

Who Gets to Audit AI?

Independent scrutiny needs a record.

The right to evaluate an AI system is incomplete without a proportionate right to keep the evidence needed to verify, challenge and correct what the evaluation finds.

Ozue C. Ezeife · Founder, AlphaCitation

AI systems now answer questions people act on: whether a medicine is approved, which product to choose, what a regulator has decided. A single answer can be checked against the record. Tracking changes requires earlier answers; investigating their causes requires further evidence.

Independent auditors therefore need more than access to a system. They need to retain evidence, compare responses over time and publish findings, including findings unfavourable to the provider.

Three of the five providers whose terms are reviewed here give the customer ownership of its outputs, and one welcomes research publications. None of their terms expressly grants an independent auditor the right to keep those answers in order to audit them, or says whether its access survives an unfavourable finding.

October showed why that matters. Kept Perplexity answers changed what one of our own published results can be taken to mean.

WHAT WE FOUND

From 17 of 54 scored answers to 40 of 54. Perplexity's naming of one medicine rose sharply between consecutive days.

A broader change in response structure. The change appeared across all four question sets examined, including categories that track no diabetes or weight-management medicine.

Retained answers made the comparison possible. The record revealed the coincidence. Its cause, and how much it explains the rise in naming, remain unestablished.

On 6 October 2026, Perplexity, queried through its developer API, named one weight-management medicine in 17 of 54 scored answers to a fixed set of brand-neutral questions framed for the United Arab Emirates. The next day, it named it in 40.

A major diabetes congress, EASD 2026, had ended on 2 October, five days before the rise. Read alone, the rise might have looked like new medical evidence reaching AI answers.

The wider record complicated that story. Across all four question sets we examined, Perplexity's responses changed structure within the same interval, between 16:34 UTC on 6 October and 13:32 UTC on 7 October, including in the 11 categories of a fixed 12-category panel that track no diabetes or weight-management medicine; the twelfth tracks diabetes and weight-management medicines, semaglutide among them. ChatGPT and Claude, measured alongside it, showed no comparable change. The rise in naming coincided with a broader change in the responses. We could not establish what caused either, and the wider record prevents us from attributing the rise to the congress alone.

We could establish that coincidence because we had kept the earlier answers.

That is the whole case for independent audit, compressed into one week. Take away the earlier answer and every movement becomes a story anyone can tell: a congress worked, a campaign worked, a correction landed, a competitor lost ground. The audit is the thing that tests those stories, and the audit is made of kept outputs.

So here is the question for every AI provider, and it is not a new one. **If AI systems increasingly influence what the world knows, should the companies that run them also decide who may keep what those systems said, and who may check whether it was right?**

It sounds like a straightforward question. Increasingly, it is not.

AlphaCitation sells measurement of AI answers, so we have an interest in how it is answered. That is why every AlphaCitation finding cited here comes from a record we have already published; the note at the end states what the paper's public evidence package contains and withholds.

01 An audit is a comparison, and a comparison needs the past

When AI answers change, an auditor must consider at least four overlapping explanations:

1. **The evidence changed:** a new approval, label, guideline or study.
2. **The information pathway changed:** the engine began retrieving or citing different sources.
3. **The engine changed:** its model, configuration, retrieval system or answer format.
4. **Nothing meaningful changed:** ordinary variation between one run and the next.

The first is usually the one that matters to a patient, a regulator or a company. The other three are what make it hard to see. Investigating these explanations requires records beyond today's answer. Even then, the cause may remain uncertain.

EXPLANATION TO INVESTIGATE	THE AUDITOR MUST ALREADY HOLD
Ordinary variation	Repeated answers to the same questions over weeks
A change in the engine	What was asked and how; what the engine returned about its own response; answers to unrelated control questions on the same days
A change in the pathway	Which sources each answer cited, and when
A change in the evidence	A dated reading of the authoritative record, such as a regulator's register

Three of the four rows depend on records of what the engine itself said or returned. A term that bars keeping or analysing outputs does not make an audit harder. It leaves the auditor holding only the fourth row, which cannot explain a change on its own.

What a ban on keeping outputs would have made impossible

Our paper on whether evidence presented at EASD 2026 changed AI treatment answers was published on 12 October. Its first version saw that Perplexity's step came with "a change in its whole source mix", called it "consistent with a change on the engine's side", and conceded that the study "cannot separate it from the sources."

The change in response structure was established two days later, from records we already held, and published on 15 October. Its cause, and how much of the step it explains, are still not established. We re-examined the structure of every Perplexity response stored between 15 August and 13 October in the study's three series and in the 60-question panel. Between 16:34 UTC on 6 October and 13:32 UTC on 7 October, the responses changed in all four question sets within the same interval:

- On each day from 20 September to 6 October, at most 2% of stored answers carried exactly 10 search results. On each day from 7 to 13 October, at least 98% did.
- In the UAE series, search results per answer fell from 19.8 (4 and 6 October) to 10.0 (7–11 October), answers lengthened from 2,741 to 4,236 characters, and the share carrying section headings rose from 5% to 98%.

Naming did not move uniformly. It rose in the UAE series, fell in the Saudi Arabian series, which tracks a different medicine, and remained at 19 of 19 in the Singapore series on every day from 5 to 11 October. In the panel's 11 categories that track no diabetes or weight-management medicine, the share of question-and-brand pairs in which Perplexity named the brand rose from 41.9% on 6 October to 47.5% on 7 October.

The structural change was broader than the UAE finding.

EVIDENCE NOTE

Saudi Arabian series (teplizumab): 12 of 55 scored answers on 6 October, equal to its highest earlier count. Its admitted cycles from 24 August to 5 October ranged from 2 of 53 to 12 of 55, and from 3 to 6 in the four from 21 September. On each day from 7 to 11 October: between 0 and 2 of 55.

Panel, 11 categories: 41.9% (181 of 432) on 6 October and 47.5% (209 of 440) on 7 October. On every day from 7 to 13 October it was between 47.5% and 49.3%, above every day from 28 September to 6 October (41.9% to 45.0%). The comparison starts on 28 September, the first day on which Perplexity answered at least 54 of the 55 questions in these categories; on 20–27 September it answered 43 to 49 a day.

Same questions, other engines: ChatGPT moved from 46.1% (203 of 440) to 48.0% (211 of 440); Claude stayed at 46.6% (205 of 440).

On the record available, our request had not changed: no software was deployed between 5 and 12 October, and for every scheduled study cycle and panel run on both sides of the change, the instrument's code sent the same model, with the same parameters, to the same endpoint. The per-request parameters themselves were not stored. We published the finding on 15 October as a post-publication note, version 1.0.1 of the paper. It changed no count, registered verdict, method or figure. It changed what those numbers can be taken to mean.

Now run the same week under a term like the one in Microsoft's standard Bing grounding terms, which forbids storing the output at all.

Without retained responses, we could not have performed this later examination of their structure. A single live answer shows 10 search results; only the stored answers from the days before show that 10 was new. A narrower term, one that let an auditor keep only the counts it had planned in advance, might have preserved the original naming result. It would not have preserved features nobody had planned to analyse: no one had planned to count answer lengths, section headings or in-text markers. A less careful reader of the same numbers could have published the congress story.

And the records we had *not* kept are one reason the cause is still open. Perplexity's documentation states that "Sonar Chat Completions support ended on September 27, 2026" and that requests "are being reformulated as Agent API requests", "rolling out gradually by model". That is a candidate explanation, nothing more. Retaining the response identifier, the model field the engine returned and the usage block could have narrowed the possibilities and improved our ability to investigate; our instrument did not store them. Even with them, only Perplexity's internal change records could establish the precise cause.

That is the second lesson, and it is the one rule-writers most need to hear. **An auditor cannot know in advance which record will decide a case.** A permission that lets it keep only what it could justify beforehand risks losing exactly that record.

Nor can one answer show an engine contradicting itself

In a second published case, ChatGPT (OpenAI's gpt-4o-mini, queried through OpenAI's API) answered the same head-to-head question thirteen times between 22 August and 25 September 2026, in a fixed surveillance battery. In six of those answers it said a weight-management medicine registered in the UAE was not approved there. That claim is testable against the regulator's register, and it was false. In four more it said the medicine was not available there, a claim about supply that needs separate evidence. In the same cycles, other questions in the same battery drew the opposite from the same engine: the medicine "is approved for weight management in the UAE", with the pharmacies that sell it named.

Each answer, read alone, is a plausible sentence. A single answer can be checked against the regulator's record. Retained answers reveal how often the error recurred and whether the same system gave incompatible answers.

02 Scrutiny is not substitution

Google's grounding restriction lists "analyze" alongside "resell", and Microsoft's standard terms bar storing output at all; neither text separates using responses to build a substitute for a service from using them as evidence about it. Microsoft's enterprise version does draw that line. Audit permissions should draw it explicitly.

Learning from an output uses it as an input to something that can stand in for the provider: training or distilling a model, building a search index, resupplying answers as a product. The output's value moves into a substitute.

Learning about an output uses it as evidence about the system that produced it: how often it is right, whether it is consistent, when it changed and why. What leaves the audit is a statement about the engine, such as "called a registered medicine unapproved in 6 of 13 answers", not a reusable copy of what the engine produced.

Analysis is not the same as resale. Measurement is not extraction. Reporting an error rate is not reproducing an index.

The first kind of use can appropriate a provider's investment. The second can embarrass a provider. The first is the core harm usage restrictions exist to prevent. Reputational discomfort alone should not justify restricting otherwise responsible evaluation. Accountability aims to establish facts and support correction. Embarrassment may follow, but it is not the objective.

That gives every provider a one-line test for any contested use, which we call **the substitution test**: *does the use substitute for the provider's system, or scrutinise it?*

The substitution test should be the starting point, not the only safeguard. Providers also have legitimate interests in user privacy, third-party rights, security and the integrity of their systems, whether or not anyone is building a competitor. Independent auditing should be presumptively

permitted where it does not substitute for the underlying service, subject to proportionate protections for those interests.

None of this makes “research” a password. High-volume extraction, unauthorised redistribution, circumventing security controls and reverse engineering cause real harm, whatever they are called. The answer is not unrestricted access. It is a rule that can tell the auditor from the copier, and says so in writing.

03 What the terms actually say

Here is what five providers' published terms say about the outputs their customers receive. We quote; readers can judge.

PROVIDER	PUBLISHED TEXT (ACCESSED 16 OCTOBER 2026)
OpenAI, Services Agreement , updated 1 December 2025, effective 1 January 2026, §4.1	Customer "owns all Output. OpenAI hereby assigns to Customer all OpenAI's right, title, and interest, if any, in and to Output."
OpenAI, Sharing & Publication Policy , updated 14 November 2022	"We believe it is important for the broader world to be able to evaluate our research and products"; "we welcome research publications related to the OpenAI API."; "In others, such as publications that pertain to security or misuse of the API, we may want to take appropriate actions to protect our users."
OpenAI, Usage Policies , effective 29 October 2025	Lists "unsolicited safety testing" among the uses it does not allow, without defining it.
Anthropic, Commercial Terms of Service , effective 17 June 2025, §B	Customer "owns its Outputs"; "Subject to Customer's compliance with these Terms, Anthropic hereby assigns to Customer its right, title and interest (if any) in and to Outputs."
Perplexity, API Terms of Service , last updated 23 January 2026, §2.3.1 and §2.5	Customer "owns all Output"; Perplexity "asserts no ownership rights in any Output". Where a Customer Application is "likely to damage Perplexity's reputation or otherwise bring Perplexity into dispute", Perplexity may "deactivate the relevant API Key(s)."
Google, Gemini API Additional Terms of Service , effective 23 March 2026, last updated 28 April 2026	Generally: "Google won't claim ownership over that content." In the section "Grounding with Google Search": "You will not, and will not allow your end user or any third party to, cache, frame, syndicate, resell, analyze, train on, or otherwise learn from Grounded Results or Search Suggestions." The same section says it is a violation "to extract or collect" the parts of a grounded answer "for another purpose", with collecting links, building an index and crawling as its examples. The terms do permit storing the text of grounded results for up to two years, for a short list of named purposes. The restriction has been in these terms since at least December 2024, and the same words appear in Google Cloud's Vertex AI terms.
Microsoft, Terms of Use for Grounding with Bing Search and Grounding with Bing Custom Search and the version for enterprise integrations , both last updated November 2025	Outside an enterprise integration, the customer may not "Copy, store, cache, archive, or create a database of Output" (§4(b)(o)). Through an enterprise integration "such as Azure Foundry", the customer "may copy, store, cache or archive Bing Services Output solely (1) to the extent permitted by copyright law and (2) as an integrated part of its own work product, so long as the work product is not created for the purpose of or intended for use as a substitute for the Services or the Bing Services Output" (§3(c)). Both versions bar using the output "to train, evaluate or improve any AI Model" (Appendix, ¶12(b)).

Set side by side, the terms differ in what they permit, restrict and leave uncertain.

- **Ownership.** OpenAI, Anthropic and Perplexity give the customer ownership of its outputs.
- **Publication.** OpenAI welcomes research publications related to its API and says that for publications that pertain to security or misuse it “may want to take appropriate actions to protect our users”. Perplexity reserves the right to deactivate the API keys of an application it determines is likely to damage its reputation, and does not say whether a published finding unfavourable to it counts.
- **Testing.** OpenAI’s Usage Policies list “unsolicited safety testing” among uses not allowed, without defining it.
- **Express restrictions on grounded outputs.** Google’s Gemini API terms place “analyze” in the same list as “resell” and “train on” for Grounded Results, and call it a violation to “extract or collect” their components “for another purpose”. Microsoft’s Bing grounding terms, in both versions, bar using the output “to train, evaluate or improve any AI Model”. Microsoft’s standard version forbids storing the output at all; its enterprise version permits storage only “as an integrated part of its own work product” that is not “a substitute for the Services”.

The pattern is not one company’s, and it is not search grounding as such: Perplexity’s API searched on every request we sent, and OpenAI’s and Anthropic’s web-search tools sit under the ownership terms above. The restrictions on outputs appear in the two grounding products built on Google Search and Bing.

Owning an output is not the same as an express right to conduct longitudinal independent audits, to retain the evidence needed to defend published findings, and to keep access when those findings are unfavourable to the provider. In the documents reviewed, we found no express provision combining proportionate evidence retention for independent evaluation with protection against losing access because of adverse findings. The absence of such a provision is not a prohibition. Where the terms are silent, it leaves the auditor’s position uncertain.

That is a material gap in the standard arrangements reviewed here. Every auditor of AI answers working under them is working in that gap. What follows is a proposal for clearer rights, not a claim that any existing term may be disregarded.

A researcher should not have to guess whether independently measuring the reliability of AI-generated information is a permitted use of an AI service.

04 We are not the first to ask, and we join those who did

The pattern is older than generative AI. By the account Databricks gave in 2021, when it removed its own version, a 1982 benchmark study by the database researcher David DeWitt “showed that Oracle’s system had poor performance”, and the licence clause that followed,

barring the publication of benchmarks, became “a standard feature of most database vendors’ license agreements” and still carries his name. Databricks, itself an industry participant, called the clause anti-competitive and asked the rest of the industry to drop it.

Platforms came next. In August 2018 the Knight First Amendment Institute at Columbia University asked Facebook to “amend its terms of service to establish a ‘safe harbor’ for public-interest journalism and research” on the platform. Facebook discussed the proposal and rejected it.

Then AI. In March 2024, Shayne Longpre, Sayash Kapoor, Kevin Klyman and twenty colleagues, among them Arvind Narayanan, Percy Liang and Peter Henderson, published [A Safe Harbor for AI Evaluation and Red Teaming](#), presented at ICML 2024. They argued that providers’ terms and enforcement can deter good-faith research through the threat of suspension or legal action, and asked AI developers to commit to a legal and a technical safe harbor. More than 350 researchers, journalists and advocates have signed the accompanying open letter.

In 2025, Peter Henderson and Mark Lemley’s [The Mirage of Artificial Intelligence Terms of Use Restrictions](#) (*Indiana Law Journal*) called the legal enforceability of such terms “questionable”, argued that they “chill good faith research” and “constrain competition”, and warned that they “create quasi-copyright ownership where none should exist”.

The principle has already been conceded, on paper. Under the EU’s General-Purpose AI Code of Practice, published in July 2025, signatories commit that they “will not take any legal or technical retaliation against the independent external evaluators as a consequence of their testing and/or publication of findings”, on five conditions, including a disclosure procedure that cannot hold publication back for more than 30 business days, unless a longer timeline is exceptionally necessary. Google, OpenAI, Anthropic and Microsoft are listed among the Code’s signatories. But that protection covers systemic-risk evaluation of frontier models, by evaluators chosen against criteria the provider publishes. It does not reach the auditor nobody selected, measuring the ordinary answers people act on.

Statutory protection is fragmented. The EU’s Digital Services Act gives vetted researchers access to data from designated very large online platforms and search engines, for research into specific systemic risks, and only if they are independent of commercial interests. It is a data-access route, not a general regime for independent AI auditing. Existing protections do not reliably accommodate a commercial organisation conducting rigorous, independently verifiable audits of the information AI systems deliver.

We add our name to that line of argument. This proposal builds on the case for a safe harbor for independent AI evaluation. It makes evidence retention explicit, including records whose significance becomes apparent only after an initial finding is published. The same terms govern something ordinary and widespread: the longitudinal measurement of the *information* AI systems deliver, by an auditor nobody selected. A medicine wrongly ruled out. A congress result that may or may not have reached the answers. An established business that stops appearing in AI comparisons while its competitors are still named. Each needs systematic observation over time to explain, and each is invisible without kept answers.

05 Independent scrutiny finds what fluency hides

In March 2025, Columbia University's Tow Center for Digital Journalism tested eight generative search tools on 1,600 queries about news articles ([AI Search Has a Citation Problem](#)).

"Collectively, they provided incorrect answers to more than 60 percent of queries." The tools fabricated links, and were generally bad at declining questions they could not answer accurately.

Those numbers concern that study's task, identifying the headline, original publisher, publication date and URL of the news article an excerpt came from, and set no general error rate. They show something more durable: reliability has to be demonstrated, not inferred from fluency, citations or the reputation of the company behind the model.

Providers know this. In August 2026, Google DeepMind [piloted a double-blind evaluation](#) with the Singapore AI Safety Institute, OpenMined, AVERI and MLCommons, using confidential computing so that the evaluator could not see the model's weights and the provider could not see the evaluator's test questions. It is a serious answer to a real problem, and it illustrates the point this essay depends on: independent scrutiny and proprietary protection can coexist.

Programmes of this kind do not settle the rights of organisations that work under standard terms and that no programme selected. An accountability system needs both. The principle is the same for each:

Protect the intellectual property. Protect the evaluation instrument. Permit responsible measurement of observable AI behaviour.

06 Uncertainty is a policy, even when nobody wrote it as one

Nobody needs to sue an auditor for a vague term to work. Uncertainty does the work on its own. Smaller organisations cannot afford to test a restriction. Researchers avoid the questions that might cost them access. Findings get published with less evidence than the science requires, or not at all.

The result is an accountability system with three tiers. Providers keep extensive internal measurement. Selected partners receive negotiated evaluation access. Everyone else works under whatever the standard terms leave unsaid.

Public institutions are moving the other way. The US [NIST AI Risk Management Framework](#), released in January 2023, makes measurement one of the four core functions of AI risk management. In June 2026 the European Commission [appointed a Scientific Panel of 60 independent experts and an Advisory Forum](#) to support AI Act enforcement; the panel's remit

includes “evaluation methodologies”. None of this overrides a private contract. All of it says independent evaluation is becoming a condition of responsible deployment, and the contracts should catch up.

The measurement industry is not exempt. A growing number of companies, ours among them, sell measurement of AI answers: visibility monitoring, citation analysis, recommendation intelligence. If one of them tells a client that its work raised a brand’s AI recommendation rate, it should be able to show the baseline, the method, the uncertainty and the evidence for attribution. An industry built on measuring AI must be able to show that its own measurements survive scrutiny. That, too, requires kept records and the freedom to publish them.

07 The proposal: an Audit Clause any provider can adopt

The 2024 safe harbor asked providers for a commitment. We are asking for a clause in their terms. Here is the text, short enough to adopt as written:

THE CLAUSE, AS PROPOSED

Independent evaluation. You may use the Services and their outputs to evaluate the accuracy, consistency, sourcing and behaviour of the Services, and to publish your methods and findings, including findings unfavourable to us.

For that purpose you may securely retain, before and after publication, the requests, outputs, cited-source records and response metadata of your evaluation, for a documented retention period proportionate to producing, substantiating, reproducing and correcting your findings and to your legal obligations.

You will protect retained material against unnecessary disclosure of personal data or third-party confidential information.

This permission does not allow you to train or improve any model on the outputs, to build an index or dataset that substitutes for the Services, to supply the outputs to others as a product, to republish outputs in full beyond short excerpts that evidence a finding, or to circumvent usage limits or security controls.

This clause prevails over any conflicting provision of these terms or of our usage and acceptable-use policies, except provisions that protect the security of any system, protect personal data or prohibit unlawful conduct; a provision that restricts testing, evaluating or analysing the Services or their outputs is not such a provision.

We will not suspend or end your access solely because an evaluation within this clause reports findings unfavourable to us. This does not limit action for a genuine breach of those excepted provisions.

The clause permits independent evaluation and proportionate retention of its evidence. It preserves protections for personal data, third-party confidential information, usage limits and system security, and the prohibition of unlawful conduct. Its precedence provision stops any other provision of the terms or usage policies, except those that protect system security or personal data or prohibit unlawful conduct, from withdrawing the permission it grants.

It permits learning *about* the outputs. It does not permit training on them, building a substitute index or dataset, resupplying them as a product or republishing them in full.

The proposal asks providers to protect an auditor's ability to keep the record and publish what it shows, including findings unfavourable to the provider. Express survival after the service ends, controlled access for independent verification and broader protection against retaliation are worthwhile further steps; they are policy choices beyond this minimum.

So the question to the industry is no longer abstract. Any provider can adopt this clause. A provider that will not should say which sentence it rejects, and why.

08 An industry that asks for trust must accept verification

AI now influences which sources people trust, which products they consider and which evidence they encounter. That influence carries a duty to let others distinguish what a system does from what its developer says it does. Independent scrutiny is how that distinction becomes visible. It exposes errors, it validates genuine improvement, it corrects inflated claims, including claims made by companies that sell measurement.

We hold ourselves to the standard we are asking for. We publish our methods, our deviations and our evidence, and when new evidence changes what a finding means we say so in public, as we did on 15 October.

No provider should be able to turn the silence of its terms into a boundary around independent scrutiny of what its systems tell the world. The issue is not whether every AI system must be open to every imaginable test. It is whether responsible independent measurement becomes a recognised, protected and ordinary activity, or remains something done at the provider's pleasure.

Retained answers did not show why Perplexity's responses changed. They showed that our own published result needed qualification.

Independent scrutiny needs the capacity to revisit the record, and explicit protection to preserve it.

HOW WE KNOW THIS

What was measured. Developer API endpoints with web search, queried in repeated cycles over several weeks. The two cases above concern Perplexity's Sonar API (model `sonar`), OpenAI's `gpt-4o-mini` and Anthropic's Claude Haiku 4.5. These are not the consumer apps, which may answer differently. The market is set in the question text.

The Perplexity case. AlphaCitation Research Series, Paper No. 01, Section 5.1 and post-publication note PN1 in Section 6.2, first published in version 1.0.1 (15 October 2026) and corrected in version 1.0.2 (16 October 2026). Version 1.0.2 corrected the note on four points: one of the panel's 12 categories does track diabetes and weight-management medicines, among them semaglutide; the study's Saudi Arabian series is its post-registration arm, not a question set unrelated to the congress; the request parameters changed on 27 September, before the change in the responses, and were the same for the scheduled study cycles and panel runs on both sides of it; and the caption of Figure 13 had widened the note's scope to every question set the instrument ran, where the note examined the study's three series and the panel. It also added the figures for the 11 categories that track no such medicine. No count, registered verdict, method or figure changed. The panel naming figures above are in `addenda/PN1-naming-panel-11.csv`; the note's 41.5% to 47.2% covers all 12 categories.

The ChatGPT case. AlphaCitation Field Note 06, "The Medicine Was Approved. ChatGPT Ruled It Out Anyway." (5 October 2026).

What can be checked, and what cannot. The paper's public evidence package holds the daily counts behind the post-publication note (`addenda/`), a per-answer outcome table without question or answer text, the coding files and their ledgers, the analysis outputs, and a verification script (`code/verify.ts`) that checks the coding files' totals against their rows and ledgers. It withholds the question batteries, full answer text, the scorer's internals, the request parameter introduced on 27 September and 13 of the study's 15 scripts, and it redacts local-business identifiers. The batteries are withheld to reduce the risk of the measurement being optimised against. This limits independent replication and review of the question wording.

What this essay does not contain. No question text, and no answer text beyond the eight-word phrase quoted from reference 2.

This Perspective states a policy position on independent evaluation. It is not a legal determination of the applicability or enforceability of any service agreement.

Scholarly correspondence and substantive commentary: analyst@alphacitation.com

REFERENCES

1. Ezeife OC. [Does New Medical Evidence Change AI Treatment Answers?](#) AlphaCitation Research Series, Paper No. 01, version 1.0.2. 16 October 2026 (version 1.0.1: 15 October 2026; version 1.0: 12 October 2026).
2. AlphaCitation. [The Medicine Was Approved. ChatGPT Ruled It Out Anyway.](#) AlphaCitation Field Note 06. 5 October 2026.
3. Perplexity. [Migrate from Sonar.](#) Perplexity API documentation. Accessed 16 October 2026.
4. OpenAI. [Services Agreement](#), updated 1 December 2025, effective 1 January 2026, §4.1; [Sharing & Publication Policy](#), updated 14 November 2022; [Usage Policies](#), effective 29 October 2025. Accessed 16 October 2026.
5. Anthropic. [Commercial Terms of Service](#), §B. Effective 17 June 2025. Accessed 16 October 2026.
6. Perplexity. [API Terms of Service](#), §2.3.1, §2.5. Last updated 23 January 2026. Accessed 16 October 2026.
7. Google. [Gemini API Additional Terms of Service](#), section “Grounding with Google Search”. Effective 23 March 2026; last updated 28 April 2026. Earlier versions: [terms archive](#). Accessed 16 October 2026.
8. Microsoft. [Terms of Use for Grounding with Bing Search and Grounding with Bing Custom Search](#), and the [version available through Microsoft Enterprise Product Integrations](#). Both last updated November 2025. Accessed 16 October 2026.
9. European Commission. [General-Purpose AI Code of Practice](#), Safety and Security chapter, Measure 3.5. 10 July 2025.
10. Olsson J, Xin R. [Eliminating the Anti-competitive DeWitt Clause for Database Benchmarking](#). Databricks. 8 November 2021.
11. Knight First Amendment Institute at Columbia University. [Knight Institute Calls on Facebook to Lift Restrictions on Digital Journalism and Research](#). 7 August 2018. Case page: [Facebook Restrictions on Digital Journalism and Research](#). Accessed 16 October 2026.
12. Longpre S, Kapoor S, Klyman K, et al. [Position: A Safe Harbor for AI Evaluation and Red Teaming](#). *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 2024: 32691–32710. Preprint: [arXiv:2403.04893](#). Open letter: [A Safe Harbor for Independent AI Evaluation](#).
13. Henderson P, Lemley MA. [The Mirage of Artificial Intelligence Terms of Use Restrictions](#). *Indiana Law Journal*. 2025;100(4):1327.
14. Regulation (EU) 2022/2065 (Digital Services Act), Article 40, [Data access and scrutiny](#).
15. Jąźwińska K, Chandrasekar A. [AI Search Has a Citation Problem](#). *Columbia Journalism Review*, Tow Center for Digital Journalism. 6 March 2025.
16. Google DeepMind. [Piloting the World’s First Double-Blind AI Evaluations](#). 27 August 2026.
17. National Institute of Standards and Technology. [Artificial Intelligence Risk Management Framework \(AI RMF 1.0\)](#). 26 January 2023.
18. European Commission. [AI Act Enforcement Gets Independent Expert Support](#). 1 June 2026.

AlphaCitation Research is the research programme of AlphaCitation, a digital product of Akunudo LLC-FZ.

© 2026 AlphaCitation (Akunudo LLC-FZ). Akunudo LLC-FZ is the copyright holder and licensor of the original material in this publication, licensed under [CC BY 4.0](#). Quoted third-party text, third-party data fields, API/provider-restricted material and third-party marks are excluded from that grant.