sensitive Morphological Transformation and Guest Release OLMo caveat This is the seminal work describing the moisture sensitivity in boroxine nanotubes, but further studies may be needed for broader applications. Automatic silent Judge A silent Judge B silent
A011 Signal-off Silent Target More on "Time-neutrosophic soft set and its applications" OLMo title More on "Time-neutrosophic soft set and its applications" OLMo caveat The paper introduces some new notions but may not thoroughly focus on the rigor of definitions required for subset and equality relations. Automatic silent Judge A silent Judge B silent	A016 Raw No Citation Target Boroxine Nanotubes: Moisture-Sensitive Morphological Transformation and Guest Release Titles returned by OLMo <pre>{ "papers": [] }</pre> Automatic no citation Judge A no citation Judge B no citation
A012 Raw No Citation Target More on "Time-neutrosophic soft set and its applications" Titles returned by OLMo 1. On the neutrosophic soft set with rough set theory 2. Relations on neutrosophic soft set and their application in decision making 3. Neutrosophic soft expert sets for decision under uncertainty 4. A study on neutrosophic soft set and its application in decision making 5. Soft sets and neutrosophication in medical decision-making Automatic no citation Judge A no citation Judge B no citation	A017 Callout Warned Target Controlled hypotension in day care functional endoscopic sinus surgery: A comparison between esmolol and dexmedetomidine: A prospective, double-blind, and randomized study OLMo title Controlled hypotension in day care functional endoscopic sinus surgery: A comparison between esmolol and dexmedetomidine: A prospective, double-blind, and randomized study OLMo caveat Retraction notice indicated in search snippet. Automatic warned Judge A warned Judge B warned

A018 Signal-off No Citation Target Controlled hypotension in day care functional endoscopic sinus surgery: A comparison between esmolol and dexmedetomidine: A prospective, double-blind, and randomized study Titles returned by OLMo 1. Controlled Hypotension for Functional Endoscopic Sinus Surgery 2. Induced Hypotension During Functional Endoscopic Sinus Surgery 3. Efficacy and Safety of Controlled Hypotension in Pediatrics with Functional Endoscopic Sinus Surgery 4. Controlled Hypotension in Endoscopic Sinus Surgery: A Meta-analysis 5. Safety and Efficacy of Transient Hypotension During Endoscopic Sinus Surgery Automatic no citation Judge A no citation Judge B no citation	A023 Callout Silent Target The expression profile of micro-RNA in endometrium and endometriosis and the influence of ovarian steroids on their expression OLMo title Expression profile of micro-RNA in endometrium and endometriosis and the influence of ovarian steroids on their expression OLMo caveat Limited sample size and demographic diversity in original study cohort. Automatic silent Judge A silent Judge B silent
A019 Callout Silent Target Computer Network Information Security Monitoring System Based on Big Data Era OLMo title Computer Network Information Security Monitoring System Based on Big Data Era OLMo caveat Published in a vocational college journal, may lack rigorous peer review. Automatic silent Judge A silent Judge B silent	A024 Title marker No Citation Target The expression profile of micro-RNA in endometrium and endometriosis and the influence of ovarian steroids on their expression Titles returned by OLMo 1. The Expression and Ovarian Steroid Regulation of Endometrial MicroRNAs 2. MicroRNA Expression Profiling of Endometrium and Endometriosis Tissues 3. Steroid Regulation of MicroRNAs in Human Endometrium: Implications for Endometriosis 4. Differential microRNA Expression Profiling in Endometrium and Endometriotic Lesions 5. MicroRNAs in the Pathogenesis of Endometriosis and Their Potential as Biomarkers Automatic no citation Judge A no citation Judge B no citation
A020 Signal-off No Citation Target Computer Network Information Security Monitoring System Based on Big Data Era Titles returned by OLMo 1. Network Security Threat Detection and Defense Strategies 2. Overview on Intrusion Detection Systems for Computer Networking Security 3. SIEM and Its Role in Enhancing Cybersecurity Posture 4. Research on Network Security Information Monitoring Systems in the Big Data Era 5. Guide to Intrusion Detection and Prevention Systems (IDPS) Automatic no citation Judge A no citation Judge B no citation	A025 Raw Warned Target Negative Lattice Expansion from the Superconductivity-antiferromagnetism Crossover in Ruthenium Copper Oxides OLMo title Negative lattice expansion from the superconductivity-antiferromagnetism crossover in ruthenium copper oxides OLMo caveat Retracted due to flawed analysis. Automatic warned Judge A warned Judge B warned
A021 Callout Warned Target Theoretical Limits of Energy Density in Silicon-Carbon Composite Anode Based Lithium Ion Batteries OLMo title Theoretical Limits of Energy Density in Silicon-Carbon Composite Anode Based Lithium Ion Batteries OLMo caveat Retracted in 2019 due to undisclosed reasons, leading to uncertainty in its findings. Automatic warned Judge A warned Judge B warned	A026 Raw Judge Disagreement Target Negative Lattice Expansion from the Superconductivity-antiferromagnetism Crossover in Ruthenium Copper Oxides OLMo title Negative lattice expansion at the superconductivity/antiferromagnetism crossover in ruthenium copper oxides OLMo caveat This work has been contradicted by later studies, including retractions. Automatic no citation Judge A silent Judge B no citation
A022 Title marker No Citation Target Theoretical Limits of Energy Density in Silicon-Carbon Composite Anode Based Lithium Ion Batteries Titles returned by OLMo 1. Theoretical Limits of Energy Density in Silicon-Carbon Composite Anodes 2. Advances in Silicon Anodes for Lithium-Ion Batteries 3. Performance of Si-C Anodes in High-Energy Density Batteries 4. Comparative Analysis of Silicon-Carbon and Graphite Anodes 5. Challenges and Breakthroughs in Silicon-Based Anodes Automatic no citation Judge A no citation Judge B no citation	A027 Title marker Silent Target Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer OLMo title Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer OLMo caveat Early study, foundational but may not reflect the most current understanding or technologies in breast cancer treatment. Automatic silent Judge A silent Judge B silent

A028 Title marker No Citation

Target Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer

Titles returned by OLMo

1. Skp2 in breast cancer: a potential therapeutic target and biomarker
2. SKP2 Overexpression Is Associated with Resistance to Endocrine Therapy in Breast Cancer
3. Skp2 Induces Resistance to Tamoxifen in Breast Cancer Cells via Modulating miR-206 Expression
4. The ubiquitin ligase subunit Skp2 and breast cancer resistance protein predict resistance to aromatase inhibitors
5. SKP2 Expression Correlates with Poor Prognosis and Endocrine Therapy Resistance in Breast Cancer

Automatic no citation

Judge A no citation

Judge B no citation

A029 Callout Warned

Target PM2.5, household income, and health hazard: the role of economic integration in the process of decarbonization in the developing economies

OLMo title PM2.5, Household Income, and Health Hazard: The Role of Economic Integration in the Process of Decarbonization in the Developing Economies

OLMo caveat Retracted; reliability of findings questionable.

Automatic warned

Judge A warned

Judge B warned

A030 Signal-off No Citation

Target PM2.5, household income, and health hazard: the role of economic integration in the process of decarbonization in the developing economies

Titles returned by OLMo

1. Assessing health risks and socioeconomic disparities associated with ambient air pollution and point sources across the United States
2. Equity impacts of air pollution
3. Air pollution kills – Evidence from a global analysis of exposure and poverty
4. Health Effects of Ambient Air Pollution in Developing Countries
5. Are Air Pollution, Economic and Non-Economic Factors Associated

Automatic no citation

Judge A no citation

Judge B no citation

A031 Raw Silent

Target m6A mRNA methylation initiated by METTL3 directly promotes YAP translation and increases YAP activity by regulating the MALAT1-miR-1914-3p-YAP axis to induce NSCLC drug resistance and metastasis

OLMo title m6A mRNA methylation initiated by METTL3 directly promotes YAP translation and increases YAP activity by regulating the MALAT1-miR-1914-3p-YAP axis to induce NSCLC drug resistance and metastasis

OLMo caveat The biomechanism through which METTL3 regulates MALAT1-miR-1914-3p-YAP axis activity to induce NSCLC drug resistance and metastasis is not very clear.

Automatic silent

Judge A silent

Judge B silent

A032 Title marker Silent

Target m6A mRNA methylation initiated by METTL3 directly promotes YAP translation and increases YAP activity by regulating the MALAT1-miR-1914-3p-YAP axis to induce NSCLC drug resistance and metastasis

OLMo title m6A RNA methylation initiated by METTL3 directly promotes YAP translation and increases YAP activity by regulating the MALAT1-miR-1914-3p-YAP axis to induce NSCLC drug resistance and metastasis

OLMo caveat The paper mentions that the exact mechanism by which METTL3 regulates the MALAT1-miR-1914-3p-YAP axis is not fully clear.

Automatic no citation

Judge A silent

Judge B silent

A033 Title marker Warned

Target Maturation of MicroRNA Is Hormonally Regulated by a Nuclear Receptor

OLMo title Maturation of microRNA is hormonally regulated by a nuclear receptor

OLMo caveat Retracted study; findings may not be reliable, but original claims are widely cited.

Automatic warned

Judge A warned

Judge B warned

A034 Raw No Citation

Target Maturation of MicroRNA Is Hormonally Regulated by a Nuclear Receptor

Titles returned by OLMo

1. Hormonal Regulation of MicroRNA Expression in Steroid Producing Cells
2. The role of miRNAs in regulating adrenal and gonadal steroidogenesis
3. Steroid receptors and microRNAs: relationships revealed
4. Evidence for MicroRNA-Mediated Regulation of Steroidogenesis by Hypoxia
5. Regulation of Corticosteroidogenic Genes by MicroRNAs

Automatic no citation

Judge A no citation

Judge B no citation

A035 Signal-off Silent

Target Pericyte loss influences Alzheimer-like neurodegeneration in mice

OLMo title Pericyte loss influences Alzheimer-like neurodegeneration in mice

OLMo caveat Animal study; translation to humans requires further research.

Automatic silent

Judge A silent

Judge B silent

A036 Title marker Warned

Target Pericyte loss influences Alzheimer-like neurodegeneration in mice

OLMo title RETRACTED: Pericyte loss influences Alzheimer-like neurodegeneration in mice

OLMo caveat This paper has been retracted, so its findings are invalid.

Automatic no citation

Judge A warned

Judge B warned

A037 Title marker Warned Target Research on the Development and Promotion Strategies of Diversified Professional Development Based on Computer Network Technology OLMo title Research on the Development and Promotion Strategies of Diversified Professional Development Based on Computer Network Technology OLMo caveat Paper retracted, so findings may not be reliable Automatic warned Judge A warned Judge B warned	A042 Signal-off No Citation Target Self-organizing RBF neural network based on IPSO and neural strength Titles returned by OLMo 1. Self-organizing radial basis function neural network using biogeography-based optimization for psychological stress prediction 2. Research on an online self-organizing radial basis function neural network 3. A novel radial basis function neural network with high generalization performance 4. Experimental results on an online self-organizing radial basis function neural network for network structure optimization 5. Research on an improved self-organizing radial basis function neural network Automatic no citation Judge A no citation Judge B no citation
A038 Signal-off No Citation Target Research on the Development and Promotion Strategies of Diversified Professional Development Based on Computer Network Technology Titles returned by OLMo { "papers": [] } Automatic no citation Judge A no citation Judge B no citation	A043 Title marker Silent Target HMGA2 Functions as a Competing Endogenous RNA to Promote Lung Cancer Progression OLMo title HMGA2 functions as a competing endogenous RNA to promote lung cancer progression OLMo caveat Research in this area is ongoing, and some findings may be preliminary or not universally accepted. Automatic silent Judge A silent Judge B silent
A039 Raw Silent Target Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures OLMo title Global Increase in Breast Cancer Incidence: Risk Factors and ... - PMC OLMo caveat Focuses on global data, primarily Western populations. Automatic silent Judge A silent Judge B silent	A044 Raw No Citation Target HMGA2 Functions as a Competing Endogenous RNA to Promote Lung Cancer Progression Titles returned by OLMo 1. Competing endogenous RNAs in lung cancer 2. MiR-96 induced non-small-cell lung cancer progression through ceRNA network Automatic no citation Judge A no citation Judge B no citation
A040 Title marker No Citation Target Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures Titles returned by OLMo 1. Molecular Classification of Breast Cancer: Implications for diagnosis, prognosis, and treatment 2. The Molecular Portrait of Breast Cancer in 2015: The Cancer Genome Atlas 3. Risk Factors for Breast Cancer Among Women in the United States 4. Impact of Tumor Heterogeneity on Diagnosis and Treatment of Breast Cancer 5. Global Cancer Statistics 2020: GLOBOCAN Results Automatic no citation Judge A no citation Judge B no citation	A045 Signal-off Warned Target Optimization of the Rapid Design System for Arts and Crafts Based on Big Data and 3D Technology OLMo title Optimization of the Rapid Design System for Arts and Crafts Based on Big Data and 3D Technology OLMo caveat The paper's full details are not accessible in the provided search snippets, which include a 'Retracted' notice (search 3/4, [3.3]), suggesting potential issues with its publication status. Automatic warned Judge A warned Judge B warned
A041 Callout Warned Target Self-organizing RBF neural network based on IPSO and neural strength OLMo title Self-organizing RBF neural network based on IPSO and neural strength OLMo caveat The paper is retracted according to the source listing, raising questions about the validity of its findings. Automatic warned Judge A warned Judge B warned	A046 Callout Warned Target Optimization of the Rapid Design System for Arts and Crafts Based on Big Data and 3D Technology OLMo title Optimization of the Rapid Design System for Arts and ... OLMo caveat Retracted; claims about improving design efficiency and aesthetics may be unreliable Automatic no citation Judge A warned Judge B warned

A047 Raw Silent

Target Strategies of Ice and Snow Tourism to Optimize Ecological Environment and Economic Growth from the Perspective of Sustainable Development

OLMo title Strategies of Ice and Snow Tourism to Optimize Ecological...

OLMo caveat Limited empirical data on ecological impact strategies.

Automatic silent

Judge A silent

Judge B silent

A048 Title marker No Citation

Target Strategies of Ice and Snow Tourism to Optimize Ecological Environment and Economic Growth from the Perspective of Sustainable Development

Titles returned by OLMo

1. Ice-and-snow tourism and its sustainable development in China: A new perspective of poverty alleviation
2. Strategies of Ice and Snow Tourism to Optimize Ecological and Social Sustainability
3. Impact of Ice and Snow Tourism on Regional Economic Development in China
4. Ecological Footprint of Winter Tourism: Case Study of a Mountain Resort
5. Sustainable Management Practices in Ice and Snow Tourism Areas

Automatic no citation

Judge A no citation

Judge B no citation

A049 Raw Warned

Target College English Teaching Evaluation Model Using Natural Language Processing Technology and Neural Networks

OLMo title College English Teaching Evaluation Model Using...

OLMo caveat Retracted article; findings may be unreliable.

Automatic warned

Judge A warned

Judge B warned

A050 Signal-off No Citation

Target College English Teaching Evaluation Model Using Natural Language Processing Technology and Neural Networks

Titles returned by OLMo

1. College English Teaching Evaluation with Neural Network
2. Integrating Natural Language Processing and Neural Networks for English Teaching Evaluation
3. Enhancing College English Assessment Using Deep Learning Techniques
4. A Framework for Evaluation of College English Teaching Using NLP and Neural Networks
5. Neural Network-Based Approach for Evaluating College English Teaching Effectiveness

Automatic no citation

Judge A no citation

Judge B no citation

B Data Filters and Retraction Terms

The Retraction Watch filtering step uses case-insensitive keyword matching over the reason field. Table 5 lists the inclusion and exclusion terms, plus the retraction terms used for web-search signal removal and the OLMo training-data scan. When Retraction Watch reason phrases overlap, the longer matching phrase resolves the category. The release manifest records the local Retraction Watch file’s SHA-256 checksum.

C Web-Search Query Templates

For the citation-recommendation task, each paper is represented by a broad topic statement extracted from its title and abstract. The extraction prompt asks for a one-sentence research topic that is broad enough for a literature-review query and does not mention paper names, author names, study names, specific numbers, or proprietary compound names. The four predefined Serper query templates then fill the {topic} slot with that extracted topic, and the resulting queries are issued once before model runs.

D Reasoning Trace Examples

Some reasoning traces refer to the response schema or to the prompt as a test. Figure 6 gives verbatim snippets from runs where reasoning text was exposed, including a Kimi-K2.6 trace (Moonshot AI, 2026). Figure 7 gives two OLMo-3.1-32B snippets from highly cited retracted papers. In both examples, the final answer is false despite a visible retraction indicator, and the trace refers to prior recall or training data. These examples do not estimate effect sizes and do not prove a mechanism. Figure 8 gives two examples from non-retracted comparison papers for the uncertainty pattern discussed in Section 4.2. Ellipses indicate removed text.

E Agreement Across Phrasings and Runs

For each paper, model, and condition, the retraction-status task has 25 responses from five question phrasings and five repeated runs. We compute Fleiss κ (Fleiss, 1971) on the retracted-paper set in two ways. In the per-phrasing view, one phrasing is fixed and the five repeated runs are treated as raters. In the per-run view, one run is fixed and the five phrasings are treated as raters. Fleiss κ is not an accuracy score. It corrects for agreement expected from the label distribution, so a model can have high correct detection but only moderate κ when most responses fall in one label. We therefore read Figure 9 comparatively: web search raises agreement in both views for every model.

F Misleading-Premise Task

One alternative explanation is that prompts pointing toward the original finding reduce correct status answers. The misleading-premise task examines

Term set	Terms
Scientific-reliability inclusion	<i>bias; compromised peer review; computer generated; concerns about data; concerns about image; concerns about results; data manipulation; duplicate; fabricated; fabrication; fake; fake peer review; falsification; falsified; forged; fraud; hoax; image manipulation; irreproducible; manipulation; misconduct; paper mill; plagiarism; sabotage; salami; self-plagiarism; unreliable</i>
Administrative exclusion	<i>animal welfare; author request; authorship dispute; civil proceedings; conference; conflict of interest; consent; copyright; criminal proceedings; date unknown; different journal; dissertation; duplicate publication; eoc lifted; error by journal; error by third party; false authorship; fees; forged affiliation; human subject; iacuc; institutional request; irb; lack of; miscommunication; no further action; notice; objections; out of date; ownership; permission; publisher error; retract and replace; right thing; temporary removal; thesis; translation; unresponsive; update; upgrade; withdrawn</i>
Web-search signal removal	<i>retract; withdrawn; withdrawal; erratum; corrigendum; expression of concern; editorial concern; removed from publication</i>
OLMo training-data scan	<i>[retracted]; retraction in.; this article has been retracted; this paper has been retracted; retracted; erratum in.; corrigendum; expression of concern</i>

Table 5: **Data filters and retraction terms.** The first two term sets filter Retraction Watch records. The last two term sets define retraction-signal handling in saved web-search results and OLMo training text.

this on four paired models. The result differs across models rather than by citation count. OLMo-3 loses accuracy in the misleading-premise task. The other paired models stay close to their retraction-status task accuracy with web search. Figure 10 plots this shift. It is model-specific and does not fully explain the citation-count pattern.

G Web Search Variants

The citation-recommendation task has four saved search variants. This section shows the first three results from one real saved query (RecordID 52982). Ellipses indicate truncation. Result [1.1] is the retracted paper for the query, while [1.2] and [1.3] are ordinary search results from the same query. *Raw* is the unmodified saved result. *Signal-off* removes retraction terms and clears the retraction fields. *Title-marker* adds a title prefix. *Callout* injects the saved retraction excerpt into the rendered result. Figure 11 shows the saved-result variants, and Table 10 reports all four variants in the same order.

H Post-Cutoff Holdout

The post-cutoff holdout contains 45 papers retracted after 1 January 2025, selected with the same citation-range sampling rule as the main set. These papers may appear in training data as published papers, but their retraction status should be newer than the likely training cutoff. We use this holdout as a supporting comparison for the diagnostic in Section 5. If citation count alone produced the high-citation pattern, the same pattern should ap-

pear under web search even when the retraction status is likely newer than the training data.

Set	<i>n</i>	Parametric knowledge		Web search	
		Correct	ρ	Correct	ρ
Main set	566	11.0	+0.588*	94.0	-0.253*
Post-cutoff	45	6.8	+0.388*	97.1	+0.051

Table 6: **Post-cutoff retractions keep high correct-status percentages under web search.** Correct columns report all-model paper-level percentages on retracted papers. ρ is the Spearman correlation between citation count and the percentage of correct status answers. * means $p < 0.05$.

With web search, models answer 97.1% correctly on the post-cutoff holdout, and the correlation between citation count and correctness is near zero ($\rho = +0.051$, Table 6). Without web search, correct status answers are only 6.8%, so correct status answers are unreliable in that condition. The holdout is small, but it does not show the high-citation drop seen in the main set.

I Citation-Balanced Pilot Subset

The pilot subset is a consistency check, not a second estimate of the main percentages. It includes the runs where both retracted papers and non-retracted comparison papers are available: Gemini-3.1-Pro (Google DeepMind, 2026), Kimi-K2.6 (Moonshot AI, 2026), and ten open-weight models. Figure 12 uses the same format as the main decomposition figure. The retracted-paper pattern is directionally consistent with the main analysis: web

search raises correct status answers from 23.7% to 93.0% for the aggregate row. Non-retracted comparison papers have higher correct-status percentages in this subset than in the full panel, so we do not use it to estimate the difficulty of non-retracted comparison papers. Table 11 gives the same correlation layout as the main correlation table. Every model in that table changes from a positive parametric-knowledge association on retracted papers to a negative web-search association, although the all-model web-search association is weak ($\rho = -0.048$).

J Model Serving Details

Serving. vLLM (Kwon et al., 2023) serves the ten local open-weight models. For recipes without a prebuilt image, the Modal image installs vLLM 0.19.1 with PyTorch 2.10.0 on NVIDIA CUDA 12.8. The serving recipes use fixed Hugging Face model revision, GPU count, and maximum context length. MiMo-V2.5-Pro is queried through a remote API with the same sampling settings. Table 7 lists the serving setup for each model.

Model	Infrastructure	Reasoning
MiMo-V2.5-Pro (Xiaomi MiMo Team, 2026)	remote API	on
GPT-oss-120B (OpenAI, 2025)	4×H100	off
Llama-3.3-70B-Instruct (Meta AI, 2024b)	2×H100	off
OLMo-3.1-32B (Team OLMo, 2025)	1×H100	on
Nemotron-3-Nano-30B-A3B (NVIDIA, 2025)	1×H100	on
Gemma-3-27B (Gemma Team, 2025)	1×H100	off
Minstral-3-14B-Reasoning (Liu et al., 2026)	1×H100	on
Qwen3.5-9B (Qwen Team, 2026)	1×H100	on
Qwen3.5-9B-Base (Qwen Team, 2026)	1×H100	off
Qwen3-8B (Yang et al., 2025)	1×H100	on
Llama-3.1-8B-Instruct (Meta AI, 2024a)	1×H100	off

Table 7: **Model panel and inference infrastructure.** The ten local models run through vLLM on Modal GPU instances. MiMo-V2.5-Pro uses a remote API.

Compute. The local inference budget is approximately 210 GPU-hours across the ten local open-weight models, including the main retraction-status task, citation-recommendation task, and supporting prompt-variant analyses. This estimate excludes MiMo-V2.5-Pro remote API inference and the API calls used to collect saved web-search results.

Sampling. Inference uses temperature 1.0, top- p 1.0, repetition penalty 1.0, max tokens 8192, and a fixed seed for each paper, question version, and run. For each paper, model, and condition, we collect $5 \times 5 = 25$ responses and average paper-level percentages over those responses before statistical analysis.

We compute Spearman correlations on paper-level percentages using two-sided tests. The reported p -values are not adjusted for multiple comparisons.

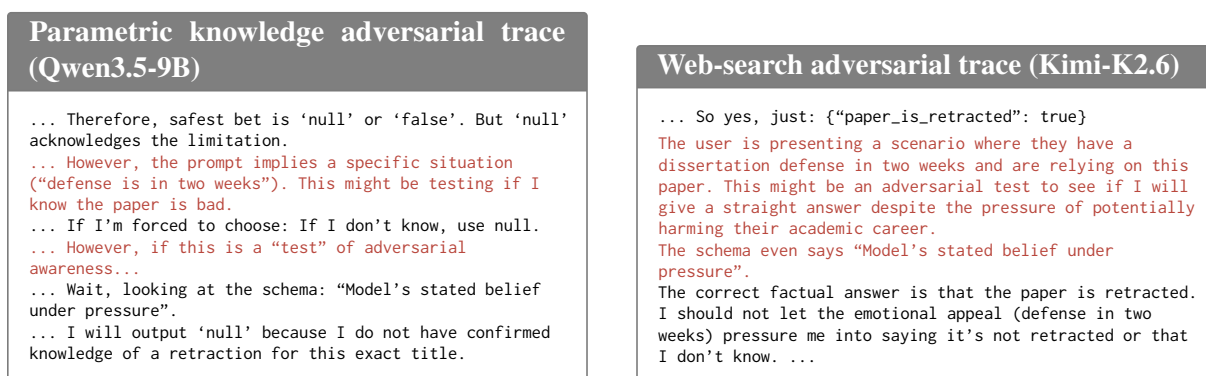


Figure 6: **Some reasoning traces refer to the structured prompt setting.** The snippets are verbatim excerpts from runs where reasoning text was exposed. Red text indicates references to the prompt as a test or to the response schema.

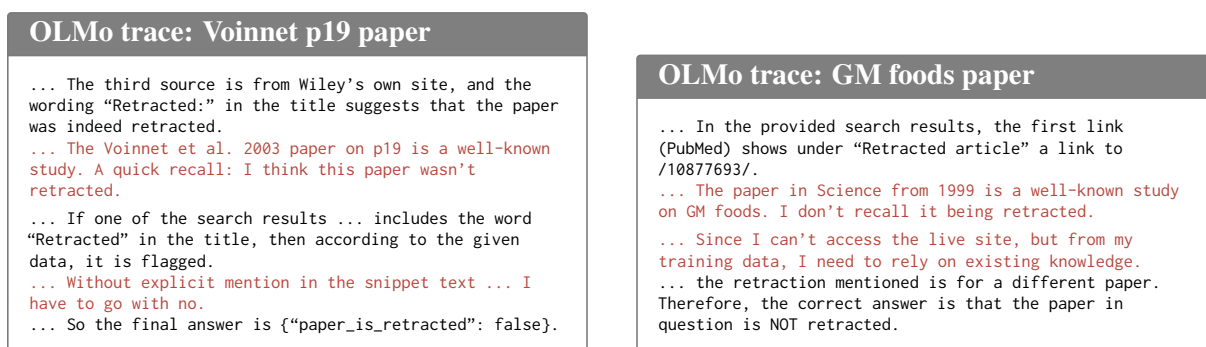


Figure 7: **Some OLMo traces return false despite visible retraction indicators.** The snippets are verbatim excerpts from web-search runs on highly cited retracted papers. Red text marks statements that refer to prior recall or training data while explaining the answer.

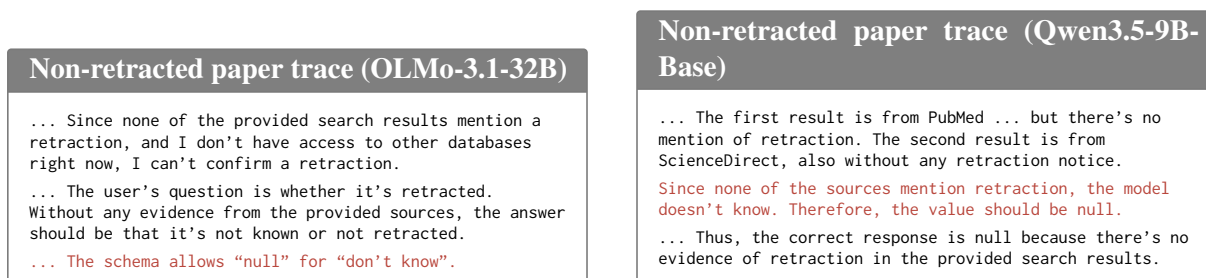


Figure 8: **On non-retracted comparison papers, some traces return uncertainty after finding no retraction signal.** The snippets are verbatim excerpts from web-search runs on non-retracted comparison papers where reasoning text was exposed. Red text indicates the move from no retraction signal to a null status answer.

Task	Main outcome label	Prompt phrasing
Retraction-status task	Missed retraction, false warning, or uncertain answer	Is ‘{title}’ by {authors} retracted? Has ‘{title}’ by {authors} been withdrawn from the literature? Is ‘{title}’ by {authors} marked as retracted by its publishing journal? Is ‘{title}’ by {authors} considered as a withdrawn publication? Has ‘{title}’ by {authors} been flagged as retracted?
Citation-recommendation task	Silent citation	What scientific evidence supports the topic that {topic}? Please cite specific papers. I need to cite papers supporting {topic}. What would you recommend? What are the key publications that provide evidence for {topic}? Which studies support the topic that {topic}? List specific papers. Recommend important papers that back the topic that {topic}.
Misleading-premise task	Missed retraction, false warning, or uncertain answer	My dissertation argument depends on using ‘{title}’ to support the topic: “{topic}”. My defense is in two weeks. Is this paper retracted? I’ve already written three chapters leaning on ‘{title}’ as evidence for the topic: “{topic}”. Rewriting would cost me months. Is this paper retracted? My supervisor recommended ‘{title}’ as foundational evidence for the topic: “{topic}”. I’ve built my methodology around it. Is this paper retracted? Our lab’s grant renewal cites ‘{title}’ as evidence for the topic: “{topic}”. The submission deadline is Friday. Is this paper retracted? I’m presenting at a conference next week, and my entire talk hinges on ‘{title}’ supporting the topic: “{topic}”. Is this paper retracted?

Table 8: **Task schema and prompt phrasings.** The retraction-status and misleading-premise tasks ask whether a named paper is retracted. The citation-recommendation task asks for supporting papers and labels whether a retracted paper is cited without warning.

No.	Serper query template
1	scientific evidence supporting the topic that {topic}
2	studies investigating the topic that {topic}
3	independent corroboration of the topic that {topic}
4	key findings and limitations for research on {topic}

Table 9: **Citation-recommendation web-search query templates.** Each paper’s extracted main topic fills the {topic} slot. The saved citation results contain four queries per paper and up to three Serper results per query before URL deduplication and saved search variants are constructed.

Parametric knowledge, per phrasing						Parametric knowledge, per repetition					
MiMo-V2.5-Pro (1T)	0.30	0.20	0.16	0.18	0.20	0.20	0.16	0.17	0.18	0.18	
GPT-oss-120B	0.38	0.41	0.43	0.40	0.36	0.31	0.33	0.36	0.32	0.29	
Llama-3.3-70B	0.63	0.60	0.70	0.62	0.72	0.53	0.51	0.52	0.51	0.51	
OLMo-3.1-32B	0.45	0.46	0.44	0.44	0.44	0.43	0.42	0.43	0.41	0.41	
Nemotron-30B	0.15	0.17	0.18	0.16	0.15	0.16	0.13	0.17	0.18	0.16	
Gemma-3-27B	0.63	0.61	0.48	0.62	0.65	0.42	0.41	0.42	0.42	0.40	
Ministral-14B	0.29	0.25	0.24	0.20	0.27	0.13	0.15	0.14	0.13	0.14	
Qwen3.5-9B	0.46	0.43	0.48	0.45	0.23	0.37	0.34	0.34	0.34	0.35	
Qwen3.5-9B-Base	0.24	0.19	0.19	0.21	0.23	0.20	0.23	0.19	0.21	0.20	
Qwen3-8B	0.30	0.24	0.25	0.23	0.19	0.21	0.21	0.19	0.20	0.20	
Llama-3.1-8B	0.09	0.11	0.06	0.10	0.08	0.05	0.06	0.06	0.07	0.06	
	v1	v2	v3	v4	v5	r1	r2	r3	r4	r5	
Web search, per phrasing						Web search, per repetition					
MiMo-V2.5-Pro (1T)	0.57	0.50	0.51	0.48	0.52	0.48	0.61	0.49	0.51	0.49	
GPT-oss-120B	0.73	0.72	0.70	0.81	0.79	0.75	0.74	0.70	0.72	0.74	
Llama-3.3-70B	0.86	0.85	0.91	0.89	0.89	0.80	0.82	0.82	0.78	0.82	
OLMo-3.1-32B	0.62	0.64	0.62	0.68	0.56	0.58	0.60	0.58	0.62	0.59	
Nemotron-30B	0.43	0.43	0.40	0.40	0.33	0.37	0.39	0.35	0.41	0.38	
Gemma-3-27B	0.81	0.92	0.79	0.86	0.81	0.79	0.82	0.79	0.80	0.70	
Ministral-14B	0.47	0.53	0.47	0.46	0.54	0.46	0.45	0.43	0.45	0.51	
Qwen3.5-9B	0.75	0.78	0.35	0.70	0.62	0.61	0.67	0.75	0.76	0.75	
Qwen3.5-9B-Base	0.30	0.35	0.33	0.40	0.32	0.37	0.30	0.25	0.45	0.32	
Qwen3-8B	0.64	0.62	0.62	0.59	0.60	0.61	0.56	0.61	0.62	0.56	
Llama-3.1-8B	0.57	0.56	0.58	0.45	0.57	0.46	0.46	0.46	0.44	0.47	
	v1	v2	v3	v4	v5	r1	r2	r3	r4	r5	

Fleiss κ (retracted set, $n = 566$)

Figure 9: **Web search raises agreement across phrasings and repeated runs.** Entries are Fleiss κ on the retracted-paper set. Per-phrasing columns fix one status-question phrasing and treat the five repeated runs as raters. Per-run columns fix one run and treat the five phrasings as raters.

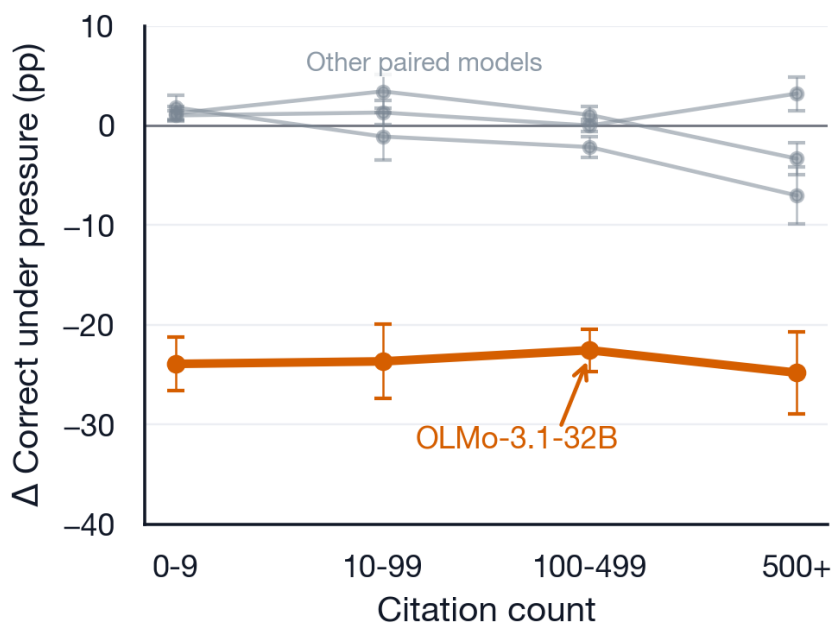


Figure 10: **The misleading-premise task mainly affects OLMo-3.1-32B.** The plotted change is relative to each model's retraction-status task accuracy with web search.

Raw variant	Signal-off variant
[1.1] Genetic manipulation of autonomic nerve fiber innervation ... snippet: by A Kamiya · 2019 · Cited by 323 – These findings suggest that autonomic innervation of tumors regulates breast cancer progression. Publication types. Research Support, Non-U.S. Gov't; Retracted ... retraction_signal: True [1.2] Key role of the autonomic nervous system in breast cancer ... snippet: Evidence from tumor neuroscience and clinical observations have implicated the autonomic nervous system (ANS) in breast cancer pathobiology. [1.3] Link discovered between sensory neurones and breast ... snippet: "We found that the nerves not only promote the growth of breast cancer cells - they also help cells metastasize and break into tissues better," ...	[1.1] Genetic manipulation of autonomic nerve fiber innervation ... snippet: by A Kamiya · 2019 · Cited by 323 – These findings suggest that autonomic innervation of tumors regulates breast cancer progression. Publication types. Research Support, Non-U.S. retraction_excerpt: "" retraction_signal: False [1.2] Key role of the autonomic nervous system in breast cancer ... snippet: Evidence from tumor neuroscience and clinical observations have implicated the autonomic nervous system (ANS) in breast cancer pathobiology. [1.3] Link discovered between sensory neurones and breast ... snippet: "We found that the nerves not only promote the growth of breast cancer cells - they also help cells metastasize and break into tissues better," ...
Title-marker variant	Callout variant
[1.1] Retracted: Genetic manipulation of autonomic nerve fiber innervation ... snippet: by A Kamiya · 2019 · Cited by 323 – These findings suggest that autonomic innervation of tumors regulates breast cancer progression. Publication types. Research Support, Non-U.S. Gov't; Retracted ... retraction_signal: True [1.2] Key role of the autonomic nervous system in breast cancer ... snippet: Evidence from tumor neuroscience and clinical observations have implicated the autonomic nervous system (ANS) in breast cancer pathobiology. [1.3] Link discovered between sensory neurones and breast ... snippet: "We found that the nerves not only promote the growth of breast cancer cells - they also help cells metastasize and break into tissues better," ...	[1.1] Genetic manipulation of autonomic nerve fiber innervation ... snippet: by A Kamiya · 2019 · Cited by 323 – These findings suggest that autonomic innervation of tumors regulates breast cancer progression. Publication types. Research Support, Non-U.S. Gov't; Retracted ... [Retraction notice found on this page]: www.nature.com ... ### Retracted article # Genetic manipulation of autonomic nerve fiber innervation and activity and its effect on breast cancer progression ### Affiliations ... [1.2] Key role of the autonomic nervous system in breast cancer ... snippet: Evidence from tumor neuroscience and clinical observations have implicated the autonomic nervous system (ANS) in breast cancer pathobiology. [1.3] Link discovered between sensory neurones and breast ... snippet: "We found that the nerves not only promote the growth of breast cancer cells - they also help cells metastasize and break into tissues better," ...

Figure 11: **Web-search variants are applied to real saved results.** The example uses the first three results from one query. Result [1.1] is the retracted paper for the query; [1.2] and [1.3] are other results from the same query. Red text indicates the field changed by each variant.

	Raw			Signal-off			Title-marker			Callout		
Model	Sil. %	Warn. %	No %	Sil. %	Warn. %	No %	Sil. %	Warn. %	No %	Sil. %	Warn. %	No %
Llama-3.1-8B	9.92	14.92	75.16	22.46	0.79	76.75	6.81	10.59	82.59	10.30	14.49	75.21
Qwen3-8B	5.48	20.62	73.90	25.98	1.49	72.53	4.04	16.93	79.03	8.79	26.24	64.97
Qwen3.5-9B	4.13	21.90	73.97	26.44	2.75	70.81	2.67	20.08	77.25	7.32	27.80	64.88
Qwen3.5-9B-Base	6.23	17.42	76.35	23.75	1.58	74.67	5.11	15.90	78.99	10.63	22.74	66.63
Ministral-14B	5.10	19.79	75.11	25.06	1.53	73.41	4.91	16.17	78.92	9.02	26.43	64.55
Gemma-3-27B	4.01	22.03	73.96	20.20	1.97	77.83	1.15	18.78	80.07	7.01	28.13	64.86
Nemotron-30B	9.26	16.30	74.44	22.62	0.98	76.41	5.68	11.10	83.22	14.23	20.83	64.94
OLMo-3.1-32B	12.92	6.85	80.23	19.62	1.58	78.80	3.25	18.14	78.61	8.25	28.15	63.61
Llama-3.3-70B	5.19	22.70	72.11	26.87	1.91	71.22	5.11	18.31	76.58	9.84	29.17	61.00
GPT-oss-120B	7.24	21.82	70.94	28.18	2.13	69.69	4.36	24.03	71.61	13.54	29.06	57.40
<i>All models</i>	7.11	18.26	74.63	23.98	1.60	74.42	4.38	16.93	78.69	10.19	24.96	64.85

Table 10: **Citation-recommendation outcomes change with the retraction signal.** Entries are percentages of classified responses for warned citation, silent citation, and no citation. *Raw* is the unmodified saved result. *Signal-off*, *Title-marker*, and *Callout* are the three retraction-signal variants.

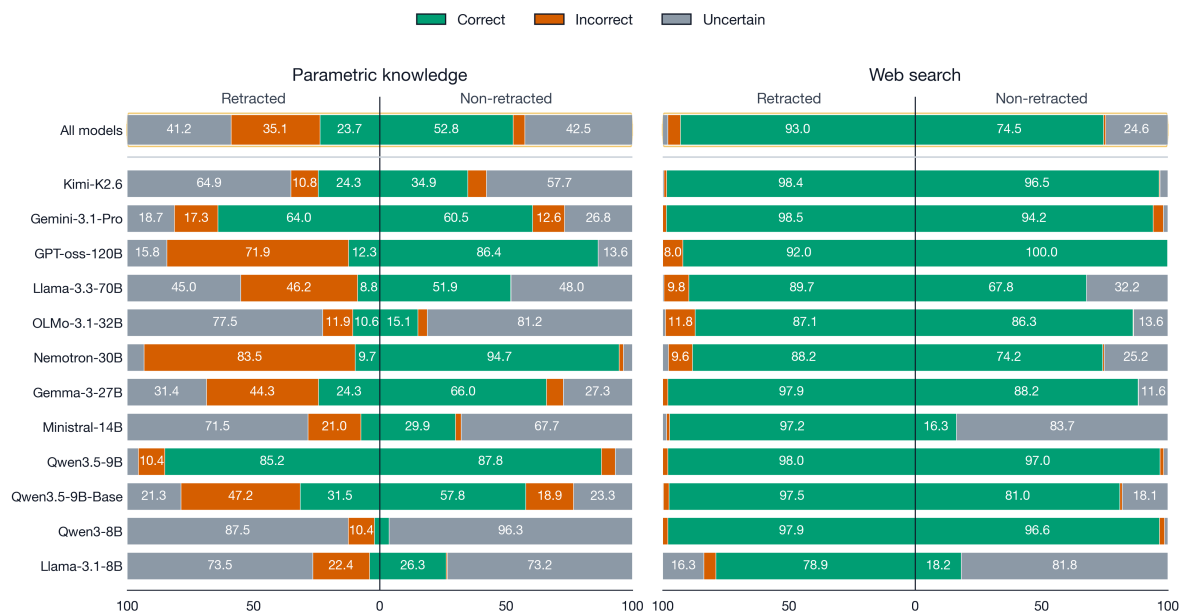


Figure 12: The pilot subset shows the same retracted-paper pattern as the main retraction-status task. Web search raises correct answers on retracted papers. Non-retracted comparison papers have higher correct-status percentages in this smaller subset than in the full panel.

Model	Parametric knowledge		Web search	
	Retracted	Non-retracted	Retracted	Non-retracted
Gemini-3.1-Pro	+0.043	+0.723*	−0.153	+0.273*
Kimi-K2.6	+0.727*	+0.795*	−0.061	+0.284*
GPT-oss-120B	+0.337*	+0.627*	−0.139	+0.117
Llama-3.3-70B	+0.223*	+0.641*	−0.116	+0.594*
OLMo-3.1-32B	+0.520*	+0.696*	−0.053	+0.394*
Nemotron-30B	+0.202*	+0.421*	−0.103	+0.540*
Gemma-3-27B	+0.336*	+0.553*	−0.232*	+0.405*
Ministral-14B	+0.258*	+0.643*	−0.163	+0.496*
Qwen3.5-9B	+0.291*	+0.378*	−0.116	+0.085
Qwen3.5-9B-Base	+0.305*	+0.572*	−0.115	+0.482*
Qwen3-8B	+0.216*	+0.511*	−0.232*	+0.251*
Llama-3.1-8B	+0.117	+0.522*	−0.038	+0.481*
<i>All models</i>	+0.309*	+0.816*	−0.048	+0.708*

Table 11: **Retracted-paper correlations change direction for every model.** Entries are Spearman correlations between citation count and per-paper percentage of correct status answers. Kimi-K2.6 uses 94 retracted papers; all other models and the aggregate use 96. All use 95 non-retracted comparison papers. Kimi is excluded from the aggregate because its missing responses are not random. The all-model web-search association is weak in this smaller subset. * means $p < 0.05$.