RESEARCH COMMENTARY / AI & INFORMATION

An AI Answer With Citations Still Needs Checking

A reference is valuable when a reader can follow it to evidence that supports the claim.

Jennifer Oriwoh

A link beside a confident sentence changes how an answer feels. It suggests that someone has done the checking. For a busy reader using an AI-generated research summary, that suggestion can be enough to move the sentence into a presentation, a proposal or a decision.

Yet several different things must be true before the reference earns that trust. The source must exist. It must contain the relevant evidence. The answer must represent that evidence accurately, and the evidence must apply to the question being asked.

Treating those as separate checks makes a source trail more useful. It also reveals why simply asking an AI system to include citations does not finish the research.

A real source can support the wrong sentence

Consider a hypothetical summary claiming that a tool improves productivity across small businesses. Its citation leads to a genuine study of a specific task, among a particular group, under controlled conditions. The link works. The broad claim still goes beyond the evidence.

The mistake may be a changed population, a missing comparison or a stronger verb than the study permits. A report that documents an association cannot automatically support a claim that one factor caused another. A vendor's account of a successful customer does not establish what a typical buyer should expect.

These are familiar research problems, now appearing inside a fast, convenient interface. Checking them requires reading the source rather than inspecting the neatness of the answer. The original date, location, sample and method may matter more than the number of links displayed.

NIST's Generative AI Profile identifies confidently stated false content as a risk and discusses fabricated supporting citations. Its treatment explains why presentation alone is weak evidence of correctness. It does not provide an error rate for every current tool or eliminate the need to test a particular use case.

Measure support, not the appearance of research

The distinction has also been examined in language-model research. In the 2023 paper *Enabling Large Language Models to Generate Text with Citations*, Tianyu Gao and colleagues introduced ALCE, a benchmark that evaluates fluency, correctness and citation quality separately. Its citation measures examine whether statements are supported and whether cited passages are relevant to that support.

That separation is the useful lesson here. A readable answer can fail a support check. A correct statement can have a poor citation. Counting references alone cannot distinguish those outcomes. The paper's experiments concern the systems and datasets studied in 2023; they should not be presented as a measurement of every product available today.

For an organisation evaluating a research assistant, the test should reflect the documents it will actually use. Include a current policy, an older version, an ambiguous request and a question the documents cannot answer. Decide what adequate support looks like before reviewing the generated responses.

Then inspect the important claims individually. Does the cited passage contain the information? Is an exception omitted? Does the answer combine separate facts into a conclusion the source never made? Record an unsupported claim even if another part of the answer is useful.

Make uncertainty usable

A research tool should help a reviewer see where its evidence ends. When documents disagree, an answer that identifies the disagreement may be more useful than one that silently chooses a side. When the answer depends on information outside the available material, a clear limitation protects the next decision.

This does not mean requiring exhaustive verification of every low-stakes sentence. Review effort should match consequence. A rough list of topics for further reading can tolerate a different process from a customer promise, a procurement recommendation or a public statement attributed to an executive.

For consequential work, a compact claim record can help. Keep the proposed statement, the original source, the relevant passage, its date and any qualification needed. The record should distinguish what the source reports from the interpretation the writer adds. It gives an editor a practical route to review the argument without starting from a pile of undifferentiated links.

Access matters as well. A reader needs to know when a source is unavailable or only an abstract has been consulted. A summary should not imply that a full study was examined when the evidence available was a short description. Source transparency includes the limits of the writer's own access.

The strongest use of an AI research assistant is to make evidence easier to find and examine. Its speed becomes valuable when the saved time supports better judgment, rather than allowing an unsupported claim to travel further before anyone opens the link.

Sources

NIST, Generative Artificial Intelligence Profile, section 2.2, Confabulation.

Gao, Yen, Yu and Chen, Enabling Large Language Models to Generate Text with Citations, EMNLP 2023, sections 2-3.