Global Journal of Scientific & Humanities Research (GJSHR)

E-ISSN:3029-9926

المجلة العالمية للبحوث العلمية والإنسانية

Trust and Reliability of Generative AI in Academic Research: An Evaluation of Large Language Models

Motasem AbuDawas1
Hasan Alhasan2
Associate Professor, Department of Computer Science1
Assistant Professor, Department of Computer Science2
Irbid National University, Jordan
motasem_dawas@inu.edu.jo 1
h.alhasan@inu.edu.jo 2

Volume: 1

Issue: 1

Year: 2026

Article No.: 010

E-ISSN: 3029-9926

Received: 11 July 2026

Accepted: 14 September 2026

Published: 30 September 2026

DOI:

https://doi.org/10.68792/010.01.2026

Website:

www.gjshr.cz

Abstract

Generative artificial intelligence (AI), particularly large language models (LLMs), is increasingly being integrated into academic research to support tasks such as scientific explanation, methodological guidance, literature identification, evidence retrieval, citation generation, and interpretation of research findings. Despite their growing use, concerns remain regarding the extent to which researchers can trust and rely on these models for scholarly tasks. This study evaluates the trustworthiness and reliability of generative AI in academic research through a multidimensional assessment of three contemporary LLMs under controlled experimental conditions. The evaluation considers key aspects of research assistance, including knowledge and answer reliability, methodological reliability, citation and evidence reliability, evidence-to-claim alignment, prompt robustness, and response repeatability. Two empirical experiments are conducted. The first evaluates 144 model responses across four scientific domains using benchmark key-point scoring and factual-error detection. The second examines 216 generated citation instances based on publication existence, bibliographic metadata accuracy, DOI accuracy, relevance, claim support, and fabrication status. The findings demonstrate strong overall performance across the evaluated models while identifying specific reliability weaknesses that are not adequately reflected by aggregate accuracy measures alone.

Keywords: Generative AI; Large Language Models; Academic Research; Trustworthiness; AI Reliability; Citation Reliability; Research Assistance; Prompt Robustness

الثقة والموثوقية في الذكاء الاصطناعي التوليدي في البحث الأكاديمي: تقييم النماذج اللغوية الكبيرة

معتصم أبو دواس[1]

حسن الحسن[2]

أستاذ مشارك، قسم علوم الحاسوب [1]

أستاذ مساعد، قسم علوم الحاسوب [2]

جامعة إربد الأهلية، الأردن

البريد الإلكتروني:

[1] **motasem_dawas@inu.edu.jo**

[2] **h.alhasan@inu.edu.jo**

الملخص

تُستخدم نماذج اللغة الكبيرة بصورة متزايدة لمساعدة الباحثين في تفسير المفاهيم العلمية، واقتراح المنهجيات، والوصول إلى الأدلة والمراجع، وتفسير النتائج. إلا أن موثوقية النموذج كمساعد بحثي لا يمكن اختزالها في دقة الإجابة فقط؛ فقد يقدم النموذج إجابة علمية مقبولة مع خطأ واقعي، أو بيانات مرجعية غير دقيقة، أو مرجع صحيح لا يدعم الادعاء بصورة مباشرة، أو يتغير أداؤه عند إعادة صياغة الطلب. تقترح هذه الدراسة إطاراً متعدد الأبعاد وقابلاً للتكرار لتقييم موثوقية نماذج اللغة الكبيرة في مهام المساعدة البحثية. يفصل الإطار بين موثوقية المعرفة والإجابة، وموثوقية الأدلة والمراجع، والموثوقية المنهجية، وموثوقية الانتقال من الأدلة إلى الادعاءات والتفسير، مع اعتبار متانة صياغة السؤال والاتساق خصائص عابرة للأبعاد. تم اختبار قابلية تطبيق الإطار من خلال تجربتين: الأولى تضمنت 144 استجابة في أربعة مجالات علمية، والثانية تضمنت التحقق من 216 حالة استشهاد مرجعي. وتبين النتائج أن ارتفاع الأداء الكلي لا يمنع وجود أنماط أدق من عدم الموثوقية، مما يدعم الحاجة إلى تقييم متعدد الأبعاد بدلاً من الاعتماد على مؤشر واحد.

الكلمات المفتاحية: الثقة، الموثوقية ، الذكاء الاصطناعي التوليدي، البحث الأكاديمي، تقييم النماذج اللغوية الكبيرة

Introduction

Large language models have moved rapidly from general-purpose language interfaces to tools used in academic and scientific workflows. Researchers can use them to obtain explanations of unfamiliar concepts, generate possible research questions, identify statistical or experimental approaches, summarize bodies of literature, draft text, and suggest references. The Transformer architecture introduced by Vaswani et al. (2017) and the scaling behavior demonstrated by later foundation models such as GPT-3 (Brown et al., 2020) established the technical basis for systems capable of performing many tasks through natural-language prompting. Subsequent models have demonstrated strong performance across broad academic and professional benchmarks (OpenAI, 2023; Touvron et al., 2023).

The usefulness of these systems, however, creates a methodological problem for researchers. Fluent language can make model output appear more certain and authoritative than the underlying evidence warrants. Earlier work has emphasized that large language models can reproduce biases and limitations contained in their training data and that users may mistake linguistic fluency for genuine understanding or factual reliability (Bender et al., 2021). Surveys of hallucination in natural-language generation similarly document that generated text can contain unsupported or incorrect information even when it is coherent and grammatically convincing (Ji et al., 2022; Rawte et al., 2023).

This problem is especially important in research assistance because scientific work is not a single task. A researcher may ask an AI system to explain a concept, recommend a method, identify peer-reviewed evidence, connect a citation to a claim, interpret a result, or identify a limitation. Failure in any one of these steps can affect the validity of subsequent work. Therefore, evaluating whether an LLM is simply 'accurate' does not provide enough information about whether it is a trustworthy research assistant.

Citation generation illustrates the problem clearly. A model can supply a publication that exists but provide an incorrect DOI; it can reproduce correct bibliographic metadata but recommend a paper that only weakly supports the requested statement; or it can fabricate a plausible-looking reference. Empirical work has documented fabrication and bibliographic errors in references generated by ChatGPT (Walters & Wilder, 2023), while retrieval-

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

augmented generation has been proposed as one means of improving provenance and factual grounding (Lewis et al., 2020). More recent citation-oriented benchmarks further show that citation generation and retrieval constitute distinct evaluation problems rather than a trivial extension of general question answering (Tilwani et al., 2024).

A second difficulty concerns prompt formulation. Researchers do not ask the same question in a standardized form. Equivalent research needs may be expressed formally, conversationally, indirectly, or with different terminology. Prior work on prompt sensitivity shows that model behavior can change under perturbations or paraphrases even when the underlying task remains similar. Consequently, a useful reliability assessment should not only ask whether a model can succeed under one carefully written benchmark prompt; it should also ask whether performance remains stable when the same need is expressed realistically in another way.

This paper proposes a multidimensional framework for evaluating the reliability of LLMs as research assistants. The framework is intended to be domain-general at the level of reliability dimensions and metrics while allowing discipline-specific benchmark content. It is also designed to be repeatable: a new model, a new version of an existing model, or a discipline-specific research assistant should be testable using the same evaluation logic. The contribution is therefore not a ranking of three particular models. The contribution is a structured way to ask what must be measured before an AI system can be considered reliable for research assistance, together with empirical evidence that several of those dimensions can be operationalized and measured consistently.

2. Research Problem, Objectives, and Questions

The central research problem is the absence of a compact but sufficiently comprehensive evaluation structure for determining how trustworthy an LLM is when it is used to assist scientific research. Existing model benchmarks often measure broad capability, task accuracy, reasoning, hallucination, or retrieval performance separately. Those evaluations are important, but they do not directly answer the practical question faced by a researcher: which parts of a research workflow can be delegated to an AI model, what kinds of failure remain possible, and how should those failures be measured in a repeatable way?

The proposed framework therefore treats research-assistant reliability as a multidimensional construct. Instead of assigning reliability a single global definition, the framework separates the research process into distinct functions and associates each function

E-ISSN:3029-9926 https://doi.org/10.68792/010.01.2026
www.gjshr.cz

with measurable risks. This design allows a model to be strong in one dimension and weaker in another without forcing all behavior into one aggregate score.

The study has seven objectives: to define core dimensions of reliability relevant to AI-assisted research; identify measurable indicators for each dimension; distinguish universal evaluation dimensions from domain-specific benchmark material; construct repeatable protocols that can be applied to different LLMs; empirically test the feasibility of those protocols; examine whether realistic prompt variation affects measured reliability; and determine whether multidimensional analysis reveals weaknesses that aggregate answer accuracy would obscure.

The current work addresses the following research questions: (RQ1) Can LLM research-assistant reliability be decomposed into independently measurable dimensions? (RQ2) To what extent do contemporary LLMs provide correct and complete research-oriented scientific answers? (RQ3) To what extent are references generated by LLMs authentic, bibliographically accurate, relevant, and supportive of the associated claim? (RQ4) How sensitive are these reliability measurements to realistic reformulation of the same underlying research request? (RQ5) Can the same evaluation architecture be applied across scientific disciplines while using discipline-specific benchmark content? (RQ6) Which reliability dimensions remain unvalidated and therefore require additional experiments before a comprehensive research-assistant profile can be produced?

3. Related Work

The modern LLM ecosystem is built on the Transformer architecture, which replaced recurrence with attention-based sequence modeling and enabled efficient scaling to large datasets and model sizes (Vaswani et al., 2017). GPT-3 subsequently demonstrated that sufficiently large autoregressive language models can perform many tasks in zero-shot and few-shot settings through natural-language prompting without task-specific parameter updates (Brown et al., 2020). Other foundation-model families, including LLaMA, reinforced the observation that broad language capabilities can emerge from large-scale pretraining (Touvron et al., 2023). These developments help explain why LLMs are attractive as research assistants: a single conversational interface can support heterogeneous tasks without requiring a separate specialized application for each one.

Capability, however, is not equivalent to reliability. The GPT-4 technical report itself notes that the model is not fully reliable and can hallucinate, particularly in contexts in which factual

E-ISSN:3029-9926 https://doi.org/10.68792/010.01.2026
www.gjshr.cz

correctness matters (OpenAI, 2023). Bender et al. (2021) emphasize broader risks of large language models, including limitations associated with training data and the tendency to over-attribute understanding to systems that generate plausible language. These concerns are directly relevant to research use because scientific work depends on traceability, evidence, methodological appropriateness, and justified inference rather than fluent text alone.

Hallucination research provides an important foundation for the proposed framework. Ji et al. (2022) survey hallucination across natural-language generation tasks and distinguish different forms of generated content that are unsupported or inconsistent with source information. Rawte et al. (2023) further emphasize that hallucination is not a single phenomenon and propose more fine-grained categorization. These studies motivate the separation of factual-error detection from overall answer completeness in the present framework. A response may contain most of the expected scientific content and still include one additional erroneous claim. Treating that response as simply correct or incorrect would lose useful information about the type of failure.

Reference generation has become a particularly visible manifestation of hallucination in academic use. Walters and Wilder (2023) systematically examined bibliographic citations generated by ChatGPT and reported fabrication and metadata errors, demonstrating that plausible citation formatting cannot be treated as evidence of bibliographic validity. Citation reliability also involves more than existence. A real publication may be inaccurately described or may not support the claim for which it is recommended. This distinction underlies the multi-component reference verification matrix used in Experiment 2.

Retrieval-augmented generation provides another relevant line of work. Lewis et al. (2020) combine parametric generation with non-parametric retrieval and show improvements in knowledge-intensive tasks and factual specificity. This is important conceptually because it demonstrates that reliability can depend on the interaction between a model and external evidence sources. The current study intentionally evaluates model-generated references under a controlled protocol without web search during generation, followed by independent external verification. This separation allows the experiment to assess the model's generated research assistance rather than conflating it with the performance of an integrated search engine.

Recent work has also begun to develop citation-focused benchmarks. Tilwani et al. (2024) introduce REASONS, a benchmark for retrieval and automated citation generation spanning multiple scientific domains, and evaluate both proprietary and public LLMs. Such work

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

supports the need to treat citation behavior as a dedicated evaluation problem. The present framework complements this direction by decomposing citation reliability into reference existence, metadata correctness, identifier correctness, topical relevance, claim support, and fabrication rather than relying on a single citation-generation metric.

Prompt sensitivity and robustness form another foundation for this study. Natural-language instructions are inherently variable, and models can respond differently to lexically or structurally altered prompts. Studies of prompt perturbation have shown that even semantically similar formulations can change model performance. From the perspective of research assistance, this matters because a benchmark that evaluates only one carefully engineered prompt may overestimate real-world reliability. The present study therefore treats prompt robustness as a cross-cutting property rather than as a separate research task. Each core dimension can be tested under multiple equivalent formulations.

The gap addressed by this paper lies at the intersection of these literatures. Work on hallucination, citation generation, retrieval, and prompt robustness typically evaluates a particular behavior. What is still needed for practical research use is an integrated architecture that connects those behaviors to the stages of scientific work and specifies how they can be measured repeatedly. The proposed framework is designed to provide that architecture while remaining extensible as new forms of AI-assisted research emerge.

4. Conceptual Model of Research-Assistant Reliability

The framework begins from the observation that researchers interact with AI across a sequence of epistemic and methodological activities. A model may first be asked to explain or retrieve scientific knowledge, then to identify evidence, then to recommend a method, and finally to interpret results or formulate claims. The reliability requirements differ at each stage. The framework therefore defines core dimensions corresponding to these functions and cross-cutting properties that can be assessed across all dimensions.

As illustrated in Figure1, the proposed framework organizes research-assistant reliability into four core dimensions—knowledge and answer reliability, evidence and citation reliability, methodological reliability, and evidence-to-claim and interpretation reliability. These dimensions are complemented by cross-cutting properties, including prompt robustness, repeatability and consistency, and uncertainty awareness, while domain-specific benchmark modules allow the same framework to be applied across different scientific disciplines.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

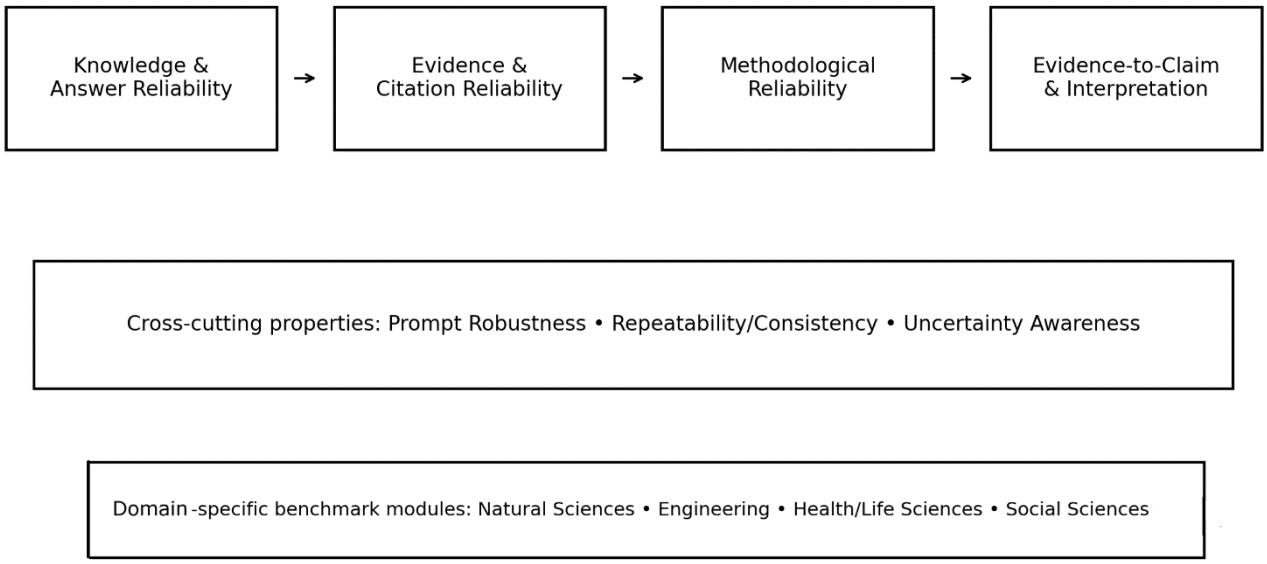

Figure 1. Proposed multidimensional framework for evaluating LLMs as research assistants.

4.1 Knowledge and Answer Reliability

Knowledge and Answer Reliability asks whether the model provides scientifically correct and sufficiently complete responses to research-oriented questions. This is the most familiar evaluation dimension, but the framework deliberately decomposes it into at least two signals: benchmark coverage and factual-error incidence. Benchmark coverage measures whether the answer contains the predefined scientific concepts required by the task. Factual-error incidence records objectively incorrect claims even when the rest of the answer is strong.

This separation avoids an important measurement problem. Suppose a model correctly explains three required mechanisms and then adds one unsupported statement. A pure completeness score may remain high, while a binary correct/incorrect score may over-penalize the otherwise useful answer. Recording both dimensions creates a more informative reliability profile.

4.2 Evidence and Citation Reliability

Evidence and Citation Reliability asks whether the sources supplied by a model can legitimately be used by a researcher. The framework treats citation reliability as a chain. First, the publication must exist. Second, the bibliographic representation must be accurate. Third, identifiers such as the DOI must correspond to the publication. Fourth, the source must be relevant to the topic. Finally, the source must actually support the claim for which the researcher requested evidence.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

This chain is important because a reference can fail at different levels. Fabrication is the most severe failure, but it is not the only one. A model may provide a real paper with a slightly corrupted title or DOI, or it may supply a real and accurately described paper that is merely adjacent to the requested claim. The proposed reference matrix therefore records each component independently.

4.3 Methodological Reliability

Methodological Reliability concerns whether the model recommends an appropriate research design, analytical method, or methodological procedure for the scenario presented. A research assistant that knows domain facts but repeatedly recommends invalid methods can still threaten research quality. Candidate metrics include method-selection accuracy, quality of justification, recognition of assumptions, identification of confounders or nuisance variables, and invalid-method recommendation rate.

This dimension is the highest-priority target for the next empirical experiment. Its inclusion in the current paper is conceptual and prospective; no final Experiment 3 results are reported here. This distinction is maintained to avoid overstating the present evidence.

4.4 Evidence-to-Claim and Interpretation Reliability

Evidence-to-Claim Reliability addresses whether a model's conclusions remain within the limits of available evidence. Researchers routinely distinguish association from causation, statistical significance from substantive importance, and absence of evidence from evidence of absence. A model that overstates conclusions may be misleading even if the data and cited literature are themselves correct.

Candidate indicators include supported-claim rate, contradiction rate, causal-overstatement rate, overclaim rate, uncertainty acknowledgement, limitation detection, and recognition of alternative explanations. Research Interpretation Reliability may ultimately be treated as a subdimension of this category or maintained separately, depending on results from future experiments.

4.5 Cross-Cutting Properties

Prompt Robustness and Repeatability/Consistency are treated as cross-cutting properties because they can be measured within every core dimension. Prompt robustness asks whether reliability changes when an equivalent research need is expressed using substantially different wording or conversational framing. Repeatability asks whether the model behaves consistently

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

across repeated runs. Neither property proves correctness by itself: a model can be consistently wrong. Their value is to show whether reliability is stable or fragile under realistic interaction condition

4.6 Framework Matrix v1

Table 1 presents the first operational version of the proposed framework by linking each reliability dimension to its research risk, representative metrics, cross-cutting tests, and validation status. As shown in Table 1, Knowledge and Answer Reliability and Evidence and Citation Reliability have been empirically evaluated through Experiments 1 and 2 (see sections 7 and 8), while the remaining dimensions are planned extensions. This matrix provides a repeatable structure for evaluating different AI models while allowing discipline-specific benchmark content.

Dimension	Research risk	Representative metrics	Cross-cutting tests	Status
Knowledge & Answer Reliability	Incorrect or incomplete scientific information	Correctness; completeness; normalized benchmark score; factual-error rate	Prompt robustness; repeatability	Empirically tested (Exp. 1)
Evidence & Citation Reliability	Invalid, inaccurate, or weakly supporting evidence	Existence; metadata accuracy; DOI accuracy; relevance; claim support; fabrication rate	Prompt robustness; reference stability	Empirically tested (Exp. 2)

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

Methodological Reliability	Invalid design, analysis, or inference procedure	Method-selection accuracy; justification; assumptions; confounder awareness; invalid-method rate	Prompt robustness; repeatability	Planned Exp. 3
Evidence-to-Claim Reliability	Conclusions exceed evidence	Supported-claim rate; contradiction; causal overstatement; uncertainty calibration	Prompt robustness; repeatability	Planned
Interpretation Reliability	Misinterpretation of results and limitations	Interpretation accuracy; limitation detection; alternative explanations	Prompt robustness; repeatability	Candidate / may merge with prior dimension

Table 1. Framework Matrix v1: dimensions, risks, metrics, and validation status.

5. Measurement Model and Metrics

A central requirement of the framework is repeatability. Each metric must be defined sufficiently clearly that another evaluator can apply the protocol to a different model without redesigning the study. The present draft therefore uses simple transparent scoring rules rather than opaque composite indices.

E-ISSN:3029-9926

https://doi.org/10.68792/010.01.2026

www.gjshr.cz

5.1 Answer Benchmark Score

For Experiment 1, each research task was associated with a predefined set of key points. Each key point received 0 when absent or incorrect, 1 when partially addressed, and 2 when correctly and sufficiently addressed. If a response contains m scored key points with scores k1 ... km, its normalized answer score is defined as NAS = (Σ ki) / (2m). NAS therefore ranges from 0 to 1. A value near 1 indicates that the response covers almost all benchmark requirements, but it does not guarantee the absence of an additional erroneous statement.

5.2 Factual Error Rate

Factual errors are recorded separately using a binary indicator for each response. The model-level Factual Error Rate is FER = E/N, where E is the number of responses containing at least one objectively incorrect factual claim and N is the total number of evaluated responses. Keeping FER separate from NAS avoids double-counting and preserves interpretability: a model can have high benchmark coverage while still producing occasional factual errors.

5.3 Citation Reliability Score

Experiment 2 evaluates each citation instance using five binary bibliographic components—authors, year, title, venue, and DOI/identifier—plus relevance and claim-support scores ranging from 0 to 2. The maximum reference score is therefore 9 points. Publication existence and fabrication are recorded explicitly as verification flags rather than hidden inside the metadata subtotal. The normalized citation reliability score is the mean reference score divided by 9. This structure makes it possible to report both an overall profile and the specific failure mode.

5.4 Prompt Robustness

For paired prompt versions P1 and P2, prompt robustness can be summarized by the absolute difference in mean score: PR-gap = |mean(P1) - mean(P2)|. Lower values indicate smaller performance degradation under realistic reformulation. The framework does not treat this gap as a universal threshold because acceptable variation may depend on task risk and domain. Instead, the metric is intended for comparisons across models, versions, or deployment settings.

5.5 Future Reliability Profile

The long-term objective is a reliability profile rather than a single universal score. A future model could therefore be reported as, for example, Knowledge Reliability = 0.96, Citation

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

Reliability = 0.91, Methodological Reliability = 0.84, Evidence-to-Claim Reliability = 0.80, with separate robustness and repeatability indicators. Such a profile is more actionable than a single composite score because it shows which research activities require additional human verification.

6. General Experimental Protocol

The empirical component uses controlled prompting and delayed scoring. Benchmark prompts are predetermined before data collection. Responses are recorded as raw outputs without editing. Scoring is then performed after the collection stage using the locked benchmark and rubric. This ordering reduces the risk that the evaluator changes the benchmark in response to particular model outputs.

Each prompt version is submitted in a fresh interaction to reduce context contamination. Where the purpose is to evaluate parametric model behavior rather than integrated information retrieval, web-search functionality is disabled during generation. External verification is performed only after generation, particularly for citation validation. This separation is essential because a model-plus-search system represents a different experimental condition from a model operating from its internal knowledge and generation process.

Three contemporary LLM services are evaluated. The purpose is not to establish a permanent leaderboard because model versions change rapidly. Instead, the three systems serve as validation cases demonstrating that the same framework and benchmark can be applied across vendors. The protocol should be rerunnable when models are updated, creating a longitudinal reliability record.

7. Experiment 1: Knowledge and Answer Reliability

7.1 Design

Experiment 1 contains 24 research-oriented tasks distributed across four domains: Natural Sciences, Engineering and Technology, Health and Life Sciences, and Social Sciences and Humanities. The task set includes scientific retrieval, scientific explanation, and methodological reasoning. This distribution is intended to test whether the scoring logic remains usable beyond a single discipline.

Each underlying task was represented by two prompt formulations, P1 and P2, and evaluated on three models. The resulting dataset contains $24 \times 2 \times 3 = 144$ responses. Each task has a locked set of key points used as the benchmark for scoring.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

7.2 Scoring

Key points were scored 0, 1, or 2, and response scores were normalized to the 0–1 interval. Factual errors were recorded independently. The evaluator did not award or deduct points based purely on style, verbosity, or phrasing unless those features affected whether a required scientific point was present and correct.

7.3 Results

The descriptive results of Experiment 1 are summarized in Table 2, which reports the mean normalized benchmark score and factual-error count for each evaluated model. These measures are presented separately because benchmark coverage and factual correctness capture different aspects of answer reliability.

Model	Mean normalized score	Factual-error count	Interpretive note
Claude Sonnet 5	0.992	0	Highest descriptive benchmark score; strong ceiling effect.
GPT-5.6 Sol	0.984	0	Very high benchmark coverage; no factual-error instances under rubric.
Gemini 3.1 Pro	0.974	2	High benchmark coverage with two separately recorded factual-error instances.

Table 2. Experiment 1 descriptive results.

All three models achieved very high benchmark scores. The principal feature of the result is therefore not the small numerical ranking among models but the presence of a ceiling effect. Straightforward research-oriented scientific questions do not strongly discriminate among current high-capability models under this benchmark. This is itself informative for framework

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

design: answer correctness is necessary but insufficient as the sole dimension of research-assistant reliability.

The separation of benchmark score and factual-error incidence is also empirically useful. Gemini obtained a mean normalized score of approximately 0.974 while two responses were marked as containing factual errors. These facts are not contradictory. The benchmark score measures coverage of expected scientific points, whereas the factual-error indicator records whether an incorrect claim appeared. A response can cover almost all expected content and still introduce an additional erroneous statement.

The ceiling effect motivates the addition of more demanding reliability dimensions rather than arbitrary difficulty inflation within a single benchmark. Citation verification, methodological reasoning, evidence-to-claim discipline, and repeated-response stability may reveal differences that are not visible in short scientific-answer tasks.

8. Experiment 2: Evidence and Citation Reliability

8.1 Design

Experiment 2 evaluates a different research-assistance function: supplying peer-reviewed literature in support of a scientific claim. Twelve benchmark tasks were constructed across several disciplines. Each task had two prompt versions, producing 24 prompt conditions. Each of the three models was asked to provide exactly three peer-reviewed references, resulting in 72 responses and 216 citation instances.

P1 used a formal benchmark-oriented formulation. P2 preserved the same underlying research claim and evidence requirement but used substantially different, more natural researcher language. This design was adopted because minor paraphrasing would test only surface wording sensitivity. Real researchers may ask the same question in very different ways; the experiment therefore aims to test robustness under realistic reformulation.

8.2 Reference Verification Matrix

Every citation instance was independently checked after generation. The verification matrix recorded publication existence, author accuracy, publication year, title, venue, DOI or identifier, topical relevance, degree of support for the requested claim, and fabrication status. The key methodological point is that these variables are not interchangeable. A citation can be genuine while containing a corrupted DOI, and a bibliographically perfect citation can still be only indirectly relevant to the requested claim.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

8.3 Results

The descriptive results of Experiment 2 are summarized in Table 3, which compares the three evaluated models in terms of overall citation reliability, bibliographic metadata accuracy, DOI accuracy, claim-support strength, fabrication incidence, and robustness to prompt reformulation. These measures are reported separately to distinguish different forms of citation reliability rather than reducing performance to a single indicator.

Model	Citation reliability	Metadata accuracy	DOI accuracy	Mean claim support /2	Fabricated	P1-P2 gap
GPT-5.6 Sol	99.54%	100.00%	100.00%	1.958	0	0.028/9
Claude Sonnet 5	98.77%	99.44%	97.22%	1.917	0	0.111/9
Gemini 3.1 Pro	98.46%	98.61%	94.44%	1.931	0	0.056/9

Table 3. Experiment 2 descriptive citation-reliability results.

No completely fabricated reference was identified in the evaluated sample. This finding is encouraging, but it does not imply perfect citation reliability. The multidimensional verification matrix identified narrower failure modes in metadata and DOI accuracy and in the strength of claim support. The result therefore illustrates exactly why fabrication rate alone is an inadequate measure of scholarly citation reliability.

GPT-5.6 Sol showed the strongest descriptive profile in this sample, with perfect metadata and DOI accuracy under the adopted verification procedure. Claude and Gemini also performed strongly but exhibited a small number of metadata or identifier errors. Mean claim-support values indicate that some references, although real and topically relevant, supported the requested statement indirectly rather than directly.

Prompt reformulation produced only small changes in mean reference score. The P1-P2 mean-score gaps were approximately 0.028 points out of 9 for GPT-5.6 Sol, 0.056 for Gemini, and 0.111 for Claude. These values suggest substantial robustness for the tested citation tasks. Nevertheless, aggregate score stability does not imply that the exact set of references remains

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

identical. A future version of the framework should therefore distinguish score robustness from reference-selection stability.

9. Cross-Experiment Interpretation

The two experiments examine different stages of research assistance and together provide stronger evidence for the framework than either experiment alone. Experiment 1 shows that current models can achieve near-ceiling performance on concise research-oriented scientific questions. Experiment 2 shows that high general capability does not make citation verification unnecessary and that citation reliability itself contains several separable dimensions.

A key result is that reliability failures can be narrow. In Experiment 1, a model can maintain a very high normalized benchmark score while still producing isolated factual errors. In Experiment 2, a model can avoid fabrication entirely while still producing a small number of bibliographic or DOI inaccuracies or selecting evidence that is only indirectly supportive. A single aggregate accuracy measure would obscure these distinctions.

This observation supports the framework's central premise: research-assistant reliability should be represented as a profile. The relevant question is not merely whether a model is 'good' or 'bad,' but where in the research-assistance process it is strong, where it is fragile, and which outputs require independent human verification.

10. Feasibility of a Repeatable Framework

The present study was explicitly designed to test feasibility as well as model performance. A framework is useful only if another evaluator can reproduce its logic without relying on subjective impressions. The first two experiments provide evidence that this is possible. Benchmark answers can be represented as scored key points; factual errors can be recorded independently; references can be decomposed into objective bibliographic fields and evidence-support judgments; and prompt robustness can be quantified through paired formulations.

The domain-general/domain-specific distinction also appears feasible. The same scoring architecture can be applied across natural sciences, engineering, health and life sciences, and social sciences/humanities, while the benchmark questions and ground truth are adapted to each domain. This is preferable to designing an entirely different framework for every discipline because it preserves comparability while respecting disciplinary differences.

Repeatability over time is another intended use. Because commercial models are updated frequently, a static benchmark result rapidly becomes outdated. The framework is designed so

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

that the same benchmark and rubric can be rerun on a later model version. A laboratory, university, publisher, or research group could therefore maintain a longitudinal reliability profile rather than relying on vendor claims or anecdotal impressions.

The framework is not intended to eliminate expert judgment. Some components, such as whether a paper directly supports a claim, require informed assessment. The goal is instead to structure that judgment, define the categories in advance, and separate them from more objective fields such as DOI correctness. This makes evaluation more transparent and auditable.

11. Planned Experiment 3: Methodological Reliability (Future Work)

In future work, the next empirical stage will target Methodological Reliability. Rather than asking for explanations or references, the model will receive realistic research scenarios and will be required to select and justify an appropriate research design, statistical procedure, or methodological decision. Scenarios should vary in domain and should include cases in which assumptions, confounders, repeated measures, blocking, dependence, multiple comparisons, or validation leakage matter.

A preliminary metric set includes Method Selection Accuracy, Justification Quality, Assumption Awareness, Confounder/Nuisance Recognition, and Invalid-Method Rate. Each scenario will have a locked expert benchmark specifying the acceptable method and the critical reasoning points. As in the first two experiments, at least two substantially different prompt formulations should be used to test robustness.

Experiment 3 is important because methodological errors can be more consequential than a minor factual omission. A researcher who follows an inappropriate statistical or experimental design may generate invalid results even if the background explanation and citations are accurate. Therefore, methodological reliability is a necessary component of any framework that claims to evaluate AI assistance across the research workflow.

12. Threats to Validity and Limitations

First, the empirical validation currently covers three LLM services. The study is not intended to generalize a permanent ranking to all models. The models are demonstration cases for the measurement framework, and future work should repeat the evaluation on additional systems and model versions.

Second, the benchmark tasks are controlled and relatively concise. They cannot represent the full complexity of conducting a real scientific project. This limitation is partly intentional

E-ISSN:3029-9926 https://doi.org/10.68792/010.01.2026
www.gjshr.cz

because repeatable scoring requires clear ground truth, but future extensions should include longer multi-step research tasks.

Third, Experiment 1 shows a ceiling effect. The benchmark therefore has limited power for separating very capable models on straightforward scientific questions. The framework responds to this limitation by adding orthogonal dimensions rather than relying only on harder question answering.

Fourth, citation verification contains both objective and judgment-based components. Existence and DOI correctness are relatively objective; claim support may depend on disciplinary expertise. Future studies should use multiple independent raters and report inter-rater agreement for relevance, support, methodological justification, and interpretation dimensions.

Fifth, model behavior is stochastic and can change with version updates, temperature settings, system prompts, and platform-specific defaults. The current study controls interaction conditions as far as practical, but future reports should document the model identifier, access date, sampling settings when available, and whether search or external tools were enabled.

Sixth, no web search was used during the generation condition in the citation experiment. This isolates parametric research-assistance behavior, but it does not represent retrieval-enabled products. The framework should therefore distinguish 'model-only' and 'model-plus-retrieval' reliability profiles rather than treating them as interchangeable.

Finally, the current framework is intentionally extensible. Methodological Reliability and Evidence-to-Claim Reliability have not yet received the same empirical validation as the first two dimensions. The present results support feasibility, not completion of a universal standard.

13. Practical Implications

For individual researchers, the framework offers a way to decide what level of verification is needed for different AI-assisted tasks. A model with high knowledge reliability but weaker citation reliability might be acceptable for concept explanation while requiring independent bibliographic checking before references are used in a manuscript. A model with strong citation performance but uncertain methodological reliability should not be relied upon to design an experiment without expert review.

For universities and research institutions, the framework could support procurement, training, and responsible-use policies. Instead of adopting a model based on general reputation,

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

an institution could test candidate systems using discipline-specific benchmark modules and report a research-assistant reliability profile. The same protocol could be rerun after a major model update.

For model developers, a multidimensional profile can identify targeted weaknesses. A drop in DOI accuracy, prompt robustness, or assumption awareness may be easier to diagnose and improve than a vague global reliability score. This also encourages evaluation procedures that reflect actual research workflows rather than generic language tasks.

For publishers and conference organizers, the framework may eventually support disclosure guidelines for AI-assisted research. Rather than treating all AI use as equivalent, disclosure could distinguish tasks such as language editing, literature discovery, methodological guidance, data interpretation, and claim generation, each associated with different verification requirements.

14. Discussion

The empirical findings are notable precisely because they are not dramatic. All three models performed very strongly in both completed experiments. A conventional benchmark paper might regard this as a limitation because there is little separation between systems. For the present research question, however, the ceiling effect strengthens the argument for multidimensional evaluation. When general answer accuracy becomes high, narrower reliability failures become proportionally more important.

Experiment 1 demonstrates that answer quality and factual errors should not be collapsed. Experiment 2 demonstrates that citation reliability cannot be represented by fabrication alone. These two observations generalize beyond the specific models tested. They show that the operational unit of reliability must be matched to the research risk. If the risk is using an incorrect claim, factual-error detection matters. If the risk is citing non-existent literature, existence matters. If the risk is citing genuine literature that does not support the statement, claim-support evaluation matters.

This risk-oriented interpretation is central to the proposed framework. The framework is not merely a taxonomy of model behaviors. Each dimension is tied to a failure that can damage scientific work. This linkage makes the metrics interpretable and helps determine where human verification is necessary.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

The domain-general structure also addresses an important practical problem. Scientific disciplines differ in methods and evidence standards, but many reliability risks are common: incorrect facts, inaccurate references, overclaiming, and unstable responses are problematic everywhere. The framework therefore keeps the dimensions stable while allowing discipline-specific benchmarks to define what counts as correct or methodologically appropriate.

The next stage of validation should focus on methodological decision-making because it represents a qualitatively different form of assistance. If the same framework logic can successfully operationalize methodological reliability, the evidence for a reusable research-assistant evaluation architecture will be considerably stronger.

15. Conclusion

This study proposes a multidimensional framework for evaluating large language models as research assistants. The framework is based on the premise that scientific reliability cannot be represented adequately by one accuracy value. Reliable research assistance involves multiple stages: producing correct scientific information, identifying trustworthy evidence, selecting appropriate methodology, and forming conclusions that remain within the limits of the evidence.

Two experiments provide initial empirical validation. Experiment 1 operationalizes Knowledge and Answer Reliability using benchmark key-point coverage and separate factual-error detection across 144 responses. Experiment 2 operationalizes Evidence and Citation Reliability by verifying 216 citation instances for existence, metadata, DOI accuracy, relevance, claim support, and fabrication. Both experiments also incorporate alternative prompt formulations to examine robustness.

The results show near-ceiling aggregate performance but also reveal failure types that would be invisible under a single measure. High benchmark coverage can coexist with isolated factual errors; absence of fabricated citations can coexist with bibliographic inaccuracies or indirect evidential support. These findings demonstrate the practical value of decomposing reliability into interpretable components.

The proposed framework should therefore be viewed as an extensible evaluation architecture rather than a completed universal standard. Its immediate contribution is to define a repeatable measurement logic and demonstrate the feasibility of applying that logic across multiple models and scientific domains. Ongoing work will extend the framework through

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

methodological-reliability and evidence-to-claim experiments, repeated-run consistency analysis, and multi-rater validation. The long-term objective is a research-assistant reliability profile that helps researchers identify which AI-supported tasks can be used with confidence and which require stronger human verification.

E-ISSN:3029-9926
https://doi.org/10.68792/010.01.2026
www.gjshr.cz

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2022). Survey of hallucination in natural language generation. arXiv:2202.03629. https://doi.org/10.48550/arXiv.2202.03629

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33.

OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774.

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A., & Das, A. (2023). The troubling emergence of hallucination in large language models: An extensive definition, quantification, and prescriptive remediations. Proceedings of EMNLP 2023, 2541–2573. https://doi.org/10.18653/v1/2023.emnlp-main.155

Tilwani, D., Saxena, Y., Mohammadi, A., Raff, E., Sheth, A., Parthasarathy, S., & Gaur, M. (2024). REASONS: A benchmark for retrieval and automated citations of scientific sentences using public and proprietary LLMs. arXiv:2405.02228.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, 14045. https://doi.org/10.1038/s41598-023-41032-5