The Footnote Fallacy: Cite-Checking in the Age of Vibe Citing

Citation-backed AI research tools invite more trust than they earn. For a fiduciary, that misplaced confidence is not a technology problem — it is a duty problem, and the duty already exists.

By Gordon Eng Traiceback Solutions LLC AIMSMART™ Series
AUGUST 2026

For lawyers who worked on a law school journal, this article may bring back memories of the yeoman duty of cite-checking each footnote of a submitted law journal article, not only for the accuracy of the cite, but equally important, verifying whether the footnote's content supports the claim being made. With the increasingly ubiquitous use of Generative Artificial Intelligence tools, this age-old practice should be emulated for anyone or any organization that posts or publishes material with citations. Apparently, in the age of generative AI, cite-checking has not become the *de rigueur* practice it should be, even among prestigious organizations whose professional practice is rooted in the verification of highly detailed information.

Over the fifteen months to July 2026, three of the four largest professional services firms in the world published research containing citations that did not support the claims attached to them. In one report since taken down by its publisher, an independent forensic review by GPTZero of forty-five references found five references that pointed accurately to a real, intact source; twenty-eight, however, carried paraphrased titles or fabricated components, and twelve could not be verified at all.[1] In another instance, sixteen of twenty-seven references were hallucinated, a nearly 60% error rate.[2] In a third instance, reported by the Financial Times on July 29,

1 GPTZero, "Chasing the Hallucinations: KPMG's AI-Powered Attempt at 'Redefining Excellence,'" June 2026, Subject report: *Total Experience: Redefining Excellence in the Age of Agentic AI* (October 2025). Available at: https://gptzero.me/news/investigations-kpmg/. Last visited July 30, 2026.

2 GPTZero investigation, May 14, 2026, into EY Canada, *Points of Attack: Uncovering Cyber Threats and Fraud in Loyalty Systems*; (an investigative report by Om Ogale, Paul Esau, and Alex Cui examining a 44-page report on cyber security published by one of the "big four" global consulting firms and concluding that, "the document is a collage of vibe citations, misattributions, fake statistics, and AI-written text"), available at: https://gptzero.me/investigations/ey; last visited July 31, 2026. See also: Stephen Foley, "EY

footnotes resolved to pages that did not contain the evidence they were offered for and, in some instances, to pages that did not exist; one cited academic study appears not to exist in the journal named or under the authors credited, and one report promotes a proprietary framework for which there is scant evidence outside the report itself while asserting that four national governments have adopted it.[3]

Ironically, each of these documents was a thought leadership piece: content produced to demonstrate command of artificial intelligence and to attract advisory work in it. Named third parties — a global bank, a regional health authority, and two public transit operators — disputed claims made about their own AI programs. One report contradicted its own firm's published survey data.[4] These are not obscure publishers. They are household brand-name organizations that compliance officers, investment committees, and boards have treated as authoritative without a second thought, and their failure mode was not analytical error. It was an unchecked footnote.

That these findings originate with a commercial AI-detection firm is worth acknowledging rather than glossing over: a vendor that sells detection has an interest in what it detects. The material claims here, however, do not rest solely on GPTZero's say-so. The *Financial Times* verified the investigations; the organizations that disputed the claims about their own AI programs did so directly to that newspaper; and the survey contradiction can be checked by anyone who opens the two published documents.

---

retracts study after researchers discover AI hallucinations", *Financial Times*, May 15, 2026 (reporting that EY has withdrawn the study entitled "Points of Attack: Uncovering Cyber Threats and Fraud in Loyalty Systems" after hallucinations were reported by research group GPTZero and noting that, "[h]allucinations have been a recurring problem across the professional services."). Available at: https://www.ft.com/content/a61cbcae-95e4-4449-86e1-ef40fb306f4e?syn-25a6b1a6=1. (Subs.req'd). Last visited: July 31, 2026.

3 GPTZero, "Chasing the Hallucinations: PwC report hallucinates product and government customers," July 28, 2026, https://gptzero.me/news/investigations-pwc/ (four reports published 2024–2026 by the firm's Middle East member entity) Last visited July 31, 2026; See also: Stephen Foley, "PwC published reports on AI marred by AI hallucinations," *Financial Times*, July 29, 2026, available at: https://www.ft.com/content/7e149ac8-2ce2-4266-8940-192f9821b33c?syn-25a6b1a6=1 (subs.req'd). The firm told the FT that it "takes the accuracy of our published research seriously and is updating a limited number of supporting citations" in the identified reports. Last visited: July 31, 2026.

4 Elizabeth Bratton and Stephen Foley, "KPMG report contained AI hallucinations on benefits of . . . AI," *Financial Times*, June 12, 2026, available at: https://www.ft.com/content/b3828e92-4961-4b39-84f0-c42f33be3c3f?syn-25a6b1a6=1 (subs. req'd) (reporting that the inaccuracies were identified by the research group GPTZero and verified by the FT, and that spokespersons for UBS, Swiss Federal Railways, Transport for London and NHS Greater Manchester each disputed the report's characterization of their organizations' use of AI; the NHS Greater Manchester assertion traced to a communiqué concerning a tool designed to combat lung cancer, which made no reference to the capabilities claimed). Last visited: July 31, 2026. On the survey discrepancy — 55% of chief executives ranking AI as a leading investment priority, against 71% in the same firm's own annual chief-executive survey published that month — see GPTZero, *supra* note 1.

> *The institutions selling assurance about artificial intelligence could not verify their own citations. The question this paper asks is narrower and more uncomfortable: on what basis is your firm confident that it can?*

That question lands on a belief that has quietly hardened among sophisticated users of AI. They have accepted that large language models hallucinate — that a fluent answer can be confidently wrong — and they have adjusted accordingly. But many have carved out an exception for a particular class of tool: the answer engine that shows its sources. If the response arrives with numbered citations, each one a clickable link to a real page, the reasoning goes, then the answer has been checked. The footnote is treated as proof of work.

This exception is where the risk now lives. It is held most firmly by exactly the people who should know better — capable professionals who have learned to distrust unsourced AI output and, in learning that lesson, have over-corrected into trusting sourced output. The citation feels like the fix. It is not. And for anyone operating under a fiduciary standard, the gap between how trustworthy these tools feel and how trustworthy they are has become a governable exposure — one that sits squarely inside obligations that already bind the firm.

I The belief, stated fairly

It is worth stating the belief in its strongest form, because it is not without merit. Tools built around live retrieval — the category that leads with citations — genuinely are more reliable on current facts than a bare model answering from memory. Grounding a response in retrieved documents, rather than in the model's parametric recall, measurably reduces fabrication. Independent testing of citation-first engines has also found them, at their best, to be the strongest of the major tools on citation accuracy and to surface a wide range of sources per query. None of that is marketing; it is real, and it is the honest foundation of the belief.

The error is not in valuing grounding. The error is in the inference drawn from it: that because retrieval helps, and because the citations are visible, the answer has therefore been verified. Two distinct things are being fused — the reliability gain that comes from retrieving, and the assurance implied by displaying a citation. The first is real. The second is largely an illusion. Separating them is the whole argument.

II What the tool is actually doing

Every answer engine of this kind — regardless of brand — runs the same basic sequence. A surrounding system takes the user's question, issues one or more searches, fetches candidate pages, and feeds the retrieved text into the model as ordinary input. The model then composes an answer over that material and attaches citations pointing back at the pages it drew from. This is retrieval-augmented generation ("RAG"), and it is not proprietary physics. A citation-first product is not doing something categorically different from any other model with web search switched on; it has simply made retrieval its default posture and citation its default presentation.

That matters because it means the citation-first engine inherits every limitation of the underlying approach. The material it can reach is not "the web"; it is a narrowed slice assembled from three separate pools — what the model already holds from training, what its retriever could reach and was permitted to fetch at that moment, and whatever content the provider has licensed. The synthesis is presented as a neutral survey. It is nothing of the kind. It is an opinionated selection, shaped by which sources rank, which are cleanly extractable, and which the retriever happens to favor.

The generation step then adds its own hazard. The model is not transcribing the retrieved sources; it is writing over them. Nothing in the architecture guarantees that the sentence the model produces is faithful to the source it then cites beside that sentence. The citation is generated too. And that is the seam where the fallacy opens.

III Where the belief breaks

A citation makes two implicit promises: that the source exists, and that the source supports the claim. Users check the first — the link resolves, the page loads — and unconsciously accept the second. But as every law school journal student knows, source existence and source accuracy are different tests, and the second one fails often. The most-cited independent audit of this behavior, conducted by the Columbia Journalism Review's Tow Center across eight generative search tools and sixteen hundred queries, found incorrect responses in excess of sixty percent overall; the strongest performer still answered incorrectly in roughly a third of cases, and the weakest in ninety-four percent.[5]

5 Klaudia Jaźwińska and Aisvarya Chandrasekar, "AI Search Has a Citation Problem," Tow Center for Digital Journalism, *Columbia Journalism Review*, March 6, 2025, https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php. (tested eight generative search tools by OpenAI's ChatGPT Search, Perplexity, Perplexity Pro, DeepSeek Search, Microsoft's Copilot, xAI's Grok-2 and Grok-3 (beta), and

An earlier Tow Center evaluation, examining attribution specifically, found one leading tool returning attributions that were wholly or partly incorrect in one hundred fifty-three of two hundred quotations drawn from twenty publications.[6] Set those figures against vendor accuracy claims well into the ninety-percent range, and the distance between the marketed number and the measured one becomes the story. Two further findings from the same work cut against intuition and deserve emphasis: paid premium tiers produced confidently incorrect answers at a *higher* rate than their free counterparts, and the tools showed a tendency to cite syndicated or republished copies in preference to the originals.[7]

The failures fall into three recognizable forms, and each is harder to catch precisely because the output looks sourced.

Misattribution

The information is right; the citation points to a source that does not actually contain it. Because the underlying fact checks out when the reader already knows it, misattribution survives casual scrutiny. Its damage is downstream: the citation enters a memo, a colleague relies on it, and the trail leads to a document that never made the claim. The forensic reviews described at the outset of this note found this form dominant — references that fused the authors of one work to the title of another or credited a commentary about an organization to the organization itself.

Phantom statistics

This is the most dangerous for anyone working with numbers. The engine attaches a specific figure to a named source; the reader clicks through; the number appears nowhere in that source. The model inferred a plausible-seeming statistic from the general content and then cited the page it was reading, rather than a page that stated the number. A well-formed figure

Google's Gemini; chose 20 news publishers with varying instances of AI access and randomly selected ten articles from each publisher; then manually selected direct excerpts from those articles for use in their study by running 1600 chatbot queries designed to verify accuracy and found collectively, "incorrect answers to more than 60% of queries"). Last visited July 31, 2026.

6 Klaudia Jaźwińska and Aisvarya Chandrasekar, "How ChatGPT Search (Mis)represents Publisher Content," Tow Center for Digital Journalism, November 27, 2024 (findingthat, in 200 quotation-attribution tests, 153 produced confidently partially correct or incorrect attributions but noting that it reached out to OpenAI for comment, and a spokesperson for OpenAI stated: "Misattribution is hard to address without the data and methodology that the Tow Center withheld, and the study represents an atypical test of our product"). Available at: https://www.cjr.org/tow_center/how-chatgpt-misrepresents-publisher-content.php. Last visited: July 31, 2026.

7 Jaźwińska and Chandrasekar, *supra* note 5 (premium tiers returned confidently incorrect answers at a higher rate than free tiers; tools cited syndicated and republished copies in preference to originals).

beside a real link is the most authoritative-looking artifact these tools produce — and one of the least trustworthy.

Three stages of citation-laundering: placement, layering, and integration

Those familiar with money-laundering will recognize the same three stages characterizing anti-money laundering and counter terrorism financing practices. In cases of outright fabrication, the information itself is wrong, paired with a citation that is irrelevant, or that resolves to a page that is itself unreliable. This is the placement stage: putting the defective citation in an otherwise honest article. This variant deserves particular attention, because it is no longer theoretical. Where a fabricated figure is published under a trusted imprint, it does not stay put. Like the layering stage of money laundering, statistics from one such report get picked up and republished by trade outlets and media before a publisher takes it down — and taking a report down does not take down the citations to it. Nor does removal amount to erasure. A document taken off a homepage persists in caches, mirrors, member-firm domains, and third-party republication. Finally, as in the integration stage of money laundering, questionable statistics of this kind are served back by language models themselves — a phenomenon GPTZero's chief executive, Edward Tian, has labelled second-hand hallucination.[8]

The fabrication has by then entered the record as an ordinary secondary source, indexable and citable, available to be retrieved and served back to a later reader as authority. That is the citation-laundering chain in operation: a machine generates plausible text, a trusted publisher lends it an imprint, a search index absorbs it, and a second machine cites it back to the user. The footnote is real. What it points to is not. One forensic team has given the underlying behavior a name — *vibe citing* — spanning wholly invented references, fusions of two genuine ones, and paraphrases altered past the point of accuracy.[9]

---

8 Edward Tian, "Second-Hand Hallucinations: Investigating Perplexity's AI-Generated Sources," June 11, 2024, GPTZero, (finding at the time of the article "GPTZero has noticed an increased number of sources linked by Perplexity that are AI-generated themselves"). Available at: https://gptzero.me/news/gptzero-perplexity-investigation/. Last visited: July 31, 2026.

9 GPTZero, *supra* note 1, (writing that "[w]e use the term 'vibe citing' to describe the accidental creation of fake references via LLM hallucinations across a spectrum of severity," which "can include references that are entirely fabricated (fake authors, fake title, and fake container/locators), fusions of two or more real references (authors of paper A paired with the title of paper B), or paraphrased or heavily altered versions of real citations. Our definition excludes common human errors.")

> *A citation that points to the wrong place is more dangerous than no citation at all, because it manufactures the appearance of verification while supplying none of the substance.*

IV Mechanism and consequence: a distinction worth keeping

Two different failures have been described so far, and it is worth resisting the temptation to merge them, because merging them would commit the error this paper exists to warn against.

The audits of answer engines describe a mechanism. They measure what happens inside the tool: retrieval narrows the source set, generation writes over it, and the citation is composed alongside the sentence rather than derived from it. That is a supply-side finding about how the artifact is manufactured.

The consultancy reports describe a consequence. No answer engine forced anyone to publish them. What happened there was the ordinary use of a generative tool to draft prose, followed by the absence of any step in which a person opened the cited pages and confirmed that they said what the draft claimed. That is a demand-side failure — a governance failure, not a technology failure. The tool behaved exactly as the mechanism predicts. The control that should have caught it did not exist.

Keeping the two apart matters for a practical reason. If the problem were purely the tool, the remedy would be procurement: buy a better engine, wait for the error rate to fall. It is not purely the tool, and so procurement will not close it. The measured error rates are the reason a verification control is necessary; the published reports are the demonstration of what happens in its absence. An organization that reads the second story as a caution about vendors, rather than as a caution about its own review process, has drawn precisely the wrong lesson.

V The trust the footnote buys

If the citations merely failed at a measurable rate, the problem would be manageable arithmetic. What makes it a trap is behavioral. The presence of a citation tends to raise a reader’s confidence in a result independent of whether the attribution is correct. The footnote signals diligence, and the trap for the unwary is to accept the signal in place of the diligence.

This inverts the safeguard. A citation is supposed to be an invitation to verify, an intellectual RSVP if you will. In practice it functions as a substitute for verifying — the very presence of the number persuades the reader that

someone already did the checking. The tool that most encourages a professional to stop checking is the one that presents its work most convincingly. That is why the sophistication of the user offers so little protection: the effect operates on trust calibration, not on knowledge. It is also why these reports cleared internal review. Nobody in the chain was incompetent. The apparatus looked done; it had the Good Housekeeping seal of approval, or so it seemed.

The failure is also most acute where reliance is heaviest. Accuracy on these tasks degrades as questions get more specialized and more recent, and as the retrieval reaches deeper into a topic — the exact conditions of serious professional research. Ironically, where a busy analyst most wants a confident synthesis is precisely where the synthesis is least trustworthy and the citations most likely to entrap.

VI The duty translation

For a registered investment adviser or an ERISA fiduciary, none of the foregoing is merely interesting. It is a compliance question, and the duty attaches without any AI-specific rule. Each citation-backed answer relied upon in a workflow is a compliance artifact, and it lands on obligations already in force.

Advisers Act §206 and Rule 206(4)-7. The antifraud obligation and the duty to adopt policies designed to prevent violations presuppose a reasonable basis for the information on which advice rests. An answer drawn from an unexamined, non-neutral source set — and accepted because it was footnoted — is not a reasonable basis. The policy failure is not that staff used an AI tool; it is that the firm had no standard requiring the footnotes to be checked.

Rule 204-2, books and records. Where an output informs advice, marketing, or a filing, the firm must be able to reconstruct its basis. A live-retrieval answer is non-deterministic: the same prompt returns different sources next week as rankings, permissions, and licenses shift beneath it. Absent captured provenance — the query, the timestamp, the sources actually retrieved — there is nothing to reconstruct. “The AI told me” is not a record.

Regulation S-P. Every live-grounded query is an outbound disclosure of the prompt’s content to the retrieval backend and its subprocessors. Where a prompt may carry client-identifying or matter-specific detail, the convenience of an answer engine has quietly created an egress event —

analytically the same category of exposure the firm already scrutinizes elsewhere, now riding on a research habit no one classified as data transmission.

ERISA PRUDENCE. For a plan fiduciary, procedural prudence turns on the quality of the process, not the outcome. Relying on AI-surfaced current information — its currency and its attribution unverified — because it arrived with citations is a defect in process. The prudent step is not to abstain from the tool; it is to verify what it produces before acting on it.

THIRD-PARTY RESEARCH AND VENDOR DILIGENCE. This is the obligation the last fifteen months have moved. Firms have long treated major-house research and consultancy thought leadership as an authoritative tier requiring no independent verification — a reasonable convention when such material carried an editorial process. That convention no longer holds on its own terms. Where a market-sizing figure, an adoption statistic, or a peer-practice claim drawn from a branded report finds its way into an investment memo, a due-diligence file, a marketing piece, or a committee minute, the firm relying on it has adopted the citation as its own. And ownership has responsibilities. Here, that responsibility is to perform the cite-checking.

VII Practice Pointers

The remedy is not prohibition. These tools are genuinely useful for what they are good at — orienting quickly in an unfamiliar area, scoping a question, discovering candidate sources. The remedy is a division of labor and a verification standard that treats the citation as the beginning of diligence rather than the end of it.

The division of labor is simple. Use the answer engine for discovery: fast orientation, finding the sources worth reading, getting the shape of a topic. By analogy, first-year law students are taught to refer to a general treatise on a topic before diving into detailed primary sources and case law. Similarly, use a raw index — ordinary search — and the primary documents themselves for verification: getting to the actual filing, checking a cited figure against its origin, confirming nothing material was filtered out of the synthesis. The answer engine cannot verify itself; something that returns unmediated pointers has to close the loop. This is the enduring role of conventional search in an answer-engine world — not because it is more accurate, but because it does the one job the synthesis structurally cannot: put the analyst’s eyes on the source.

PRACTICE POINTERS FOR AI-ASSISTED RESEARCH

1. Existence is not accuracy. Confirm not that the cited source loads, but that it actually contains the specific claim, quote, or figure attributed to it.

2. Every number to its origin. Treat any statistic as a phantom until located in the primary source. Do not carry a figure forward on the strength of a footnote alone.

3. Tier the sources — and retier them. Distinguish authoritative primaries, such as regulator filings and official registers, from everything else. "Everything else" now includes branded research and consultancy thought leadership, which has demonstrably carried fabricated citations. An imprint is not a primary.

4. Preserve provenance. For anything that will inform advice or a filing, capture the query, the timestamp, and the sources relied upon, so the basis can be reconstructed.

5. Label currency. Mark whether a fact reflects the model's training cutoff or was retrieved live, and when.

6. Evidence the check, not just the result. A verification that leaves no trace is indistinguishable from one that never happened — to a reviewer, to an examiner, and in hindsight to the firm itself.

Would the Practice Pointers have caught it?

The test of a control is whether it would have stopped the thing that happened. Applied to the defects catalogued in those reports, each pointer engages a distinct failure:

Pointer 1 catches the footnote resolving to a live page that does not contain the claim — the single most common defect found, and invisible to anyone who only confirms that the link opens. Pointer 2 catches the adoption statistic attributed to a survey that the cited page never mentions. Pointer 3 catches the "success story" sourced to an anonymous personal blog post, and the market claim resting on a study that does not exist in the journal named. Pointer 4 catches the tell that no reviewer would otherwise notice: the same assertion appearing three times in two pages, footnoted once, then not at all, then twice to two different sources — a pattern no human author produces. Pointer 5 catches the case of a genuine initiative, real and

correctly described, whose provenance predated generative AI by five years and which therefore could not evidence what it was cited to evidence. Pointer 6 catches all of them a second time, by making the absence of any check visible on the face of the file.

None of these pointers are sophisticated. Not one requires a tool, a vendor, or a budget line. What they require is that somebody be accountable for having opened the page — and that the opening be recorded. The reports in question failed not because the standard was hard, but because no one had written it down.

Nor is this a consultancy problem. In April 2026 a white-shoe law firm acknowledged that a filing it had submitted in a bankruptcy case contained numerous AI-generated inaccuracies, including misreadings of the bankruptcy code.[10] If members of a profession whose apprenticeship begins with cite-checking can commit this error in a document filed with a court, then no firm should assume its internal processes are immune from the same mistake.

Written into an AI-use policy, this reduces to a single line worth adopting verbatim: **a citation is where verification begins, not proof that it happened.** Any AI-surfaced fact, statistic, or citation destined for advice, marketing, or a filing is checked against its primary source, and the check is evidenced, before it leaves the analyst's desk.

And a single question is worth raising at the next compliance meeting, because it usually surfaces an exposure no one had named: *which of our workflows already relies on an AI research tool — and who is verifying its footnotes?*

The reassurance offered by a citation-backed answer is doing psychological work, not epistemic work. Recognizing the difference — and building the modest discipline that supplies the epistemic work the footnote only pretends to — is the whole of the obligation. It is not novel law. It is old duty, meeting a new and unusually persuasive interface.

10 Bratton and Foley, *supra* note 4 (reporting that Sullivan & Cromwell acknowledged in April 2026 that a filing it submitted in a bankruptcy case contained numerous AI-generated inaccuracies, including misreadings of the US bankruptcy code). Last visited: July 31, 2026.