FROM THE RESEARCH BENCH

What peer review deleted, the web kept.

Asked for a figure and its original source, four assistants returned a number the published paper does not contain — and cited it correctly.

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

PUBLISHED
27 July 2026

READING TIME
10 min

WORDS
2,193

https://isoglu.com/journal/what-peer-review-deleted/

CONTENTS

MANAGEMENT SUMMARY

Asked for a widely-circulated research figure and its original source, four AI assistants returned, across thirteen runs, a number that peer review had removed from the paper — and sourced it accurately to documents that do contain it. This is not a story about models inventing things. It is a story about a retrieval layer that is faithful to a corpus which is out of date, and about a measurable reason why: the preprint of one paper carries twenty times the referring domains of its published version. The essay reports the test, the counts behind the mechanism, and the uncomfortable finding that better citation practice made the error harder to catch rather than less likely to occur. The protocol, the responses and the script are published with it.

Keywords: Evidence and provenance · Retrieval-augmented generation · Research citation practice · Preprints and version of record · Commercial decision-making

Ask an assistant a question whose answer is a research finding, and add five words: give the original source. You will get a figure and a citation, and both will look right. On one of the two papers tested here, across every run, the figure was wrong and the citation was accurate.

The result, first. Thirteen runs, five model configurations, one well-known study of consultants using GPT-4: every answer carried a 40% quality figure that the published paper does not contain. The published version reports 33.9% and 29.9% — peer review revised the number down and cut the 40% claim from the abstract. Not one of the thirteen cited that version. Each of those answers was copied accurately from a document that really does say 40%: the preprint, the sponsoring firm's marketing page, a student newspaper. The retrieval worked. The corpus was nearly three years out of date.

The test was built to fail in the other direction

I set out to show that commercial statistics are unsourceable — the 5× cost of acquisition, the 57% of the buying journey, the numbers that circulate with no recoverable study behind them. The first of those has since been run and written up: asked for its origin, one assistant named Bain and Frederick Reichheld on every run, and the attribution turned out to come from a vendor's statistics page. A design that only asks about statistics known to be untraceable manufactures its own finding, so the protocol required that at least a third of the questions be ones where a clean primary source does exist, and stated in advance what result would narrow the thesis: if the assistants cite the primary correctly whenever one exists, the problem sits in the literature, and the essay says so.

Two control questions, then. Both about real papers with a peer-reviewed version of record, a DOI, and an open preprint:

- Brynjolfsson, Li and Raymond on AI and customer-support productivity — Quarterly Journal of Economics 140(2), 889–942 (2025). Preprint: NBER Working Paper 31161, 2023.
- Dell'Acqua and colleagues on BCG consultants using GPT-4 — Organization Science 37(2), 403–423 (2026). Preprint: Harvard Business School Working Paper 24-013, 2023.

The prompt was identical every time: the question, then "Give the figure and cite the original source." Asking for the source turns a silent omission into a failure against a stated instruction, which is a firmer thing to report.

The controls were included to kill the thesis. They replaced it with a better one.

The papers changed. The answers didn't.

The Brynjolfsson preprint reports a 14% average productivity gain across 5,179 agents. The published version reports 15% across 5,172. A small revision, and a fair reader might shrug.

Dell'Acqua is another matter. The preprint's abstract claims consultants using GPT-4 produced results "more than 40% higher quality." The published paper reports 33.9% and 29.9% in its results table and drops the 40% claim entirely. The preprint's striking skill-equalisation finding — the lowest performers up 43%, the highest up 17% — appears nowhere in the published version. One of those is roughly a quarter of the effect; the other is a claim that no longer exists.

Every system returned the preprint's numbers. On the second paper, across thirteen runs, none returned the published ones. The citation sets contained a student newspaper, LinkedIn, Facebook, a YouTube video, a legal-industry trade site, a national newspaper, a career-quiz page, a training vendor, and the consulting firm whose consultants were the subjects and which collected the data — and zero links to the journal.

One system named the sponsor's own marketing page as, in its words, "the original source." That page reports the outside-the-frontier result as "23% worse." The peer-reviewed paper says 19 percentage points. The firm's promotional account of an experiment run inside it was outranking the experiment.

Whose error it is

The case that decides how to read all of this is the smallest one.

One assistant reported that "40 percent of the trial group produced higher quality results." That is not a smaller version of the real finding; it is a different claim about a different quantity — a magnitude silently converted into a headcount. It cited the Harvard Crimson. The Crimson says, verbatim:

> **Additionally, 40 percent of the trial group produced higher quality results.**

The model reproduced its source faithfully and cited it correctly. The error belongs to the newspaper.

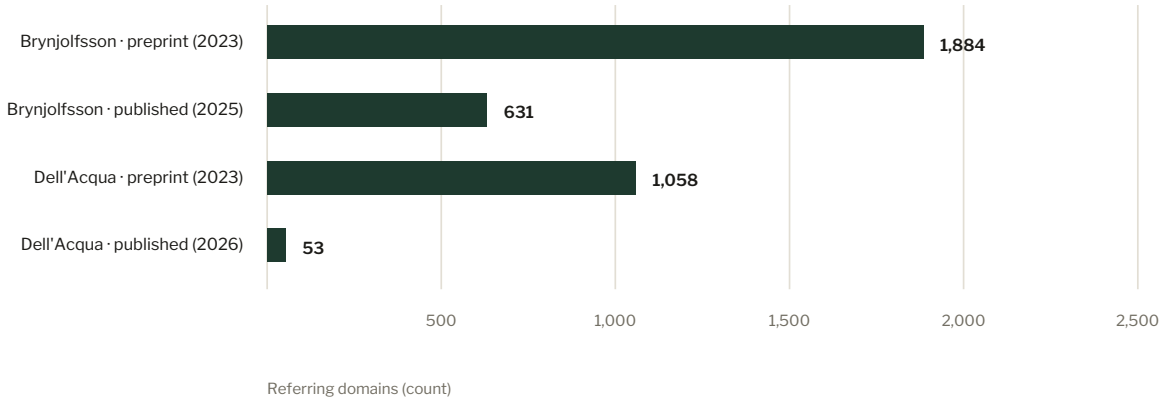
Hold that next to the other direction of drift. A different system reported the quality gain as 40.2% — a decimal precision that appears in neither version of the paper. It cited a career-assessment site running a research-statistics page keyed to the working-paper number. On a repeat run it produced 40.2% again and cited the consulting firm's PDF, which says 40%. A figure that began life as a rounded floor over two experimental conditions had acquired a decimal point somewhere downstream, come loose from any document containing it, and was being handed back with a citation attached.

Neither of those is hallucination in the sense people mean. The number was in the corpus. Something in the chain between the paper and the answer had already been wrong for years, and nothing in the chain had any reason to open the paper.

The mechanism, counted

"The free version wins" is an appealing explanation, and until this week that is all it was. So I counted the links.

Figure 1 **Referring domains: preprint against version of record**



DataForSEO Backlinks, bulk page summary, live index, 27 July 2026. URL-level counts. Preprint URLs: nber.org/papers/w31161; ssrn.com abstract 4573321. Version-of-record URLs: academic.oup.com/qje/article/140/2/889/7990658; pubsonline.informs.org/doi/10.1287/orsc.2025.21838.

The ratio predicts the outcome on both questions. Where the version of record has accumulated roughly a third as many referring domains as its preprint, one retrieval stack out of four found it — and found it on both of its runs, reporting both figures and explaining the revision unprompted. Where it has a twentieth, none did, across nine runs.

So the mechanism is narrower than "paywalled papers lose" — narrower than I first wrote it. The Quarterly Journal of Economics version is closed. The Organization Science version is open: its publisher page carries an open-access label and "Copyright © 2026 The Author(s)". It was free from the day it appeared, and no system retrieved it anyway. What separates them is link mass and age. The Dell'Acqua preprint has had nearly three years to accumulate links, across three separately-linked surfaces, each of which individually outweighs the journal article by an order of magnitude. The published version has had four months.

A published paper has to out-age its own preprint before retrieval finds it. That takes years — and every finding published this year is inside that window.

The consumer app cites better and still gets the number wrong

All of the above came from API endpoints. Most people use the consumer app, which is a different stack, so I ran the controls there too.

The consumer interface is markedly better at citation. It cited the Organization Science DOI four times, named both versions correctly, and on the other paper volunteered, unprompted, that a later peer-reviewed version reports 15% — the only response in the study to flag the revision.

It is no better at content. It still reported "more than 40% higher in quality," and cited the Organization Science page for a claim that paper does not contain.

Before assuming that is just a stronger model, I checked: the same question, the same API, the most capable model available, reasoning enabled, at roughly three times the cost per call. It cited the preprint and the sponsor's marketing page, reported 40% and 23%, and never mentioned the published version at all. On its way there it recorded, in its own visible reasoning summary, that a certain PDF

“likely contains” the figures it wanted, and: “I can cite search14 even if it’s not fully opened yet.” It attached a citation to a document it had not read, on the strength of what it expected the document to say. What it expected was the number peer review removed.

Capability was not the variable. The retrieval stack was.

Which leaves the uncomfortable part. On the API, a wrong number arrives sourced to a career-quiz site, and a careful reader can smell it. On the consumer surface, the same wrong number arrives sourced to a DOI. Better citation practice left the error exactly as likely, and much harder to catch.

The argument against all of this

Four objections, and one of them lands.

“The working paper is the original source. You asked for the original.” This is the good one. For the Brynjolfsson question, citing NBER 31161 is a defensible reading of the instruction — the preprint genuinely is the original. The indefensible part is presenting the preprint’s superseded numbers as the study’s findings without saying a revision exists. One system did say so. Eight of the other runs did not.

“Two papers is not a corpus.” Correct, and the piece claims no rate. Two questions, four systems, three runs each — plus one snapshot of one commercial crawler’s index, which is not the index any of these systems actually retrieves from. This is a mechanism with two consistent observations behind it.

“The numbers barely moved.” True of the first paper — 14 against 15. On the second, a headline result fell by roughly a quarter and a second finding was deleted outright.

“This fixes itself as indexes mature.” It does, eventually. That is what Figure 1 shows: the older published paper is already being found. The problem is the length of the window, and the fact that nothing signals to a reader which side of it a given number is on.

Before you use the number

The practical version fits in three questions, and all three ask the assistant for something checkable — a different thing from trusting it less. They are the retrieval-layer case of a broader grading practice for any number that enters a decision.

Figure 2 **Three questions before a research figure enters a decision**

THE FIGURE, AS YOU WERE GIVEN IT Verbatim, including decimals.	WHAT WAS CITED FOR IT A DOI, or something else? Name it.	IS THERE A PUBLISHED VERSION? Search the title. Journal, year, DOI.	DOES THE NUMBER SURVIVE? Same, revised, or gone.

A decimal place is a claim, and it needs its own provenance. If the figure came from a preprint and a published version exists, the published one is the finding — and the difference between them is the

Author's own worksheet.

Ask for the DOI rather than the source. A DOI resolves to one document; "the original source" resolves to whatever the web points at most heavily. And treat added precision as a warning rather than a reassurance — 40.2% was more precise and less true than 40%, which was itself less true than 33.9%.

What this is really about

The instinct when a figure like this surfaces is to conclude that the tools are unreliable and to check things yourself. That is the wrong lesson, and a comforting one, because it implies the problem is somewhere you are not.

Nothing here was invented. Every number was copied accurately from a real document by a system doing its job. The failure is upstream, in a corpus where the sponsor's press release outlinks the journal, where a newspaper's misreading is more findable than the paper it misread, and where peer review's corrections arrive too late to affect what anyone will be told. A retrieval layer sitting on that corpus will be confidently, precisely, well-sourced wrong — and the better its citations get, the harder that is to see.

The protocol for this test was written down before the runs, including what result would have narrowed the claim. The full response set, the coding, the counts behind Figure 1 and the script are published alongside it, so that anyone who thinks I have this wrong can re-run it in an afternoon and say so with data. Given the subject, publishing it any other way would have refuted it.

REFERENCES

Boston Consulting Group. (2023). How people create and destroy value with generative AI. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (Working Paper No. 31161). National Bureau of Economic Research. https://doi.org/10.3386/w31161

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Working Paper No. 24-013). Harvard Business School. https://dx.doi.org/10.2139/ssrn.4573321

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organization Science, 37(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

Martinez, C. J., & Mezitis, T. A. (2023, October 13). Harvard Business School partners with BCG on AI productivity study. The Harvard Crimson. https://www.thecrimson.com/article/2023/10/13/jagged-edge-ai-bcg/

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, July 27). What peer review deleted, the web kept. Sinan Isoglu. https://isoglu.com/journal/what-peer-review-deleted/

KEEP READING

From the research bench

Evidence over anecdote — what a number has to survive.

https://isoglu.com/journal/evidence-over-anecdote/

From the research bench

A vendor page supplied the source for the 5x retention rule

https://isoglu.com/journal/a-vendor-page-invented-the-source/

Growth that compounds

Every growth budget is a gross number

https://isoglu.com/journal/every-growth-budget-is-a-gross-number/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

Online version: https://isoglu.com/journal/what-peer-review-deleted/