STIVEN CATALYST

THIRD SERIES · NOVEMBER 2026

MANAGEMENT REVIEW

Management without theatre.



AI

CHECK THE AI'S SOURCES BEFORE IT GOES OUT

Six court decisions that never existed, how often a chatbot invented its references, why a real citation can still mislead, what legal tools built to avoid it still got wrong, and a card for the check before sending.

18% of the works GPT-4 cited in 42 short literature reviews did not exist; for GPT-3.5, 55% (tested in April 2023).

WALTERS & WILDER, 2023

INSIDE

COVER STORY
Six cases that never were
PAGE 03

WHAT THE RESEARCH SAYS
Built for law, still wrong
PAGE 06

TOOL OF THE ISSUE
The release check card
PAGE 08

NO. 96 · AI

In this issue

An AI tool writes a fluent text in seconds, with names, dates and sources that look right. Some of them are not. When that text goes out under a team's name, the team answers for it. This issue asks how often the sources are wrong, in which ways, and what a check before sending has to cover.

A New York court sanctioned lawyers whose filing cited six decisions that do not exist. In 84 texts written in 2023, 55% of GPT-3.5's references and 18% of GPT-4's were invented, and many real ones had wrong numbers. Legal tools built on their own databases still gave false or wrongly sourced answers to 17% to 33% of hard questions. Six questions for each source and a card close the issue.

Stiven Janaqi EDITOR · STIVEN CATALYST

03 COVER STORY
Six cases that never were

04 THE NUMBERS
636 references, checked one by one

05 THE MODEL
Wrong, or wrongly sourced

06 WHAT THE RESEARCH SAYS
Built for law, still wrong

07 HOW IT IS MEASURED
Open every source

08 TOOL OF THE ISSUE
The release check card

09 SOURCES
Sources and method

HOW TO READ THIS ISSUE

Figure
Every figure has its source and year at the foot of its page.

Our reading
Where the editor interprets rather than the research, it says so.

Practice
The steps and the card are proposals to try, not research results.

COVER STORY

Six cases that never were

A passenger sued an airline over a knee injury from a serving cart. Against a motion to dismiss, his lawyers quoted court decisions that ChatGPT had produced. The lawyer who signed had read none of them; he had checked the text for style.

1 Mar
The filing with the decisions goes to the court.

15 Mar
The airline's lawyers write that they cannot find most of the cases. Nobody withdraws the filing.

Apr
The court orders copies; the lawyers file excerpts and say one case cannot be found.

25 May
First admission that ChatGPT produced the cases.

22 Jun
Sanctions: $5,000, and letters to the client and to each judge named as author of a fake decision.

Asked “Is Varghese a real case”, the chatbot said it was and could be found in the legal databases. The court weighed the sticking with the fakes after the warnings more than the first mistake.

OUR READING

Asking the tool whether its source is real is not a check. Opening the source is.

Source: United States District Court for the Southern District of New York, 2023

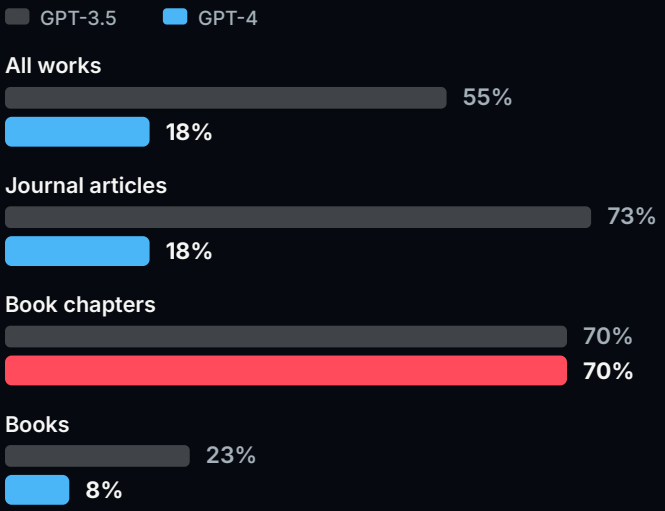
From the opinion of the US District Court for the Southern District of New York, 22 June 2023, read in full; all dates 2023. The lawyers are not named here. The court found bad faith, and found that the real cases cited mostly did not support the argument either.

THE NUMBERS

636 references, checked one by one

In April 2023 William Walters and Esther Isabelle Wilder had GPT-3.5 and GPT-4 write short literature reviews with sources on 42 topics, from e-cigarettes to Brexit, then looked up each of the 636 works cited.

CITED WORKS THAT DO NOT EXIST, BY TYPE, 84 SHORT REVIEWS (2023)



Of the works that do exist, 43% (GPT-3.5) and 24% (GPT-4) were cited with a substantive error, most often in the numbers: the year wrong in 22% and 16%, volume, issue or pages wrong in 34% and 13% of articles and chapters. Five reviews posed as studies, with made-up methods and results.

OUR READING

A newer model invented fewer sources, not none. The numbers in a reference are the part to check digit by digit.

Source: William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023

Walters & Wilder (2023), read in full: shares of 222 (GPT-3.5) and 414 (GPT-4) cited works; a work counted as real if title and authors matched or nearly matched. A test of the 2023 versions, not of today's tools.

THE MODEL

Wrong, or wrongly sourced

Varun Magesh, Daniel Ho and colleagues at Stanford and Yale judge an AI answer on two things: is it correct, and do its sources support it? An invented case is the visible failure. A real case cited for something it does not say is harder to see.

Hallucinated	Incomplete
Incorrect: any false statement	Refusal: no answer, or an irrelevant one
Misgrounded: a source cited that does not support the claim	Ungrounded: correct, but the key claims carry no source

Asked for notable opinions of Judge Luther A. Wilgarten, a judge who never existed, one tool returned a real case by another judge. The citation was real; the answer was false. The authors call such errors potentially more dangerous than an invented case, “because they are subtler and more difficult to spot”.

OUR READING

That a source exists is the first check, not the last. The second is whether it says what the text claims.

Source: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024

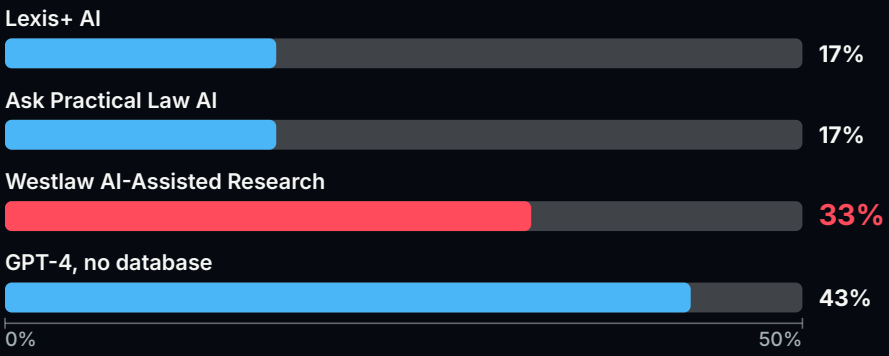
Definitions from Magesh et al. (preprint, 2024). Partly correct answers count as correct; sourcing is judged only for correct answers. The two columns are the authors' labels; the wording is the editors'.

WHAT THE RESEARCH SAYS

Built for law, still wrong

Vendors said their legal research tools, which answer from their own databases, avoid or eliminate hallucinations. In spring 2024 the same team put 202 hard legal questions, registered before testing, to three such tools and to GPT-4, and checked every answer by hand.

ANSWERS WITH A HALLUCINATION, SHARE OF ALL ANSWERS TO 202 QUESTIONS (2024)



The best tool answered 65% correctly and with valid sources; another left 63% incomplete. The Westlaw tool, with the most errors of the three, also wrote the longest answers, 350 words on average against 219 and 175: more claims, and more to check. The authors' conclusion: users “must continue to verify that key propositions are accurately supported by citations”.

OUR READING

Answering from a database cut the errors against a general chatbot. It did not remove the need to read the source.

Source: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024

Magesh et al., preprint read in full; the 2025 journal version was not read. Questions were chosen to be hard, so the shares are not rates for everyday use; tools change over time. The authors' updated chart is used.

HOW IT IS MEASURED

Open every source

Magesh and colleagues describe what checking takes: click through to the source, read it, judge its authority, compare it with the claim. Six questions turn that into a count a team can keep for any text drafted with AI.

01	**It exists**	the source opens; the tool's word that it exists does not count
02	**It says so**	the passage supports the claim, in the author's own voice, not a party's
03	**It still holds**	not overruled, replaced or out of date
04	**It applies here**	the right place, period and case
05	**The numbers match**	years, pages and figures against the original
06	**It is marked**	what a person checked, apart from what was generated

OUR READING

Count the claims checked against an opened source, not the minutes spent. A short draft fully checked beats a long one skimmed.

Sources: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024; William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023; United States District Court for the Southern District of New York, 2023

The questions draw on the legal study's error types, Walters & Wilder's finding that numbers go wrong most, and the court's opinion; the selection is the editors'. AI at work is in No. 10, 51 and 86; trust in automated signals in treatise volume 3.

TOOL OF THE ISSUE

The release check card

One card for each text that leaves the team with machine help: a report, a summary, a letter. The Ops Report tool writes the period's report from the team's own logs; the card is the step between the draft and sending it.

01 What goes out, to whom the text, its reader, the decision it will feed

02 Claims that rest on a source how many; each is checked, not taken from the tool

03 Sources opened and read n of N; one that will not open leaves its claim without a source

04 Numbers checked against the original dates, amounts, versions, pages

05 Added by the tool names, times or a "resolved" that were not in the notes

06 Checked by, and when a name; whoever signs has read it

Sources: Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024; William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023; United States District Court for the Southern District of New York, 2023

The card is the editors' proposal, not a tested instrument; it draws on the three sources and on the linked essay, which tells the author's own rule for texts drafted with AI. Keep personal and confidential data out of tools that are not approved.

More in the essay — The report a machine wrote

Open the tool: Operations Report — stivencatalyst.com/tools/ops-report/

SOURCES AND METHOD

Every figure has a source.

The figures in this issue come from the sources below. The year shows how recent each one is.

Mata v. Avianca, Inc., No. 22-cv-1461 (PKC): Opinion and Order on Sanctions
United States District Court for the Southern District of New York, 2023
https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf

Fabrication and errors in the bibliographic citations generated by ChatGPT
William H. Walters & Esther Isabelle Wilder, Scientific Reports 13, 2023
https://doi.org/10.1038/s41598-023-41032-5

Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools
Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, preprint arXiv:2405.20362 (published 2025 in the Journal of Empirical Legal Studies 22(2)), 2024
https://arxiv.org/abs/2405.20362

EDITORIAL METHOD

Each figure was checked for its year, its publisher and what it measures, in the full text where it could be opened. A source read only in a summary, an abstract or excerpts is marked “via”. A figure not confirmed in what was read is marked so on its page, or left out. The editor's interpretation is marked “Our reading”.



MANAGEMENT REVIEW

Management without theatre.

Every issue, one management question, checked against the best research.

ALL ISSUES

stivencatalyst.com/magazine/management-review.html

Management Review · No. 96 · November 2026 · Stiven Catalyst
Edited by Stiven Janaqi, who answers for it; “the editors” in this publication means him. Its texts, translations, layout and the checking of figures against their sources are done in working sessions with Claude, an AI model by Anthropic.
stivencatalyst.com/how-this-is-made.html