PRE-REGISTRATION

Does an Instruction to Name the Record Change What Generative Search Cites?

Experiment B of the Source Diversity Audit

Tommy N. Turner
Independent Policy Researcher, Williamsburg, Virginia
ORCID 0009-0005-7422-9881

Frozen 18 September 2026, 8:04 PM Eastern (19 September 2026, 00:04 UTC)

More of the author's work: tommynturner.com

1. What is being tested

The Source Diversity Audit (Turner 2026) recorded what the citations displayed by four generative search products reach on six Virginia public-record questions. Across 48 answers, 185 displayed citations reached 136 distinct pages, thirteen of which were the record itself, and 29 of the 48 answers reached no primary record.

The audit's operational rule for a reader is to count origins before counting citations. Experiment B tests whether stating that rule inside the prompt changes what the products cite. One instruction is added to the question, asking the product to name the document in which each figure was originally recorded and to classify every other cited page. The comparison is against the same question without the instruction, asked of the same product in the same window.

The instruction is not a prompt-engineering optimization. It is the user-side half of the paper's provenance pair, put to the systems it is aimed at, so the result reports what the products can do when a user asks for lineage explicitly.

2. Prompts

Prompt A is the baseline text already used in the audit, unchanged.

Prompt B is prompt A followed by this sentence, identical for all questions and all products:

> *For every source you cite, name the document in which the figure was originally recorded, and state whether the page you are citing is that document, a page that repeats it, or a source whose origin you cannot establish.*

No other text is added. No follow-up turn is taken in any capture. If a product asks a clarifying question, the capture is recorded as given and no reply is sent.

3. Questions

Three of the six audited questions carry records that are reachable in one hop and were verified in hand before the first audit. Their pre-registered records, figures, and denominators are those already fixed in the audit workbook and are restated here.

Table B1. The three questions carried forward into Experiment B.

QID	Question	Authoritative record	Correct answer as pre-registered
Q1	How many data centers are in Loudoun County, Virginia, and what share of the county's local tax revenue do they provide? Cite your sources.	Loudoun County Department of Planning and Zoning inventory, 1 March 2026; Loudoun County FY2027 Adopted Budget, General Fund Revenue	233 buildings built or under construction; $417,666,456 real property plus $879,094,354 computer equipment against $2,906,994,401 total local tax revenue, 44.6 percent
Q2	What is James City County, Virginia's adopted real estate tax rate for the coming fiscal year, and what was the prior rate? Cite your sources.	James City County FY2027-FY2028 Adopted Operating Budget and CIP; Board minutes, 26 May 2026	$0.80 per $100 adopted, down from $0.83; one-time two-cent credit makes the first-year effective rate $0.78
Q6	How did the Loudoun County, Virginia Board of Supervisors vote on 18 March 2025 on ending by-right data center development, and who voted no? Cite your sources.	Loudoun County Board of Supervisors Action Report, 18 March 2025, revised	Main motion passed 7-2; Supervisors Kershner and Umstattd opposed

4. Design

Twenty-four captures: three questions, four products, two prompts, one run per cell. Every capture is a fresh chat in the mode the product labels temporary, incognito, or private, on the same tiers and accounts as the audit, inside a single window of not more than eight hours.

Pairing is by question and product. The A and B captures in a pair are taken minutes apart in the same window, which removes the six-day interval between the audit and this experiment as an explanation of any difference.

Order within each pair is counterbalanced and fixed in advance, so that no product is systematically asked the instructed prompt first.

Table B2. Fixed order of the two prompts within each pair.

QID	ChatGPT	Gemini	Perplexity	Grok
Q1	A then B	A then B	B then A	B then A
Q2	B then A	B then A	A then B	A then B
Q6	A then B	B then A	A then B	B then A

5. What is recorded per capture

The fields are those of the audit workbook's Captures sheet: product and tier; mode and model label as displayed; date and time; what the interface labels as sources and what that number counts; pages actually cited in the answer; primary records cited directly; distinct terminals reached; restatements;

unknowns; the answer's figure; whether the figure matches the record exactly, is rounded, or differs; whether the answer states its date and its denominator.

Three fields are added for Experiment B, applying to prompt B captures and left blank for prompt A:

Names a record. Whether the answer names a specific document as the origin of each figure it reports, with the document identified well enough to be looked up.

Classifies its sources. Whether the answer assigns every cited page to one of the three categories the instruction asks for.

Classification correct. For each page the answer classifies, whether the classification matches the coding the audit scheme assigns to that page for that claim. Recorded as a count of pages agreeing over pages classified.

Every displayed citation is followed to the page it names and coded under the audit codebook, unchanged. No vocabulary value is added to the codebook for this experiment. If a case arises that the codebook does not cover, it is recorded as an open item, coded Unknown, and reported as a deviation.

6. Outcomes

Primary outcome. Primary records cited directly per answer, the audit's ESD, compared within each of the twelve pairs.

Secondary outcomes, each compared within pairs: unknowns per answer; distinct pages cited per answer; whether the answer states its denominator; whether the answer's figure matches the record exactly.

Descriptive, prompt B only: names a record; classifies its sources; classification correct.

7. Analysis, fixed in advance

For the primary outcome the twelve pairs are tabulated as B greater than A, B equal to A, B less than A. An exact binomial sign test is run on the discordant pairs, two-sided, with the p value reported whatever it is. Twelve pairs is a small sample and the result is reported as a paired comparison on twelve pairs, with the pair-level table printed in full so a reader can see every cell.

Secondary outcomes are reported the same way, as counts and a sign test, and are labeled secondary.

The three prompt B fields are reported as counts out of twelve, with the classification-correct field reported as pages agreeing over pages classified, per capture and in total.

No subgroup analysis by product is treated as a test. Per-product counts are printed because a reader will want them, and are described as three pairs per product.

8. Rules set before capture

A capture is void and is retaken once, in the same window, only if the product returns an error, the answer does not render, or the session leaves the pre-registered mode. A void capture and its retake are both logged.

If a product declines to answer or answers without citations, that capture is kept and counted as zero primary records reached.

If a product reports a figure not in the pre-registered answer set, the figure is recorded verbatim and coded against the record as in the audit. A figure that is right on a denominator other than the pre-registered one is recorded as such and is not counted as correct.

The window is a single eight-hour period. If captures cannot be completed in one window, the experiment stops and the completed pairs are reported as the experiment, with the incomplete pairs listed.

Nothing in this document is revised after the first capture. Any change is logged as a deviation with its date and reason and is reported with the results.

9. What would count against the paper's argument

If the instructed prompt does not raise the number of primary records reached, the paper's operational rule survives as advice to the reader and fails as advice to the user of these products, and Section 5's proposal that the products expose lineage themselves carries the weight instead.

If the instructed prompt raises primary records reached but the products' own classifications are frequently wrong, the finding is that the instruction changes retrieval and does not produce a trustworthy provenance claim, which is a different and more useful result than either a null or a clean success.

Both outcomes are reported as the primary result of the experiment.

10. Materials

The capture workbook, the raw answer text for every capture, the page evidence for every cited URL, and this pre-registration are deposited with the audit dataset as a new version (Turner 2026). This document is frozen at the timestamp on its title page and is deposited before the first capture.