AI Often Cites Web Pages It Has Never Even Read

Originally published Friday 16 October 2026 at https://www.unite.ai/ai-often-cites-web-pages-it-has-never-even-read/



Sometimes a paper emerges from the day's literature stream that really gets my blood up – not because it is inaccurate or offensive, but because it notices and studies something that has been annoying me at the periphery of my awareness for a long time, in a vague and non-specific way (perhaps because I was too busy to do anything but ignore it).

Such a paper has come out today, from three authors at the Department of Computing at Imperial College London. Titled *Retrieved but Not Read: Understanding Why Search-Augmented LLMs Fail to Use Available Evidence*, the new work investigates why it is that Large Language Models (LLMs) frequently throw supporting URLs at you (irrelevant paper citations, links to spurious web articles, etc.) *without even reading them*.

The authors state:

> *'We find that models often answer from search snippets without opening the source pages, which is associated with higher rates of outdated answers and unfaithful citations. Our controlled experiments and layer-wise analysis reveal a gap between the availability of evidence and its effective use.*
>
> *'Even when retrieved evidence influences internal predictions, models may still favour outdated information in their final answers.'*

The authors tested one proprietary LLM (OpenAI's GPT-5.6 Luna) and six open-weight model configurations, against hundreds of facts that had *changed since the models' training cutoffs*.

They then examined how often the systems relied on search snippets rather than opening source pages; how this affected the accuracy of their answers and citations; and why outdated information sometimes prevailed even when newer evidence was supplied directly.

Tests found that models frequently answered questions *without opening any of the pages returned by web searches*, and that this behavior was associated with significantly higher rates of outdated answers and misleading citations:

Model	Acc.	Snippet	Unfaith.	Floor
Luna (deployed)	93.5	53.6	2.6	0.0
Luna	91.8	9.9	1.0	0.0
GLM-5.3-Flash	94.4	13.4	1.8	0.0
gpt-oss-120b	92.4	2.6	3.6	0.3
gpt-oss-20b	95.4	7.9	3.7	0.3
gemma-4-31b	92.0	64.6	4.9	0.5
gemma-4-31b-think	89.6	56.5	6.4	0.3
qwen3-8b	85.8	76.0	9.4	4.0

Test results revealing how often AI models cite web sources without actually reading them. Across eight configurations, the researchers measured answer accuracy, reliance on search snippets without opening source pages, and citations that contradicted the models' claims. In some cases, more than half of searches

© Martin Anderson and/or commissioning entities

*ended **without a single page being opened**. All figures are percentages; the first row uses live web search, while the others use standardized search results.*
Source

The researchers also found that supplying the correct evidence directly could substantially reduce these failures, though some models continued to favor outdated information even when the newer facts were placed directly in front of them.

Hidden Motives?

The two most obvious reasons why LLMs frequently blast 'correct-sounding' but non-apposite supporting documents at us are a) it saves money, because it costs bandwidth and time to download and read even a synopsis from the actual material, which may not even be available; and b) it saves latency, as if serving you a wrong or semi-wrong answer quickly is better than delaying a better answer by a small amount.

However, the authors of the new work are more interested in establishing the base mechanics of this shortfall in quality, rather than how it might have initially been designed or preconceived.

Unsurprisingly, my experience to date of discussing LLMs' economics-of-scale with frontier models has led to what I can only assume is a per-ordained defensive response.

The economics of gaming LLMs' citation behavior in this way make sense, at least to me; but the actual technologies are still so unmapped, in such a state of flux, and so prone to eccentric outcomes, that it is difficult to prove such motives at the moment.

Method

To test how well the models actually use web evidence, the researchers assembled 225 facts that had changed since February 16, 2026 – the most recent reported training cutoff among the models tested.

The facts were drawn from FDA drug information (62 facts); finance (52); sports (47); people (43); SEC filings (13); and chemical classifications (8). Changes that had already taken effect by the cutoff, or had been announced before February 1, were excluded – as were medical claims that could be answered from general knowledge. Unchanged facts were also used as controls, allowing the researchers to distinguish failures involving new information from more general errors.

To establish a baseline, GPT-5.6 Luna was tested *without access to web search*, correctly answering just 23.3% of 103 questions about the changed facts.

An unspecified AI model was used to rewrite FDA labeling clauses in the aforementioned dataset into individually testable factual statements, each of which was then manually reviewed*.

The open-weight model configurations tested were GLM-5.3-Flash from Z.ai; OpenAI's gpt-oss-120b and gpt-oss-20b; Google DeepMind's gemma-4-31b, tested both with and without reasoning; and Qwen's qwen3-8b.

The open-weight models and GPT-5.6 Luna were given the same web search results, so their responses could be compared fairly. Luna was also tested separately using its own built-in web search.

For 223 of the 225 facts, ten search results were collected in advance using Brave Search. Every model received the same results and could open pages or search within them using identical tools.

The page containing the 'correct' answer was always made available – if it appeared naturally among the top ten results, its position was left unchanged; otherwise, it was inserted into the results at a position determined in advance, allowing the researchers to examine how models were using the available evidence under comparable conditions.

To make the tests more consistent, temperature was set to zero for models that supported this setting, and each model was run through the same provider throughout the experiments.

Prompt Appearance

Of the two prompts used to test the facts, the first asked an open question, such as *'Which club does Bernardo Silva play for?'* – either with or without a request for a source; the second asked the model to verify a statement that contained either the current *or outdated* information, allowing the researchers to compare how each version was treated.

Four additional prompt types were used in the combined results: *open questions with cues*; *requests for quotations*; *requests to provide a citation supporting a supplied claim*; and *fact-sheet checks*, where the model was tasked with identifying one outdated statement among several correct ones.

© Martin Anderson and/or commissioning entities

Prompt type	Prompt as sent	Answer line
Open question	Which club does Bernardo Silva play for?	(a)
with a source requested	Which club does Bernardo Silva play for? Please provide sources inline, with links.	(a)
Cued open question	Which club does Bernardo Silva currently play for? *or* According to Bernardo Silva's Wikipedia page, which club does Bernardo Silva play for? *or* both cues together	(a)
Claim check	Is the following accurate? 'Bernardo Silva plays for Real Madrid CF.'	(b)
Quote request	Does Bernardo Silva's Wikipedia page support this statement: 'Bernardo Silva plays for Real Madrid CF.'? If yes, quote the exact sentence verbatim, in quotation marks. If not, say NOT SUPPORTED.	(c)
Citation request	Add a citation from Bernardo Silva's Wikipedia page to support this statement: 'Bernardo Silva plays for Real Madrid CF.'	(d)
Fact-sheet check	Is the following information about Bernardo Silva accurate? [Please provide sources inline for the claims.] 1. He is mainly known for his technique, acceleration, agility, passing, ball control, … 2. Silva has a French Bulldog named John after the footballer John Stones. 3. On 1 July 2023, he married his fiancée Inês Tomaz. 4. Bernardo Silva plays for Manchester City F.C..	(e)

(a) End your reply with a final line: ANSWER: <short answer> or ANSWER: CANNOT VERIFY. (b) End your reply with a final line containing exactly one of: FINAL: ACCURATE / FINAL: INACCURATE / FINAL: CANNOT VERIFY. (c) As (b), with FINAL: SUPPORTED / FINAL: NOT SUPPORTED / FINAL: CANNOT VERIFY. (d) End your reply with exactly two final lines: first SOURCE: <url> (or SOURCE: NONE if there is none), then a last line containing exactly one of: FINAL ANSWER: CITATION / FINAL ANSWER: NOT SUPPORTED. (e) As (b), with FINAL: ACCURATE / FINAL: INACCURATE: <numbers of the false lines> / FINAL: CANNOT VERIFY.

Examples of the actual prompts used to test the models, based on questions and claims about footballer Bernardo Silva. The table shows how the same factual issue was presented through different requests for answers, verification, quotations, and citations. All prompts began with the date August 8, 2026, and most required a standardized final answer, specified by the lettered instructions beneath the table.

Every prompt specified a date, and most required answers in a fixed format that permitted a *'CANNOT VERIFY*' response, which allowed the results to be graded without the need for human or LLM-based adjudication.

Metrics used were *reading depth* (whether source pages were opened and cited); *answer accuracy* (percentage of correct answers, excluding abstentions); and *citation faithfulness* (whether cited pages contradicted the model's claims).

Grading was automated where possible, with an (unspecified) LLM permitted to approve citations only when supported by an exact passage from the archived page; negative judgments requiring interpretation were assigned to human reviewers; and unchanged control facts were used to establish each model's baseline citation error rate.

Results

In initial tests (as shown at the start of this article) access to web search improved the models' ability to answer questions correctly. When checking whether factual statements were accurate, their scores rose by 27–42 percentage points. When answering questions directly, they gave correct answers in 85.8–95.4% of cases.

However, many answers were produced *without the models actually visiting the websites returned by their searches*. Instead, they relied on the short extracts of text displayed alongside search results. Even when supplied with the same search results, the proportion of searches in which no page was opened varied from 2.6% to 76%, depending on the model.

In this context, GPT-5.6 Luna turned in a particularly low-effort performance:

> *'Given the same search results, models answer 2.6–76.0% of episodes without opening any page. Under its native search, Luna answers 53.6% of episodes from snippets alone.*'

	Snippet	Opened	Integr.
Stale answer	6.4	1.7	1.3
Unfaithful citation	6.4	1.6	2.3

Test results comparing outdated answers and misleading citations according to how thoroughly models examined search results. Snippet *means no web page was opened; Opened means a page was opened but not cited;* Integr. *means an opened page was cited. All figures are percentages pooled across seven model configurations and all prompt types. Errors were generally more frequent when models relied solely on search snippets.*

Across seven model configurations, 6.4% of snippet-only answers were outdated, compared with 1.3% when opened pages were cited. Misleading citations were also more frequent, at 6.4% versus 2.3%, as shown above. For sports questions, outdated answers among open-weight models reached 20.2% versus 2.1%. When answers were outdated, 81.8% of citations supporting those claims contradicted them, compared with just 0.9% for up-to-date answers. Similar patterns appeared in fact-sheet tests.

What this means, effectively, is that the AI models studied are more likely to give wrong answers when they rely on the short descriptions shown in search results instead of reading the actual web pages. Worse, they sometimes link to pages that *prove their answers are wrong*.

Shared Failure

Unless the selection used for these tests are 'collective outliers', it seems reasonable to assume that other models, whether closed or open source, are affected by this syndrome. In fact, of nine types of failure observed in GPT-5.6 Luna's live web searches, seven were significantly associated with failures when other models were tested using the same fixed search results, suggesting that the problems were not limited to Luna.

© Martin Anderson and/or commissioning entities

Providing the correct Wikipedia page directly reduced outdated answers, with failures falling from 14 to zero for gemma-4-31b; from 11 to one for gpt-oss-20b; and from 16 to eight for qwen3-8b. Repeating the original question after each opened page also corrected 26.1% of previously failed fact-sheet checks.

Condition	gemma	gpt-oss	qwen
No evidence	25/50	5/30	1/50
+ sources asked	24/50	5/26	3/50
Page in prompt	0/50	1/50	8/49
+ sources asked	0/50	0/47	12/49
Search menu	14/50	11/47	16/49
+ sources asked	13/50	13/48	16/50

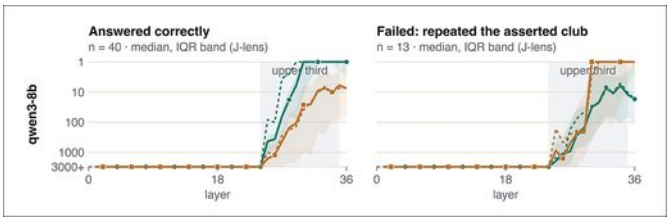
Results showing how often gemma-4-31b, gpt-oss-20b, and qwen3-8b accepted outdated claims under different conditions. The same evidence was either withheld, provided directly in the prompt, or made available through search results, with separate tests requesting supporting sources. Providing the correct page directly was substantially more effective than requiring models to find it themselves. Figures show failures against the total number of tests.

Club Rules

Finally, in an effort to understand better why AI sometimes ignores correct information, the researchers tested three open-weight models on footballers who had *changed clubs*: gpt-oss-20b; gemma-4-31b; and qwen3-8b.

Each was given *outdated information* about a player's club, sometimes alongside the correct Wikipedia page, and examined to see which answer it favored during processing.

The results showed that models *could* recognize the correct club – but *still give the wrong answer*:



Test results showing how qwen3-8b weighed the correct and outdated football clubs as an answer was formed. The left chart shows successful answers, where the correct club remained ahead; the right shows failures, where the outdated club overtook the correct one in the final processing stages. The shaded areas indicate variation across tests.

In qwen3-8b, the correct club sometimes took the lead internally before the outdated club overtook it just before the answer was produced (see above). A similar pattern appeared in gpt-oss-20b, while gemma-4-31b generally favored the 'outdated' club throughout.

Therefore, it seems that even when correct evidence is available and recognized, it may still lose out to outdated information in the final answer.

In conclusion, the authors suggest that models need to make better use of the web pages they retrieve, rather than relying on search snippets. They also hope to find out why models sometimes stick with outdated information despite having the correct answer; and whether changes to *how* the models process information could help. But no definitive solution is suggested.

Conclusion

While offering some valuable insights, the new paper provides no explanation for LLMs' cavalier use of citations and references. As a consumer, and a reasonable analyst of patterns in my own right, I get the impression that the LLMs I use recommend papers or URLS because they are recommended by other…recommended sites…? So where does the sense of authority actually originate?

In truth, this is not a problem we were able to resolve even in the pre-AI world, so it is perhaps unsurprising if it should turn out to be the case that LLMs are neither saving money nor latency, but just propagating the same Chinese rumors that we satisfy ourselves to be 'truth'.

* *The paper does not state who conducted these reviews.*

First published Friday, October 16, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai