SEPTEMBER 2026

Evaluating Evidence Sufficiency in Frontier Language Models

Implications for AI Evaluation, Deployment, and Youth

"[AI] risk management can enable AI developers and users to understand impacts and account for the inherent limitations and uncertainties in their models and systems, which in turn can improve overall system performance and trustworthiness and the likelihood that AI technologies will be used in ways that are beneficial."

National Institute of Standards and Technology, *AI Risk Management Framework 1.0* (2023), p. 4.

A joint publication of the **Institute for Youth in Policy** and the **Frontlines Foundation**.

Contact Email: nathanielnadler@yipinstitute.org

Executive Summary

Artificial intelligence systems are often asked to answer questions from retrieved webpages, uploaded documents, research papers, and institutional records. In these instances, access to a relevant source is only part of the reliability problem. Particularly, the model must also determine what the source actually supports. That is, a claim may be directly supported, directly contradicted, or simply left unresolved by the available evidence. The third case is far easier to miss than the former two because an unsupported conclusion can be coherent, plausible, and even closely related to the source.

This study tests whether two frontier language models can reliably distinguish among those three claim categories. A controlled benchmark paired short, headline-like passages that were labeled as either *True*, *False*, or *Insufficient Evidence*. GPT-5.6 Sol and Claude Opus 5 evaluated the items across repeated, stateless runs. Due to the fact that the relevant evidence appeared directly in each prompt, the benchmark was designed to examine evidence interpretation rather than pure factual recall.

Both systems performed strongly overall, with an accuracy of 92.7 percent for GPT-5.6 Sol and 92.8 percent for Claude Opus 5. Nevertheless, the aggregate result (i.e., the accuracy score alone) conceals important features of the results. Accuracy neared 99 percent when a claim was directly supported, remained above 92 percent when it was contradicted, and fell to 83.6–86.1 percent when the correct judgment was that the evidence was insufficient. The direction of error was also consistent, with roughly 14 percent of insufficient-evidence claims being classified as true by each model, while incorrect rejection was uncommon.

The central policy concern is not that the systems usually failed; instead, it is that high overall accuracy concealed a repeatable tendency to convert plausible inference into affirmative judgment. This distinction matters for evaluation, procurement, product design, research, and education. More broadly, a model should not receive full credit for displaying a relevant citation if the sentence attached to that citation goes beyond what the source establishes.

Due to the scope of the study, the recommendations in this brief are correspondingly narrow. Ultimately, this study's results suggest several key recommendations across various different applications of LLMs:

1. Evaluators should treat insufficient evidence as a distinct category, report the direction of error, and separate retrieval quality from evidence-use quality.

2. Product teams should make claim-to-source relationships easier to inspect and permit explicit statements of insufficiency.

3. Schools and youth-oriented institutions should teach evidence boundaries as part of normal research practice.

1 Introduction

Many discussions of factual reliability begin with a sensible question: did the model get the fact right? For models that answer from a bounded set of sources, that framing is inherently incomplete. In practice, a claim can be supported by the evidence, contradicted by it, or left unresolved. The last category is not simply another form of falsehood or a nebulous state between truth and falsehood; rather, it marks the point at which the available material stops.

Consider the following illustrative example: suppose a source states that a state health agency launched a vaccination program across 23 pharmacies. The claim that 23 pharmacies joined is supported; a claim that 32 pharmacies joined is contradicted; and a claim that the program increased vaccination rates belongs in a different category. It may very well be a reasonable hypothesis, and later evidence may confirm it, but the source itself does not establish the result.

The demonstrated distinction becomes more significant as language models work from search results, uploaded files, legal materials, academic papers, and internal records. Providing the relevant document removes one obvious source of error, as the model no longer has to rely entirely on information stored in its parameters. However, access to the document does not guarantee that every conclusion drawn from it is warranted, as a relevant citation can accompany an unsupported inference with just as much ease as a supported statement.

The problem can be divided into two contributing questions. First of all, did the system obtain the evidence needed to answer? Secondly, once the evidence was available, did the system limit its conclusions to what that evidence justified? While retrieval models primarily address the first question, this brief is more focused on the second. Generally, it asks: *when frontier language models are given the relevant evidence, how reliably do they distinguish what the evidence supports, what it contradicts, and what it simply does not establish?*

The objective is not to rank the two models against each other. In practice, the benchmark uses them to examine whether evidence sufficiency is a distinct point of weakness that aggregate accuracy can hide. That question relates most directly to model evaluation and deployment, and it also has consequences for users who may have limited ability to independently check an answer that merely sounds correct, including students encountering material that they are new to or unfamiliar with.

2 Evaluation Design

Generally speaking, the benchmark was designed to isolate evidence interpretation from general factual recall. Each example contains a short, neutral, news-style event description and an associated claim (in the form of a news headline). The claims fall into three categories: a **True** claim is directly supported by the passage, a **False** claim conflicts with the passage, usually because a limited detail such as a number, date, institution, location, or action has been changed, and an **Insufficient Evidence** claim remains plausible but introduces a causal interpretation, motive, future consequence, broader effect, or other detail that the passage does not establish.

Consequently, the insufficient-evidence category is central to the design. Furthermore, these claims were not intended to be absurd or obviously fabricated. Instead, they were written so that plausibility and evidentiary support could come apart. That is, the model must recognize both the contradiction and also the absence of a sufficient basis for judgment.

GPT-5.6 Sol and Claude Opus 5 each received the evidence passage and candidate claim and returned one of the three labels. Evaluations were stateless and repeated independently. The primary measures were overall accuracy, class-specific accuracy, the direction of error on insufficient-evidence items, and agreement across repeated runs.

Table 1: Evaluation design

Component	Specification
Models	GPT-5.6 Sol and Claude Opus 5
Task	Three-way claim classification
Input	Controlled evidence passage and candidate claim
Output	True, False, or Insufficient Evidence
Context	Stateless evaluation; independent repeated runs
Primary measures	Overall accuracy, class-specific accuracy, error direction, and run-to-run agreement

3 Results

Interestingly, at the level of aggregate results, the two systems performed almost identically. Specifically, overall accuracy was 92.7 percent for GPT-5.6 Sol and 92.8 percent for Claude Opus 5. Overall, those figures indicate fairly strong performance, but they do not offer much substantive information about the types of errors that occurred.

Table 2: Accuracy by ground-truth category

	GPT-5.6 Sol	Claude Opus 5
Overall	92.7%	92.8%
True	98.1%	98.8%
False	95.8%	92.8%
Insufficient evidence	83.6%	86.1%

As suspected, the class-specific results reveal a noticeable asymmetry. Specifically, both models were nearly perfect when the evidence directly supported the claim. Similarly, performance on false claims remained high. On the other hand, the largest drop in accuracy appeared when the best judgment was that the evidence did not resolve the claim. This is evident in figure 1, which shows that the evidence-sufficiency category was the central source of weakness for both models.

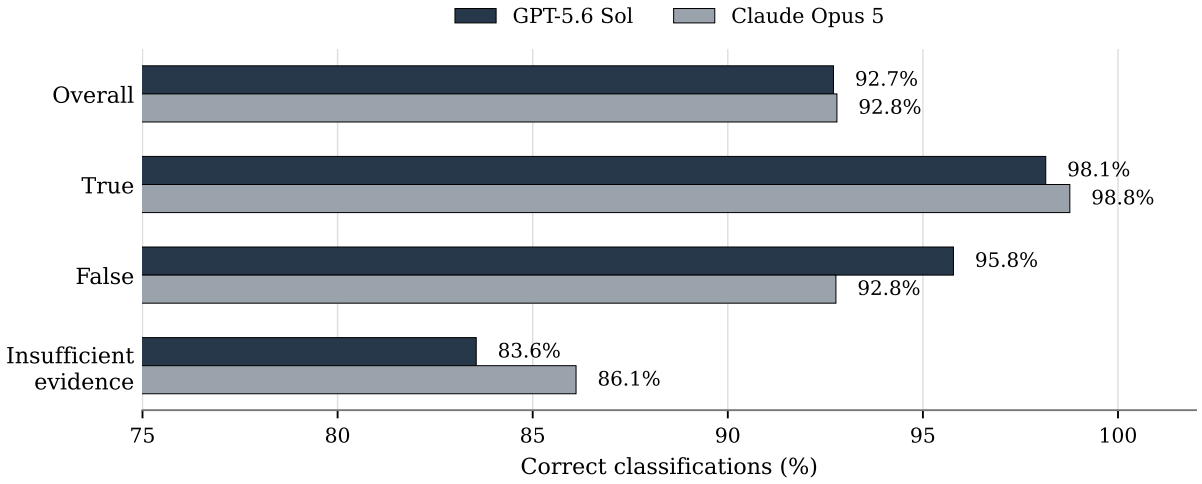


Figure 1: Accuracy by ground-truth category. Both systems performed near ceiling on directly supported claims and materially worse when the evidence did not establish the claim.

To be clear, this pattern does not suggest that the models generally failed to read straightforward evidence. In fact, that is directly contrary to the results: their performance on directly supported statements was extremely strong. The bigger difficulty was deciding when a reasonable-sounding conclusion exceeded the information provided. For example, changing a number creates a very explicit conflict between the claim and the evidence provided. On the other hand, a claim about a program's effects, a company's motive, or a policy's future consequences may very well be coherent, even if they do not fit into the story told by the evidence provided. In those cases, the central challenge for the model is that it must resist finishing the story for the user, as that jumps over the bounds of what is substantiated.

Further, the direction of error adds a second asymmetry. This emerges because an unresolved claim can be mishandled in two ways. For example, a model may call a claim false, treating an absence of evidence as evidence against the claim, or it may call it true, treating plausibility as if it were active support. In this study, the second error tended to dominate. Concretely, GPT-5.6 Sol classified 13.8 percent of insufficient-evidence claims as true and 2.6 percent as false, and Claude Opus 5 classified 13.9 percent as true and none as false in the supplied responses.

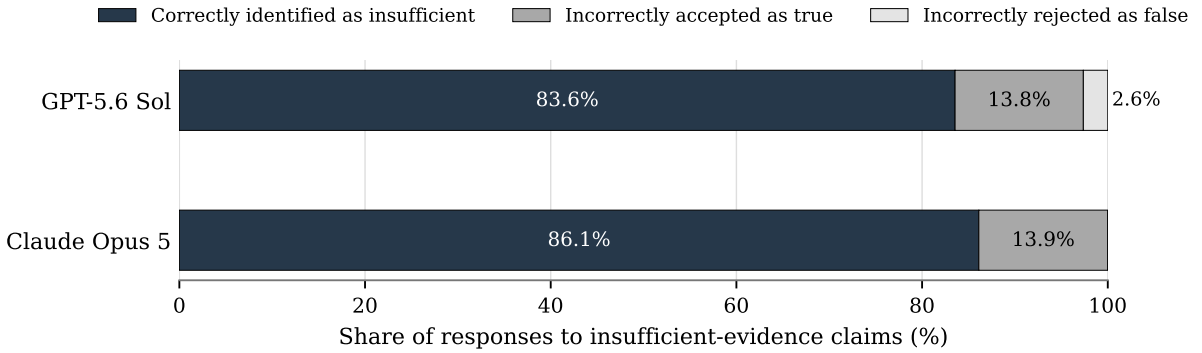


Figure 2: Responses to claims whose correct label was *Insufficient Evidence*. The principal error was acceptance: approximately one in seven unresolved claims was classified as true.

Run-to-run agreement was also high in both models: 98.8 percent for GPT-5.6 Sol and 97.0 percent for Claude Opus 5 in the repeated evaluations. While consistency is certainly useful, it is not the same as correctness. This can imply that a model is predictable, but it does not necessarily point to such a model being reliable. At the same time, the agreement rates suggest that the observed pattern cannot be dismissed as a result of sampling noise; that is, some errors appear to reflect stable judgments about what the evidence permits.

4 Policy Analysis

Search, retrieval, uploaded documents, and institutional bases of knowledge are often treated as important safeguards against factual error because they place relevant material in front of a model before it answers. To be clear, that is a meaningful improvement over asking a model to simply rely on information encoded in its parameters. Still, it is important to acknowledge that retrieval and evidence use are different stages of the same process. The worry is that a model may locate the right document and still draw a conclusion that the document does not establish. As a result of the benchmark supplying the relevant evidence directly, it is able to isolate this second problem: whether the model can recognize the boundary between what a source says and what a reasonable interpretation adds.

That distinction matters for AI governance due to the fact that source-grounded systems are often evaluated through visible signs of grounding—retrieved documents, citations, links, or quoted passages. Those features can show that relevant evidence was available, but alone, they do not show that each material claim was substantiated. A more complete evaluation should ask two varying questions: whether the system had access to the information needed to answer, and whether the answer remained within the evidentiary limits of that information. This demarcation is consistent with broader approaches to risk-management that treat testing, evaluation, and documentation as continuing parts of model governance rather than as a single check of accuracy.[1]

The emergence of class asymmetry in this benchmark also demonstrates why aggregate accuracy can be an incomplete description of reliability. For example, two models can receive nearly identical overall scores while making different kinds of mistakes. In practice, a model that occasionally misses an explicit contradiction behaves differently from one that tends to turn unresolved propositions into affirmative claims. The latter failure is easy to overlook because the output may remain understandable, relevant, and reasonable. For policy and institutional evaluation, the important underlying question is what kind of uncertainty a model is most likely to erase when it is wrong.

The direction of error is especially significant in settings where users ask models to synthesize rather than simply retrieve. It goes without saying that questions about causes, likely effects, motives, legal implications, scientific interpretation, or policy consequences often require reasoning beyond a single quoted sentence. Thus, in such cases, inference is not a defect or concern; rather, it is a requisite of useful analysis. Still, the governance problem arises when an inference is presented with the same evidentiary status as a fact that is directly supported. A system that can distinguish "the source states this" from "this is a reasonable inference from the source" gives users far more information about how much trust to place on the answer.

This distinction also changes how procurement and deployment testing should be understood. A model's behavior in a controlled evidence task is not a complete measure of an end-to-end product resulting from the fact that deployed systems also depend on retrieval, prompting, system instructions, interface design, and the documents supplied to them. At the same time, a controlled evidence-use test can help identify where a failure occurs. On one hand, if the correct source was never retrieved, the problem lies upstream. Conversely, if the source was present and the system still overstated what it established, the relevant weakness lies in evidence interpretation or in the way the product encourages the model to express uncertainty. Separating these failure modes makes evaluation more diagnostic and diminishes the allure of treating every error as the same kind of "hallucination."

Citation design also deserves similar scrutiny. For example, a list of sources on the paragraph level can be

[1]See National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023), and NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (2024), both of which frame measurement and evaluation as components of AI risk management.

useful for navigation, but it can also create an appearance of verification when only part of the paragraph is directly supported. The user will often see a real source, recognize that it concerns the same topic, and may plausibly assume that the source substantiates the facts supplied. On the other hand, traceability on the level of claims provides a stronger signal because it lets the user inspect the relationship between a particular proposition and the passage offered in support of it. Accordingly, this concern sits alongside existing transparency principles that emphasize giving users meaningful information about the basis of AI outputs.[2]

The policy relevance is wider than any one sector. The identified evidentiary distinction can have significance wherever AI is used to summarize records, compare documents, explain research, assist with analysis, or answer questions from a bounded body of information. Nevertheless, the appropriate safeguards will differ by context: a brainstorming tool and a model used in a consequential institutional workflow do not need to be evaluated with the same rigor or tools. The notion that transfers across settings is the value of knowing whether a generated proposition is directly supported, contradicted, or unresolved by the material the system was asked to use.

Youth and education remain an important application of that broader point, but they should not define the entire analysis. This benchmark did not test young users and does not establish that students are uniquely susceptible to unsupported inference. Nevertheless, the relevance is deeply practical: learners often use AI when they do not yet possess enough subject knowledge to independently assess every claim. The familiar advice of "checking the source" is useful, but it stops one step early. A more helpful habit is to ask what exactly the source establishes, where the model begins to synthesize, and what additional evidence would be needed before an inference should be treated as a concrete fact. This has always been a conventional research skill rather than an AI-specific rule, but it remains all the more important even as model strength improves.

5 Recommendations

The findings do not support sweeping conclusions about every model or deployment. However, they do support a set of focused changes to evaluation, system design, procurement, and education. Ultimately, these recommendations preserve the original distinction at the center of the benchmark: supported, contradicted, and unresolved are different evidentiary states and should be treated with such distinction in mind.

1. Treat insufficient evidence as a separate evaluation category. Model evaluations should distinguish between claims that are supported, contradicted, and non-determinable by the supplied material. Binary true-or-false benchmarks certainly remain appropriate for some tasks, but they are often incomplete when a system is expected to reason from bounded sources. Thus, forcing every claim into a binary label removes the behavior that matters most in this setting: recognizing when the evidence stops. Hence, evaluators should report class-specific performance alongside overall accuracy and should include plausible, non-absurd claims that require the model to abstain from a definitive truth value rather than opting for falsehood.

Practical application. Benchmark developers can add an insufficient-evidence class to document-grounded tasks, publish the item-writing rules used to separate unsupported inference from contradiction, and conduct adjudication on the level of items before releasing definitive estimates.

2. Report the direction of error. Evaluations should demonstrate what happens when a model misses an unresolved claim. At a minimum, reports should demarcate correct insufficiency judgments, unsupported acceptance, and unsupported rejection. The issue is that, in practice, these outcomes are not interchangeable. For example, over-acceptance indicates that a system is willing to convert plausibility into affirmation; over-rejection indicates that it is treating missing evidence as disproof. Each distinct pattern requires different prompting, training, calibration, or interface responses. Reporting only a single error rate conceals that distinction. The ramifications of specific error types vary across model implementations, making it all the more important that such error directions are clearly communicated.

Practical application. Reports should include a small confusion matrix or an error-direction table for source-grounded tasks. When sample sizes permit, the same breakdown should be reported by claim subtype, such as

[2]The OECD AI Principles, updated in 2024, call for transparency and meaningful information that can help affected users understand AI-generated outputs and, where feasible, the sources or factors underlying them.

causal, predictive, motivational, quantitative, or legal inference.

3. Evaluate retrieval and evidence use separately. Organizations testing retrieval-augmented or document-grounded systems should divide the evaluation into two distinct parts. The first should ask whether the system found the material that was needed to answer. The second should ask whether, given that material, the system made claims that were warranted by it. This distinction matters because a failure at the first stage points toward search, indexing, chunking, or ranking, whereas a failure at the second cannot be solved merely by retrieving another relevant document. Hence, combining the stages into a single score makes it difficult to identify which part of the model requires improvement.

Practical application. Procurement and internal assurance processes should include a controlled evidence-use test in which the relevant source is supplied directly. This isolates inference quality from retrieval quality and provides a clearer basis for comparing configurations, prompts, and model updates.

4. Design for claim-level traceability and calibrated insufficiency. Products that present answers that are grounded in a source should make the relationship between material claims and supporting passages easier to inspect. To be clear, this does not require every interface to resemble a formal research platform. For example, a product can link a sentence to a specific excerpt, distinguish direct support from synthesis, or mark a conclusion as an inference–which some models have partially started implementing. It should also allow the model to clearly state that the supplied materials are insufficient. Specifically, models that reward confident completion while hiding evidentiary limitations make the observed failure a bit harder for users to detect.

Practical application. Product teams should test whether users can identify the source passage for each major claim and can distinguish a directly supported statement from an inference generated by a model. A useful design review should ask whether citations are proposition-specific and whether unsupported conclusions are specifically qualified.

5. Teach evidence boundaries as a general research skill. Schools and youth-serving organizations should place evidence evaluation within AI literacy, but the recommendation is broader than youth policy. The relevant habit is not "never trust AI," nor is it simply "always check the source." Those conclusions would be imprudent. The central lesson is that users should learn to separate topical relevance from evidentiary support. For instance, a source can discuss the same subject without proving the conclusion attached to it. This is a normal research skill whose importance increases when a system can produce reasonable-seeming synthesis faster than any user can verify it.

Practical application. Assignments and exercises can require users to identify the exact passage supporting a generated claim. When no passage can be located, the statement should be labeled as an inference, hypothesis, or unsupported claim, and the user should specify what additional evidence would be needed.

6 Limitations and Next Steps

This study is deliberately narrow. The benchmark uses short, synthetic news-style passages so that the relevant evidence can be controlled with a high degree of accuracy. That design makes it far easier to determine what the model was entitled to infer, but it does not reproduce the messiness of real information environments, in which supplied documents may be long, incomplete, internally inconsistent, written for different purposes, or distributed across various different sources. As a result, the demonstrated performance on bounded passages should not be assumed to describe performance on open-ended research, web searches, or large document collections.

The study also examines two frontier models in one three-way classification setting. The observed pattern may vary across families of models, versions, prompting strategies, languages, domains, or product configurations. Real deployments add additional layers—retrieval systems, system prompts, citation interfaces, document ranking, and user interaction—that can either reduce or amplify the underlying issue that was identified. Accordingly, the benchmark is best interpreted as evidence of a specific failure type that is worth measuring, but not as a universal estimate of how all generative AI models will behave.

A second limitation relates to the idea of "sufficiency" itself. The benchmark assigns each claim to a discrete

category, while real evidence often comes in degrees. For instance, a source may partially support a proposition, support it only under an unstated assumption, or provide evidence that is relevant but not particularly concrete. The benchmark also asks whether the supplied text warrants the claim. This means that it does not independently evaluate whether the source is credible, representative, current, or complete – which are separate questions that are incredibly important in real-world research and policy use.

The youth discussion is similarly constrained. Since there were no young people that participated in the experiment, the study does not measure how students interpret citations, respond to uncertainty, or detect unsupported inference in practice. The educational implications in this brief result from the structure of the model error as opposed to from evidence about youth behavior. Hence, user studies would be needed to determine whether claim-level citations, explicit insufficiency labels, or particular forms of instruction actually improve users' ability to distinguish between source support and model synthesis.

Future work should test how durable the pattern is under more natural conditions. Some useful extensions include longer and multi-document evidence, conflicting sources, domain-specific materials, additional models and languages, and end-to-end retrieval systems. Furthermore, it would be valuable to compare interface designs and prompting strategies that make the boundaries of evidence more explicit. Those studies would help determine when the gap observed here is most consequential and which interventions meaningfully address it.

7 Conclusion

The central finding of this benchmark is not that frontier language models performed poorly. That is simply false; both systems achieved high overall accuracy and were nearly perfect when the supplied evidence directly supported the claim. The more informative result is in which instances performance was weakened: both models were materially less accurate when the evidence did not establish the proposition before them.

When the systems missed those cases, they generally moved toward acceptance. Specifically, roughly one in seven insufficient-evidence claims was classified as true in the data. Consequently, that pattern should influence how source-grounded AI is evaluated. Further, supplying the correct document is valuable, but it does not settle the question of whether the model will use the document carefully. The fact remains that retrieval and evidence use are connected problems, but not identical ones.

For users, including young ones, the practical lesson is certainly familiar but increasingly urgent. Good research requires far more than locating a source related to a claim; it requires determining what the source actually proves, what remains an inference, and where additional evidence is needed. AI can make the first step extraordinarily fast, but it is imperative that evaluation, product design, and education preserve the second.

Authors

Institute for Youth in Policy

Nathaniel Nadler

Felix Lee

Bhavya Thakur

Wanyi Li

Mateo Santalucia

Dyuthi Dhawan

Frontlines Foundation

Anshi Bhatt

Rithika Pendurthi

Ayesha Afghan

Anvi Jha

Grace Kweon

Rithika Pendurthi

Rida Akhtar

Veda Kumaresan