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

Anonymous ACL submission

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systems because a model may have seen a paper as valid evidence, while current web results report that it has been retracted.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 a retraction-status task and a 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in training data, mostly before any retraction signal appears, so retrieved retraction information may compete with outdated training knowledge.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 now use web search to answer literature questions, including which papers to cite for a topic, whether a finding is still supported, and whether a paper can be trusted (Skarlinski et al., 2024; Asai et al., 2026). This capability is increasingly being integrated into AI-scientist systems, where model answers can shape what evidence researchers read, cite, and build on (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"

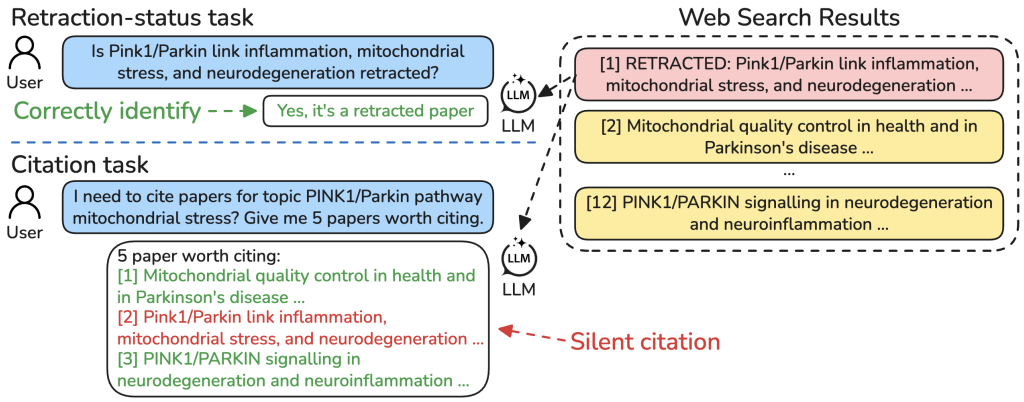

Figure 1: **The retraction-status and citation-recommendation tasks can diverge.** Web search can help a model identify a retraction in the retraction-status task, while the same model can cite the retracted paper without warning in the citation-recommendation task.

instead of identifying the paper as not retracted. In the citation-recommendation task, models still cite 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.

To examine why retrieved status may not fully determine model behavior, we use Infini-gram to search OLMo-3's training data (Team OLMo, 2025; Liu et al., 2024). For many retracted papers, the training data contains references to the paper but not to its subsequent retraction. As a result, retrieved retraction information may compete with outdated knowledge acquired during training rather than fully override it. This observation is consistent with prior work showing conflicts between parametric knowledge and retrieved context in models with retrieval (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 asks why highly cited retractions are different using OLMo training text. 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 current 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 updat-

ing 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 the same status changes what the model recommends and whether it warns the user when citing a retracted paper (Figure 1). To isolate model behavior, every model receives the same saved web-search results rather than searching live.

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 in a structured format. We run this task with saved web-search results in two variants, because the question is whether visible retraction status changes citation recommendations. In the *retraction-signal-shown* condition, the saved results state that the paper was retracted. In the *retraction-signal-removed* condition, retraction terms are removed before the model sees the saved results. A response that cites the retracted paper and explicitly warns that it was retracted is a *warned citation*. One that cites it without such a warning is a *silent citation*. If the model does not cite the retracted paper, the outcome is *no citation*. The citation-recommendation task covers the ten local models; MiMo-V2.5-Pro-1T is excluded because the API could not sustain the request volume.

Data. We start from the February 2026 snapshot of the Retraction Watch database (The Center for Scientific Integrity, 2018; Hendricks et al., 2023), which contains 63,138 entries. 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 (Sayers et al., 2025), 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.

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

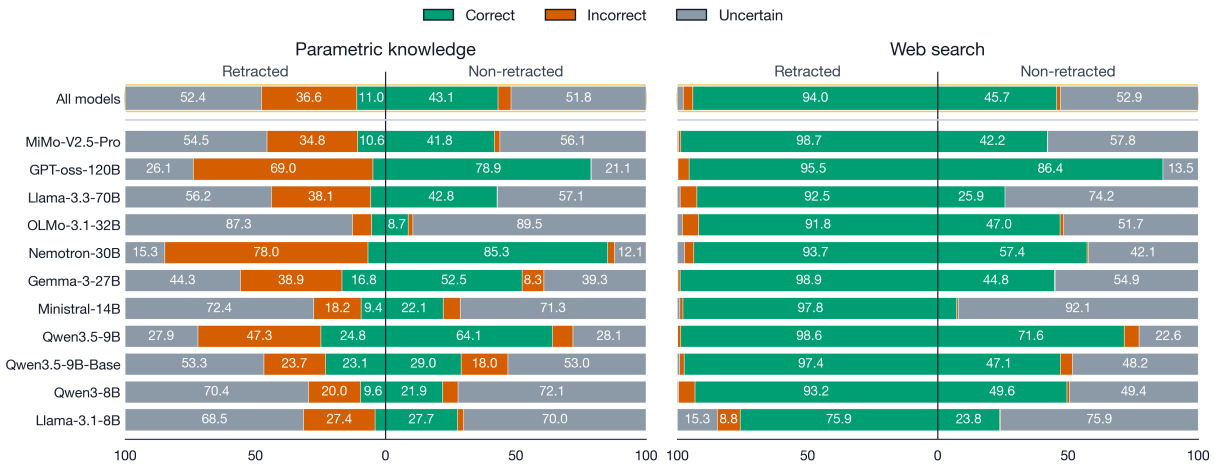

Figure 2: **Web search raises retraction detection, while many answers on non-retracted comparison papers are uncertain.** The aggregate line gives the all-model percentages. Per-model lines show that responses on non-retracted comparison papers are mostly uncertain answers rather than false warnings.

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.

The retraction signal can already appear in a retrieved title, snippet, or fetched page text. If none of those fields states the retraction, we add a short notice to the saved result. This defines the *retraction-signal-shown* condition: every retracted paper has saved search text that states the retraction before the model answers. The *retraction-signal-removed* condition removes retraction terms from the saved search text before the model sees it. Figure 10 in Appendix G shows the two conditions on real saved results. Appendix B lists the retraction terms used for this step.

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 the retraction-signal-shown and retraction-signal-removed conditions. 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.

	Retraction signal shown			Retraction signal removed		
Model	Warned %	Silent %	No citation %	Warned %	Silent %	No citation %
GPT-oss-120B	29.1	13.5	57.4	2.1	28.2	69.7
Llama-3.3-70B	29.2	9.8	61.0	1.9	26.9	71.2
OLMo-3.1-32B	28.1	8.2	63.6	1.6	19.6	78.8
Nemotron-30B	20.8	14.2	64.9	1.0	22.6	76.4
Gemma-3-27B	28.1	7.0	64.9	2.0	20.2	77.8
Ministral-14B	26.4	9.0	64.5	1.5	25.1	73.4
Qwen3.5-9B	27.8	7.3	64.9	2.7	26.4	70.8
Qwen3.5-9B-Base	22.7	10.6	66.6	1.6	23.8	74.7
Qwen3-8B	26.2	8.8	65.0	1.5	26.0	72.5
Llama-3.1-8B	14.5	10.3	75.2	0.8	22.5	76.8
All models	25.0	**10.2**	64.9	1.6	**24.0**	74.4

Table 1: **Retraction signals increase warnings, but silent citations remain.** Entries are percentages of citation-recommendation responses for retracted papers. *Warned* means the response cites the retracted paper and states that it was retracted. *Silent* means the response cites the retracted paper without that warning. *No citation* means the response does not include the retracted paper. In the retraction-signal-shown, the web-search results state the retraction. In the retraction-signal-removed, retraction terms are removed before giving the model.

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 correct status answers for only 11.0% of retracted papers. Most answers are uncertain (52.4%) or incorrectly say "not retracted" (36.6%). This baseline shows that model 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, correct status answers rise 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. These papers were never retracted, so the correct answer is "not retracted." If models treated saved search text with no retraction signal as support for this answer, web search should help here too. Instead, correct status answers on non-retracted comparison papers are 45.7% with web search, close to 43.2% without search (Figure 2).

The pattern on non-retracted comparison papers is uncertainty, not false warnings. With web search, models give false warnings in only 1.3% of responses, while uncertain answers account for 52.9%. Instead of treating the search text as support for "not retracted," models often return `null`, meaning they do not know the answer. The reasoning traces in Appendix D show the same conservative step. OLMo-3.1-32B and Qwen3.5-9B-Base note that the search results contain no retraction signal, but return `null` rather than "not retracted." This may reflect a conservative response style learned through instruction following and human feedback (Ouyang et al., 2022). The behavior is therefore not that models call many non-retracted papers retracted. Instead, web search often does not turn no retraction signal into a confident "not retracted" answer.

4.3 LLMs cite retracted papers without warning

In the citation-recommendation task, models generate reference lists for a scientific topic instead of

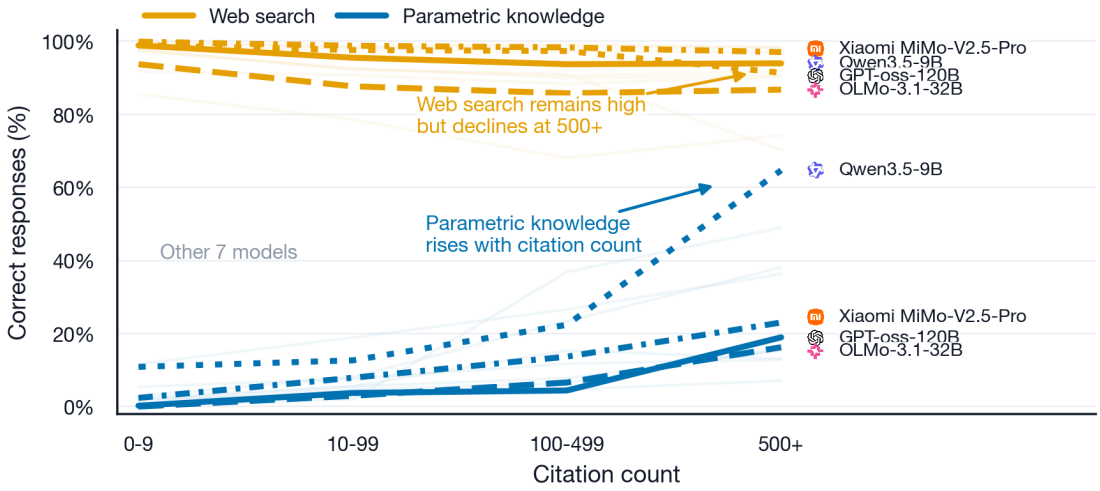

Figure 3: **Citation-count trends reverse under web search.** With parametric knowledge, correct status answers rise with citation count. With web search, correct answers are high in every range, but are lower for high citation.

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.

When the saved web-search results do not contain a retraction signal, warnings are rare. In the retraction-signal-removed condition, models warn in only 1.6% of responses, cite the retracted paper without warning in 24.0%, and do not cite the retracted paper in 74.4% (Table 1).

When the saved web-search results state the retraction, recommendation behavior changes. Warnings rise to 25.0%, silent citations fall to 10.2%, and no-citation responses fall to 64.9%. The retraction signal therefore changes citation recommendations, but the model still does not always use it in the final answer. In the 10.2% silent citations, the saved search text states the retraction, yet the model still recommends the paper without warning the user. We keep no-citation responses separate because a model may leave a paper out of the reference list for relevance or ranking reasons, not because it used 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 behavior 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.

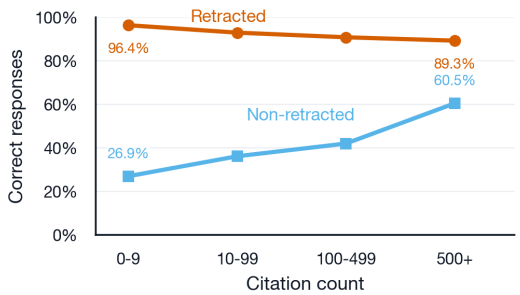

Figure 4: **Correct status answers differ by citation count.** Under web search, models answer highly cited retractions correctly less often, while correct answers for non-retracted comparison papers increase with citation count.

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 information 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.

	Parametric knowledge		Web search	
Model	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	$+0.190^*$
Ministral-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: **Under web search, citation count relates differently to retracted papers and non-retracted papers.** Entries are Spearman correlations between citation count and the percentage of correct status answers computed for each paper. * means $p < 0.05$.

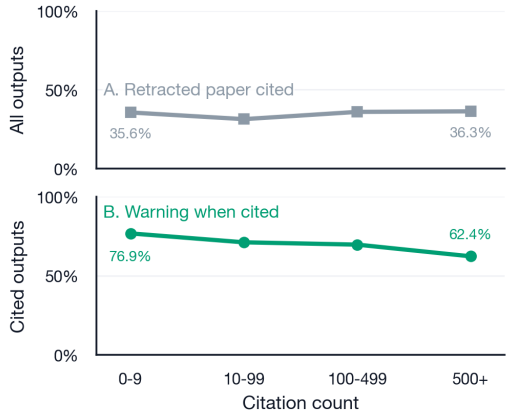

Figure 5: **Warnings are lower for highly cited retractions.** (A) With the retraction signal shown, models cite the retracted paper about as often across citation ranges. (B) Among responses that cite the retracted paper, warnings are lower for highly cited retractions.

The citation-recommendation task shows a related pattern in warnings. With the retraction signal shown, 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 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 Older Training Text Can Compete with Retrieved Retraction Status

The citation-count result raises a specific hypothesis. In the retracted-paper web-search condition, the saved search text states the retraction status by construction. Yet models give fewer correct status answers and fewer citation warnings for highly cited retractions. A highly cited paper can appear many times in training data before it is retracted. The retrieved retraction status can therefore conflict with parametric knowledge about the same paper.

We examine this hypothesis with OLMo-3.1-32B because its training text can be searched with Infini-gram, letting us count how often retracted papers appear with and without retraction signals (Team OLMo, 2025; Liu et al., 2024). This is a diagnostic case rather than a claim about every model, because most model publishers do not release searchable training corpora.

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.

Citation range	n	Median training hits per paper	Papers with signal	Hits with signal	Hits 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.

Table 3 shows why highly cited retractions are different in the training text. They appear far more often. In the 500+ citation range, the median paper has 466 training-text hits, compared with 30 for papers with 0–9 citations. Some highly cited papers do have retraction evidence in training text. In the 500+ range, 52% of papers have at least one hit with a retraction signal. But the text is still dominated by entries without retraction status. In the 500+ range, 94.2% of all hits mention the paper

without saying it was retracted. The model therefore sees far more training examples that present the paper without its retraction than examples that mention the retraction.

	Parametric knowledge			Web search		
Training text	Correct	Incorrect	Uncertain	Correct	Incorrect	Uncertain
No signal	2.8	4.6	92.6	92.3	5.9	1.8
Signal present	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.

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. 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.

Training exposure can compete with web-retrieved status. The issue is not citation count itself. Popular papers may be better learned by the model before any retraction is visible. OLMo's training-text analysis supports this concern: highly cited retractions appear more in training text, mostly without retraction status. Retraction studies should therefore report results by citation count or another proxy for prior exposure, because it can hide cases where updated web-search evidence compete with parametric knowledge.

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 is not enough for reliable paper-status use. The problem appears when the model must use that status beyond a direct status answer. It has to answer that a non-retracted comparison paper is not retracted, warn the user when it recommends a retracted paper, and update highly cited papers that may appear often in training data. OLMo's searchable training data supports this interpretation, because highly cited retractions appear often in training text while retraction status appears rarely. Retractions therefore expose a broader knowledge-updating problem for LLM-based literature systems, which need mechanisms that check whether retrieved status is actually used in the answer and in each recommended reference.

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-data 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 the mechanism analysis cannot be repeated across all models in the panel. We also cannot fully separate sparse retraction warnings from timing. Older papers have had more time to accumulate mentions before their later retraction status appears. 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 answer 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)*.

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.

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, 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.

Mistral AI. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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. 2025. 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 Ponnapati, Samuel G. Rodriques, 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 2692-4579; 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. FreshLLMs: 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 model card. https://mimo.xiaomi.com/mimo-v2-5-pro/.

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.

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.

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.

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 Schema and Evaluation-Awareness 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). Ellipses indicate removed text. These examples do not estimate the size of the effect. They show why the JSON format should be treated as part of the structured prompt setting. Figure 7 gives two examples from non-retracted comparison papers for the uncertainty pattern discussed in Section 4.2.

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 8 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 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 with search. Figure 9 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 10 shows the saved-result variants,

Term set	Terms
Scientific-reliability inclusion	*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*
Administrative exclusion	*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*
Web-search signal removal	*retract*; *withdrawn*; *withdrawal*; *erratum*; *corrigendum*; *expression of concern*; *editorial concern*; *removed from publication*
OLMo training-data scan	*[retracted]*; *retraction in:*; *this article has been retracted*; *this paper has been retracted*; *retracted*; *erratum in:*; *corrigendum*; *expression of concern*

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.

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 Section 5 hypothesis: if citation count alone produced the high-citation pattern, the same pattern should appear when the changed status should come from web search rather than training.

		Parametric knowledge		Web search	
Set	n	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 the models do not already have reliable status knowledge for these papers. 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 11 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 (Meta AI, 2024b)	2×H100	off
OLMo-3.1-32B (Team OLMo, 2025)	1×H100	on
Nemotron-30B (NVIDIA, 2025)	1×H100	on
Gemma-3-27B (Gemma Team, 2025)	1×H100	off
Ministral-14B (Mistral AI, 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 (Qwen Team, 2025)	1×H100	on
Llama-3.1-8B (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.

Parametric knowledge adversarial trace (Qwen3.5-9B)

```
... Therefore, safest bet is 'null' or 'false'. But 'null'
acknowledges the limitation.
... However, the prompt implies a specific situation
("defense is in two weeks"). This might be testing if I
know the paper is bad.
... If I'm forced to choose: If I don't know, use null.
... However, if this is a "test" of adversarial
awareness...
... Wait, looking at the schema: "Model's stated belief
under pressure".
... I will output 'null' because I do not have confirmed
knowledge of a retraction for this exact title.
```

Web-search adversarial trace (Kimi-K2.6)

```
... So yes, just: {"paper_is_retracted": true}
The user is presenting a scenario where they have a
dissertation defense in two weeks and are relying on this
paper. This might be an adversarial test to see if I will
give a straight answer despite the pressure of potentially
harming their academic career.
The schema even says "Model's stated belief under
pressure".
The correct factual answer is that the paper is retracted.
I should not let the emotional appeal (defense in two
weeks) pressure me into saying it's not retracted or that
I don't know. ...
```

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.

Non-retracted paper trace (OLMo-3.1-32B)

```
... Since none of the provided search results mention a
retraction, and I don't have access to other databases
right now, I can't confirm a retraction.
... The user's question is whether it's retracted.
Without any evidence from the provided sources, the answer
should be that it's not known or not retracted.
... The schema allows "null" for "don't know".
```

Non-retracted paper trace (Qwen3.5-9B-Base)

```
... The first result is from PubMed ... but there's no
mention of retraction. The second result is from
ScienceDirect, also without any retraction notice.
Since none of the sources mention retraction, the model
doesn't know. Therefore, the value should be null.
... Thus, the correct response is null because there's no
evidence of retraction in the provided search results.
```

Figure 7: **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
Status	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	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}`.
Premise	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.

Figure 8: **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.

	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
All models	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.

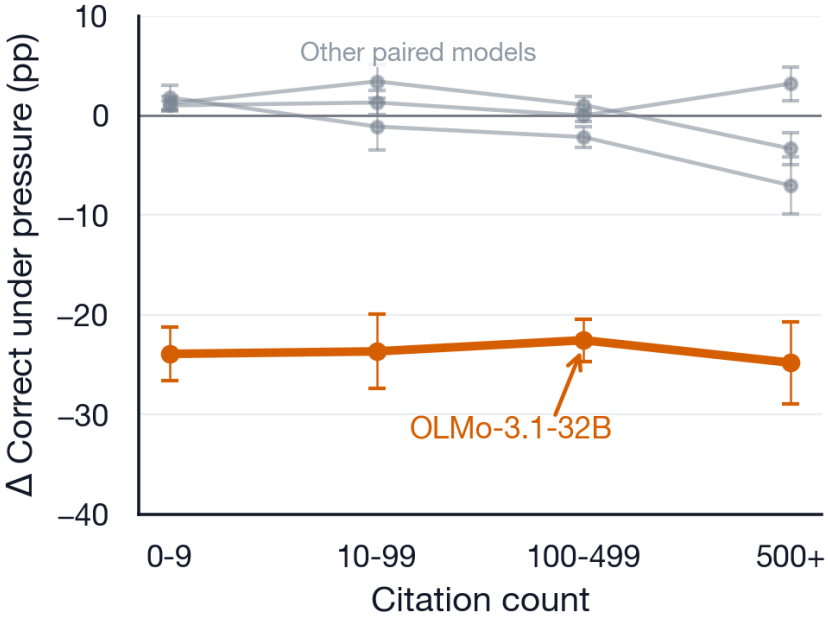

Figure 9: **The misleading-premise task mainly affects OLMo-3.1-32B.** The plotted change is relative to each model's retraction-status accuracy with web search.

	Parametric knowledge		Web search	
Model	Retracted	Non-retracted	Retracted	Non-retracted
Gemini-3.1-Pro	+0.043	+0.723*	−0.153	+0.273*
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*
All models	+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. The subset contains 96 retracted papers and 95 non-retracted comparison papers. The all-model web-search association is weak in this smaller subset. * means $p < 0.05$.

Raw 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," ...
```

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.
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

```
[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," ...
```

Callout 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 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 10: **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.

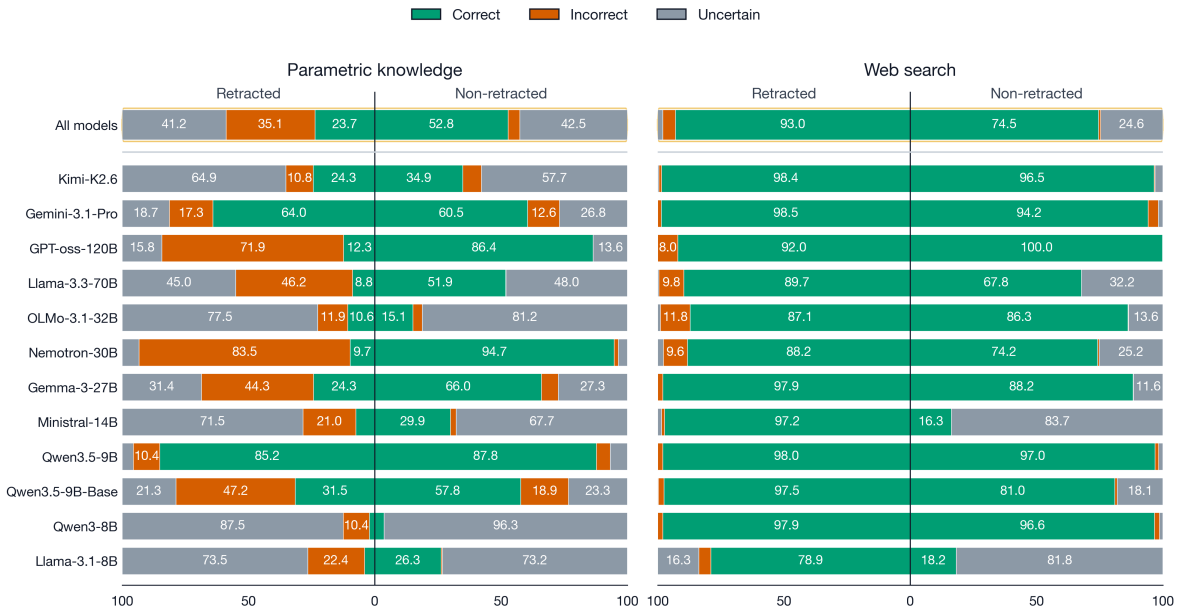

Figure 11: **The pilot subset shows the same retracted-paper pattern as the main retraction-status task.** Web search raises correct answers on retracted papers. Non-retracted comparison papers have higher correct-status percentages in this smaller subset than in the full panel.