Proceedings of Machine Learning Research ***318*** *(2026)*
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25—29 May 2026, Vancouver, BC

Citation Constraints and Reference Hallucinations in Large Language Models

Kimberly Davis[†,*] and Qusay H. Mahmoud[†]
[†]Department of Electrical, Computer and Software Engineering
Ontario Tech University, Oshawa, Ontario L1G 0C5, Canada

Abstract

This paper investigates reference hallucinations in large language models (LLMs) under different prompting constraints. Thirty-six academic-style documents were generated across four systems: Gemini 3, ChatGPT 5.1, ChatGPT 4o, and Microsoft 365 Copilot, and evaluated using an automated citation verification method that cross-checks references against Crossref, OpenAlex, and arXiv. The results show that stricter citation requirements are associated with higher rates of invalid or inconsistent references, whereas unconstrained prompts more frequently produce unsupported conceptual claims rather than fabricated citations. These findings indicate that hallucination behaviour depends on task structure rather than simply topic difficulty, highlighting the importance of prompt design and verification when LLMs are used for research-style writing and literature assistance.

Keywords: Large Language Models, hallucination, reference accuracy, prompt design, citation verification.

1. Introduction

Large Language Models (LLMs) have become widely adopted across education, research, industry, and creative practices. Their ability to produce coherent, well-structured text often makes their outputs feel trustworthy. However, these systems can also generate confident but incorrect or fabricated information, commonly referred to as hallucinations [1, 2].

Hallucinations are typically seen as failures of reliability because they can introduce false facts, misleading reasoning, or made-up citations [3, 4]. However, emerging research is beginning to suggest that hallucinations might be more complex than just errors. Prior work has also noted that hallucinations can produce novel or unexpected outputs, though such effects are difficult to evaluate objectively and may vary by application context. [5, 6]. This dual nature creates a tension: while hallucinations can be harmful in contexts that demand accuracy, they might be beneficial in more generative or exploratory scenarios.

As LLMs increasingly shape academic writing, problem-solving, and knowledge creation, understanding their behaviour with respect to hallucinations is crucial. Rather than seeing hallucinations solely in a negative light, researchers now advocate for a more nuanced perspective that recognizes when hallucinations pose risks and when they could foster insight or innovation [7].

This paper examines how hallucination behaviour changes under different prompting constraints, with particular focus on reference integrity in long-form academic-style generation. Building on prior theoretical, taxonomic, and survey-based work, this study presents an empirical comparison of hallucination behaviour across different models, topics, and interaction modes, with particular attention to reference integrity [1, 2, 8]. Rather than viewing hallucinations solely as failures of reliability, this work also explores how they may contribute to creative synthesis when models operate beyond well-established knowledge areas. To this end, this paper's primary contribution is demonstrating that citation requirements influence the form of hallucinations rather than only their frequency.

[*]kimberly.davis@ontariotechu.net

The rest of this paper is organized as follows. Section 2 presents the background and related work, defines hallucinations, discusses their underlying causes, and outlines their advantages and disadvantages. Section 3 describes the experimental design and methodology. Section 4 presents the results and discusses the findings. Section 5 concludes the paper and outlines directions for future work.

2. Background and Related Work

2.1. What is LLM Hallucination?

Hallucinations in Large Language Models refer to outputs that are incorrect, fabricated, or logically inconsistent, often presented with unwarranted confidence [1, 2]. Prior work shows that hallucinations are not a single failure mode but a collection of related behaviours, including factual, reference, reasoning, and contextual errors [1, 7]. Factual hallucinations involve fabricated information such as non-existent studies or incorrect explanations [3, 9], while reference hallucinations produce citations that appear legitimate but do not correspond to real publications [4, 10]. Reasoning hallucinations arise when models generate arguments that appear coherent but contain invalid steps or unsupported assumptions [2, 11], and contextual hallucinations occur when models infer details not grounded in the input [7, 8].

Rather than being isolated issues, these forms of hallucination often interact in practice. For example, fabricated references may support incorrect claims, or flawed reasoning may be reinforced by seemingly credible citations. This interplay is particularly relevant in academic-style generation, where both conceptual correctness and reference integrity contribute to perceived reliability.

2.2. Causes of Hallucinations in LLMs

Hallucinations arise from a combination of data limitations, model architecture, and inference processes [1, 2]. Training data may contain incomplete, outdated, or conflicting information, and when models encounter knowledge gaps, they often generate plausible responses rather than express uncertainty [7, 8]. At a structural level, LLMs generate text by predicting likely token sequences, which can favor fluency and coherence over factual accuracy [2].

In addition, alignment and instruction-following objectives can reinforce this behaviour by encouraging models to produce complete and confident responses, even when the underlying knowledge is uncertain. During inference, prompt design further shapes outcomes: vague prompts may lead to unsupported elaboration, while highly constrained prompts can push models to satisfy formatting or content requirements even when reliable information is unavailable [10, 11]. These interacting factors make hallucinations not simply random errors, but context-dependent behaviours influenced by both training and task design.

2.3. Risks of Hallucinations

Hallucinations pose significant risks when outputs are interpreted as reliable. Models may generate incorrect facts, fabricated references, or misleading explanations that appear credible [3, 4]. These errors can distort reasoning and introduce flawed assumptions, often presented with high confidence [2, 7]. In high-stakes domains, such inaccuracies can lead to harmful decisions if not carefully verified [1, 11].

In academic contexts, reference hallucinations are particularly problematic because they can undermine the credibility of an entire document. Even when the surrounding narrative appears coherent, invalid or unverifiable citations weaken trust and make it difficult to distinguish supported claims from fabricated ones. As LLMs are increasingly used for research assistance, ensuring the reliability of both content and references becomes a central concern.

2.4. Generative Effects of Hallucinations

Despite these risks, hallucinations may also contribute to creative and exploratory tasks. Prior work suggests that LLMs can generate novel ideas, connections, and hypotheses when operating beyond well-established knowledge [5, 6]. In such contexts, hallucinated content can support brainstorming and early-stage exploration, where novelty is often valued over strict correctness [5, 7].

However, this potential benefit is highly context-dependent. The same generative behaviour that produces useful speculation in exploratory settings can introduce misleading or unsupported claims in accuracy-critical applications. This tension highlights the need to distinguish between harmful inaccuracies and potentially useful speculative outputs, and to consider how task design influences which form emerges.

3. Methodology

To better understand how hallucinations behave in practice, especially in academic-style writing, a controlled experiment was conducted using the Large Language Model's Deep Research feature. This mode was selected because it generates long, structured outputs that resemble full research papers with references, making it ideal for examining hallucinations in an academic context. The use of this feature also reflects a realistic scenario, as LLMs are increasingly used for research assistance and literature-style writing. Three research paper prompts were designed for this study:

- **Prompt A:** a well-established software engineering topic.
- **Prompt B:** a niche emerging topic.
- **Prompt C:** the same niche topic with explicit reference-generation guidance.

These conditions were selected to examine the effects of topic familiarity and prompt constraints on hallucination behaviour. The decision was made not to tell the model how many references to include for Prompt A and B. The idea was to see what the model would naturally do, and whether the topic itself influenced how many references it provided and how accurate they were. Prompt C included explicit guidelines for the references to test what the model would do when given stricter instructions. Table 1. Shows the full prompts used.

Prompt A - Well-known Topic	**Prompt B - Niche Topic**	**Prompt C - Niche Topic with Explicit Reference Guidelines**
Write a full research paper intended for submission to a top-tier software engineering conference. The topic of the paper should be **Code Review Practices and Their Impact on Software Quality**. The paper should follow a standard academic structure with sections such as Abstract, Introduction, Background and Related Work, Method/Approach, Analysis or Experimentation, Results, Discussion, and Conclusion. You may choose any research questions, methods, examples, or frameworks you find appropriate. Include in-text citations and references in IEEE format.	Write a full research paper intended for submission to a top-tier software engineering conference. The topic of the paper should be **Adaptive Bug Prediction Using Multi-Agent Large Language Models**. The paper should follow a standard academic structure with sections such as Abstract, Introduction, Background and Related Work, Method/Approach, Analysis or Experimentation, Results, Discussion, and Conclusion. You may choose any research questions, methods, examples, or frameworks you find appropriate. Include in-text citations and references in IEEE format.	Write a full research paper intended for submission to a top-tier software engineering conference. The topic of the paper should be **Adaptive Bug Prediction Using Multi-Agent Large Language Models**. The paper should follow a standard academic structure with sections such as Abstract, Introduction, Background and Related Work, Method/Approach, Analysis or Experimentation, Results, Discussion, and Conclusion. You may choose any research questions, methods, examples, or frameworks you find appropriate. **Include at least 20 references, and all of them must be peer-reviewed sources. Each reference should have corresponding in-text citations. Format all references according to IEEE style, which requires the following details for each entry: the author(s), the year of publication, the title of the work, and the DOI.**

Table 1. Full deep research prompts used in the experiment.

The prompts were tested using the Deep Research feature on four LLM systems:

- Gemini 3.
- ChatGPT 5.1.
- ChatGPT 4o.
- Microsoft 365 Copilot.

These four models were chosen to compare how hallucinations show up across different platforms. ChatGPT 4o was included as a slightly older model to explore whether hallucination behaviour changes between model versions, while the other models reflect newer systems commonly used for research and academic-style writing. This mix helps determine whether the observed patterns are specific to certain models or are consistent across tools. All experiments were conducted between November 2025 - February 2026 using publicly available web interfaces, and outputs were archived immediately after generation, prior to verification. Exact model versions are determined by the provider platforms and may change over time; results should therefore be interpreted as a snapshot of deployed systems during the evaluation period. At the time of writing, ChatGPT 4o is no longer available in ChatGPT.

Each prompt was executed three times per model in independent sessions to account for generation variability while maintaining a controlled comparison across systems and to avoid reinforcement effects from repeated sampling. All generations used each platform's default Deep Research configuration, as sampling parameters such as temperature and top-p were not user-configurable through the official interfaces. This resulted in a total of thirty-six research-style outputs.

In addition to the research-based conditions, a constrained bibliography-only condition was introduced as a stress test to examine hallucination behaviour under strict citation requirements. In this condition, models were prompted to generate bibliographies under explicit constraints (e.g., fixed reference counts and formatting requirements) using standard chat interaction without the Deep Research feature enabled. Prior work has documented that generative models frequently produce fabricated or incorrect bibliographic citations when asked to generate references, motivating the use of constrained bibliography generation as a stress-test scenario [4, 13]. This condition was not intended to reflect typical user behaviour, but rather to probe how models cope with strict citation requirements. The same well-known and niche topics were reused to maintain comparability across experimental conditions. The prompts used are shown in Table 2.

Prompt D - Well-known Topic	**Prompt E - Niche Topic**
Generate a bibliography of exactly 20 scholarly references on **Code Review Practices and Their Impact on Software Quality**. Every reference must be a real, verifiable publication. Use IEEE reference format, which requires the following details for each entry: the author(s), the year of publication, the title of the work, and the DOI. Only include works published between 2018 and 2025.	Generate a bibliography of exactly 20 scholarly references on **Adaptive Bug Prediction Using Multi-Agent Large Language Models**. Every reference must be a real, verifiable publication. Use IEEE reference format, which requires the following details for each entry: the author(s), the year of publication, the title of the work, and the DOI. Only include works published between 2018 and 2025.

Table 2. Constrained bibliography prompts.

After the papers and bibliographies were generated, the reference lists were extracted and evaluated using an automated reference cross-checking tool developed for this study. As shown in Figure 1, the tool queries three scholarly data sources, Crossref, OpenAlex, and arXiv, and when a DOI is present in the reference text, the system also attempts a direct DOI lookup. It then calculates an Integrity Score ranging from 0 to 100. Details on the metrics used are covered later in this section. The Integrity Score reflects how closely the metadata of a generated reference aligns with the best-supported scholarly record returned by these sources.

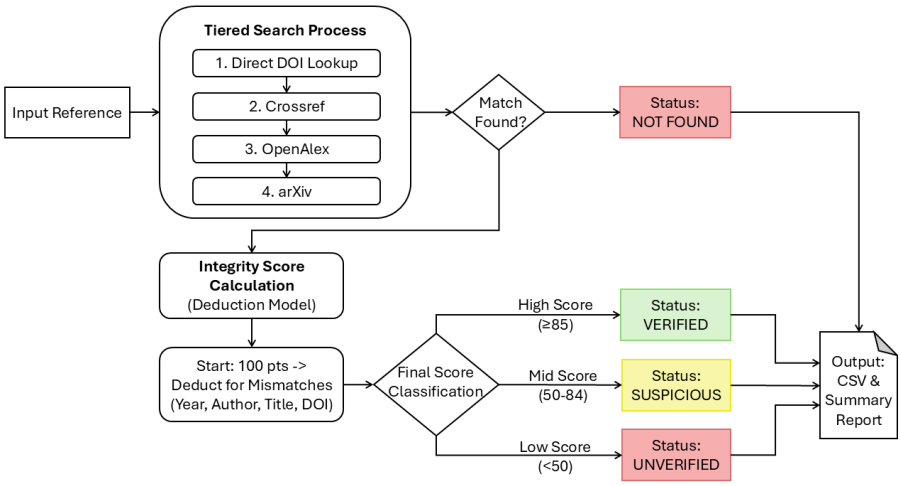

Figure 1. Reference cross-checking and integrity scoring workflow.

The process begins with metadata extraction from the reference text. For example, when a reference such as "Smith, J. (2021). Machine learning applications. Journal of AI, 15(3), 123-145." is provided, the system extracts key fields including the publication year (2021), the first author's last name (Smith), title keywords (such as "machine," "learning," and "applications"), and the DOI when available.

Following the initial extraction, the system performs a tiered, non-cascading search across Crossref, OpenAlex, and arXiv. Crossref serves as the primary scholarly metadata source, OpenAlex provides broader coverage, and arXiv provides coverage for preprints in fields such as computer science, mathematics, and physics. The records returned from these sources are compared against the extracted reference metadata, and the best-supported match is used for integrity scoring.

Once the database information is retrieved, the calculation of the Integrity Score begins. This score starts at 100 and deducts points for specific discrepancies found during the verification process, as shown in Table 3. For instance, a significant year mismatch (a difference of more than 1 year) results in a 25-point deduction, while a minor year mismatch (a difference of 1 year) results in a 10-point deduction. Title mismatch carries the largest penalty (-40 points), as the title serves as the strongest semantic indicator of whether the generated citation aligns with a scholarly record. In practice, title mismatch is assessed using both non-common title keyword overlap and fuzzy title similarity. Other deductions include author mismatches, DOI mismatches, missing DOI information in matched records, and very low Crossref scores for ordinary bibliographic Crossref matches. Fuzzy matching is also applied to author names and titles to accommodate minor variations, such as formatting differences, typographical errors, and altered word order.

Problem	Points Deducted	Example
Year mismatch (>1 year difference)	-25	Reference says 2021, database says 2019
Minor year difference (±1 year)	-10	Reference says 2021, database says 2020
First-author mismatch	-25	first author does not match the author list
Partial title mismatch	-15	Title only partially matches the database title
Title mismatch	-40	Title keywords do not match the database title
DOI mismatch	-15	Reference DOI differs from database DOI
Missing DOI in the database	-5	Reference has a DOI, but the database does not
Very low Crossref score (<5.0)	-10	Crossref has very low confidence in its match

Table 3. Integrity score.

Next, the system determines the reference's status based on its Integrity Score and source support, as shown in Table 4. These statuses range from higher-confidence verified matches to suspicious, unverified, or not-found cases. Direct DOI resolution can support a high-

confidence classification when the matched record aligns strongly with the generated citation. A downgrade rule is applied when the selected Crossref search result has a very low Crossref score, even if the Integrity Score would otherwise support a stronger classification. In this way, the verification status reflects the degree of database support for a generated reference, while the Integrity Score captures the degree of metadata consistency between the generated citation and the matched scholarly record.

Integrity Score	Source Support	Status	Meaning
85-100	Crossref score ≥ 10.0 or direct DOI resolution	VERIFIED - HIGH CONFIDENCE	Strong match with strong database support
85-100	Below high-confidence support threshold	VERIFIED	Good match but lower confidence
70-84	Crossref score ≥ 8.0 or direct DOI resolution with score ≥ 75	VERIFIED - MEDIUM CONFIDENCE	Reasonable match, minor issues
70-84	Below medium-confidence support threshold	SUSPICIOUS - Review Recommended	Borderline match requiring review
50-69	Any	SUSPICIOUS - Likely Mismatch	Major metadata inconsistencies
0-49	Any	UNVERIFIED - Poor Match	Very poor match, weak database support
Not found	-	NOT FOUND (All APIs)	No queried source returned a usable match

Table 4. Reference status. Verified Crossref search matches with Crossref scores below 5.0 are downgraded to a suspicious category. This downgrade does not apply to direct DOI-resolved matches.

The results of this verification process are presented in several formats. Real-time console output provides status indicators together with extracted and matched metadata, highlighting discrepancies where they occur. In addition, a CSV export is generated containing the original reference text, verification status, integrity score, source score when available, and reasons for any deductions. A summary report is also generated to provide statistics such as the percentage of verified, suspicious, unverified, and not-found references, along with a list of references requiring review.

For analysis, references were grouped as Valid or Problematic according to their final verification status. References classified as Verified were treated as Valid, whereas references classified as Suspicious, Unverified, or Not Found were treated as Problematic. Because the verifier relies on heuristic metadata extraction and public scholarly databases, problematic references should be interpreted as citations with weak database support or significant metadata inconsistency, rather than as definitive proof of fabrication in every individual case. Borderline cases, especially unusually formatted citations, preprints, books, and web-based sources, were treated as cases where automated results should be interpreted cautiously.

4. Results

Thirty-six research papers were generated across four LLMs using three prompting conditions. Reference verification revealed consistently high integrity scores across all models, indicating that, contrary to expectations, hallucinated or fabricated citations were relatively rare in the produced outputs.

4.1. Deep Research Results

Figure 2 summarizes average reference integrity scores across models for Prompts A, B, and C under the Deep Research condition, illustrating overall variation by prompt and model. Additionally, Figure 3 shows the distribution of valid and problematic references across models for the same prompts under the Deep Research condition.

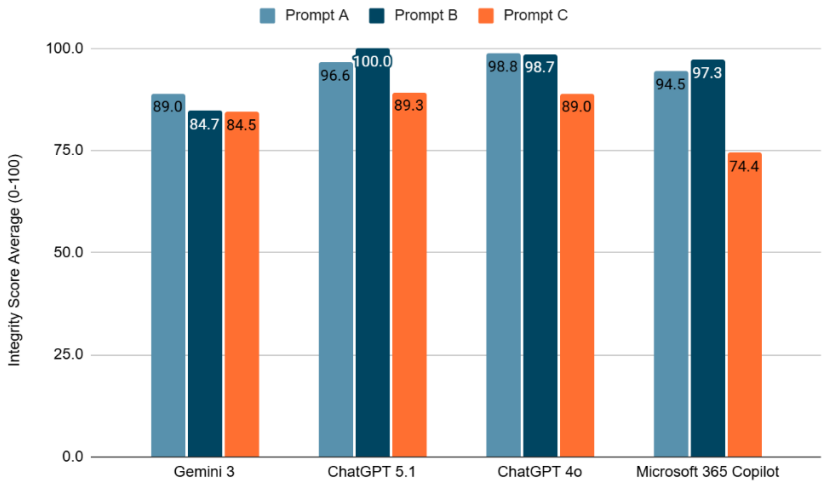

Figure 2. Average reference integrity scores across models for prompts A–C (deep research).

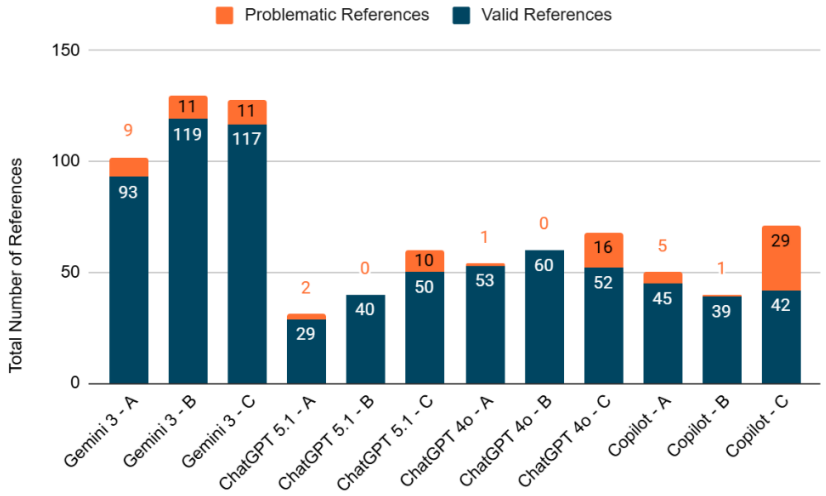

Figure 3. Distribution of valid and problematic references across models for prompts A–C (deep research).

Across the 834 total references analyzed, 95 were classified as problematic, yielding an overall hallucination rate of 11.39%. Integrity scores ranged from 74.4 to 100, indicating generally strong bibliographic accuracy overall, with noticeable variation depending on both model and prompt type.

4.1.1. Gemini 3

Gemini 3 generated the largest number of references, averaging 40 per paper, and showed relatively consistent performance across the three Deep Research prompts. Across its nine outputs, 31 problematic references were identified. Its integrity scores ranged from 84.5 to 89.0, showing consistent but comparatively moderate accuracy.

4.1.2. ChatGPT 5.1

The performance of ChatGPT 5.1 showed a moderate number of references per paper, ranging from 7 to 23, while maintaining strong consistency with relatively low error rates. Throughout all outputs, only 12 problematic references were identified. Integrity scores remained high, ranging from 89.3 to 100.0, with the strongest performance observed for Prompts A and B. The increase in problematic references under Prompt C indicates some decline in reference reliability when stricter reference instructions were introduced.

4.1.3. ChatGPT 4o

ChatGPT 4o produced the most text with good accuracy. Out of the nine papers reviewed, five had no problematic references. However, the papers generated for Prompt C, which included specific reference instructions, had 16 problematic references. The integrity scores for the papers ranged from 89.0 to 98.9, showing strong performance on open prompts. Still, there was a clear weakness when the reference requirements were explicitly stated.

4.1.4. Microsoft 365 Copilot

Copilot produced between 5 and 29 references per output and showed greater variability across prompts. A total of 35 problematic references were identified across nine papers, with the highest number occurring under Prompt C. Integrity scores ranged from 74.4 to 97.3, showing a generally strong performance on Prompts A and B, but a notable decline under Prompt C. As with other models, stricter reference instructions appeared to increase the likelihood of reference-level inaccuracies.

4.2. Effect of Prompt Type

The expected trend of higher hallucinations on niche topics was only partially supported by the findings. Prompt B, which focused on a niche topic, produced lower numbers of problematic references across most models (e.g., 0 for ChatGPT 5.1 and ChatGPT 4o, and 1 for Copilot), and maintained relatively high integrity scores. In contrast, Prompt C, which combined a niche topic with explicit reference guidance, resulted in the highest counts of problematic references across all models, particularly in ChatGPT 4o (16 problematic references) and Microsoft 365 Copilot (29 problematic references). This was associated with higher rates of problematic references in this experimental setting, whereas simply presenting a niche research question allows models to determine the number of references to use, leading to higher accuracy. For the well-known topic (Prompt A), the models displayed moderate levels of hallucinations, with each model producing between 1 and 9 problematic references. Notably, the highest accuracy was observed under Prompts A and B for ChatGPT 5.1 and Prompt B for ChatGPT 4o. These results challenge the initial hypothesis that niche topics inherently produce more hallucinations.

4.3. Topic Familiarity vs Model Behaviour

The data indicate that different models employ specific strategies when it comes to reference use. Gemini maximizes its reference volume, which in turn increases exposure to hallucinations. In contrast, ChatGPT 5.1 generated a smaller number of references and maintained consistently high integrity scores, suggesting a more conservative reference selection approach. ChatGPT 4o exhibited strong performance under Prompts A and B but showed a sharp increase in problematic references under Prompt C, indicating sensitivity to explicit reference constraints rather than topic familiarity alone. Microsoft 365 Copilot showed the greatest variability, sometimes taking a different approach by minimizing reference counts, which reduces the risk of hallucinations but may affect the academic richness of its responses. Overall, Copilot performed relatively well under Prompts A and B but showed a substantial increase in problematic references under Prompt C, resulting in the lowest integrity score among all models (74.4).

4.4. Stress Test

Prompts D and E were designed as constrained bibliography stress tests to examine how models handle strict citation requirements. Both prompts required the generation of exactly 20 scholarly references in IEEE format, limited to 2018–2025, with a DOI included for every entry. These constraints removed narrative context and required models to meet fixed bibliographic criteria.

Figure 4 summarizes average reference integrity scores across models under the constrained bibliography stress-test condition, highlighting variation between Prompts D and E.

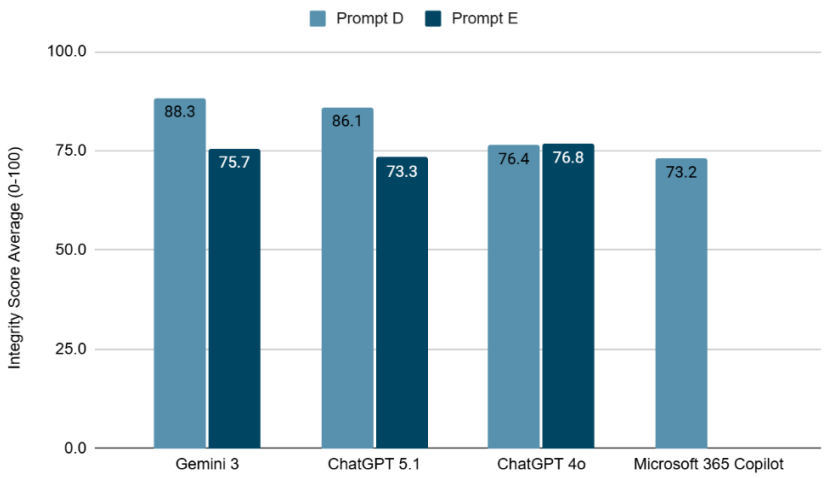

Figure 4. Average reference integrity scores across models for prompts D–E (bibliography stress test). For prompt E, Microsoft 365 Copilot returned no references, citing the absence of verifiable sources that met the prompt constraints.

Figure 5 summarizes the number of valid and problematic references per output. Under the stress test, models generated 9-32 problematic references per bibliography. Many of the errors involved incorrect or mismatched DOIs, suggesting that DOI requirements increase the likelihood of citation-level inaccuracies.

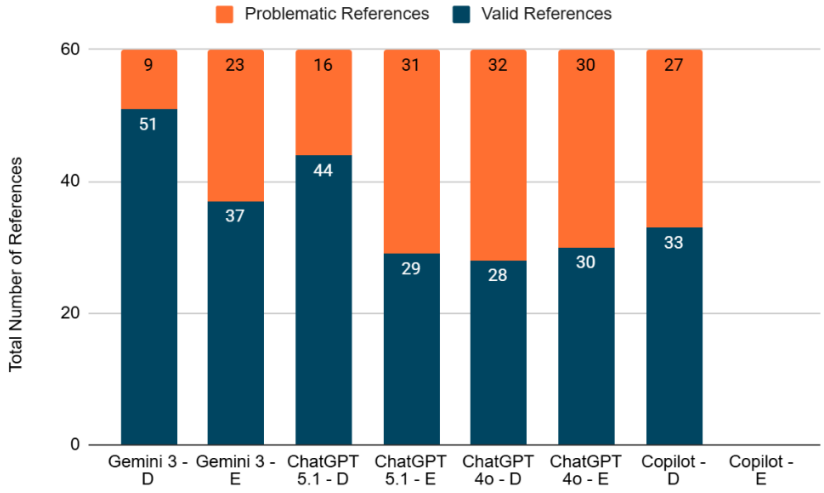

Figure 5. Distribution of valid and problematic references across models for prompts D–E (bibliography stress test). For prompt E, Microsoft 365 Copilot returned no references, citing the absence of verifiable sources that met the prompt constraints.

Although Prompt D covered a well-known topic, some problematic references were still observed. This suggests that reference inaccuracies under the stress test are influenced not only by topic familiarity but also by the requirement to satisfy fixed formatting and metadata constraints.

An additional outcome was observed for Microsoft 365 Copilot under Prompt E during both trials, rather than producing partially verifiable or fabricated references, Copilot returned no bibliography entries for the niche topic, indicating that it could not retrieve usable sources that satisfied the prompt constraints. The system explicitly stated that generating references without verifiable DOIs would violate its citation rules. As a result,

Copilot produced zero problematic references for Prompt E but also failed to meet the requirement to generate a complete bibliography.

In comparison with the Deep Research condition, the stress-test results suggest a shift in where hallucinations appear rather than a large increase in overall error. In Deep Research, reference integrity remained relatively high, with hallucinations more often appearing at the conceptual level. Under the stress test, reference-level inaccuracies were more common, indicating that output format and citation constraints influence how hallucinations manifest.

The proportion of problematic references was consistently higher under the stress-test prompts than under the Deep Research prompts across the three runs. This repeated pattern suggests a modest and relatively stable effect of bibliographic constraints on reference accuracy, although more extensive repeated sampling and formal statistical testing would be needed to confirm the strength of this effect.

4.5. Threats to Validity

This study has several limitations affecting internal validity. Results are based on a fixed set of prompts and models, with three runs per condition; because LLM generation is probabilistic, observed differences should be interpreted as trends rather than precise statistical estimates. Deep-research style outputs were time-intensive to generate, limiting repeated trials and experimental scale. The verification method evaluates bibliographic existence and metadata consistency but cannot determine whether cited sources actually support the generated claims, and therefore underrepresents deeper conceptual or reasoning-level inaccuracies. In addition, the Integrity Score thresholds used in this study should be interpreted as operational thresholds for this experimental setting rather than as universal standards. Although the scoring framework is intended to be broadly applicable, the cutoffs used to distinguish verified, suspicious, or high-confidence references may require recalibration in other domains, particularly in higher-stakes fields such as medicine or pharmaceuticals. Differences in reference volume and platform-specific behaviours across models may also influence the number of problematic references detected.

External validity is limited by the academic writing context. The use of long-form research prompts, Deep Research mode, and constrained bibliography tasks represents controlled evaluation conditions that may not reflect typical real-world usage. Additionally, only software-engineering topics and a small set of systems were examined, so hallucination behaviour may differ across domains, interaction styles, or future model versions.

4.6. Discussion

The most notable outcome was that hallucination rates were lower than anticipated, particularly for niche topics. Models often avoided citing uncertain sources and, in some cases, shortened their bibliographies when confidence was low, a behaviour notably observed in Copilot. This suggests that modern LLMs often avoid citing uncertain sources.

Although reference hallucinations were relatively rare, conceptual hallucinations occurred consistently in the niche-topic condition. All models generated papers that introduced fictional frameworks related to the emerging research area. These frameworks appeared credible, often including structured components and quantitative claims, yet were entirely invented. Rather than fabricating external citations, the models constructed internally coherent knowledge structures to fill conceptual gaps and maintain an academic narrative when factual grounding was limited.

This pattern is consistent with prior observations that hallucination behaviour varies with knowledge availability [14]. Across conditions, hallucination behaviour was shaped more by task structure and prompt constraints than by topic familiarity. Narrative, research-style prompts encouraged more cautious reference generation, whereas constrained bibliography tasks increased citation-level inaccuracies, consistent with prior findings that hallucination form depends strongly on task design [8].

The deduction-based Integrity Score was chosen for interpretability, since it makes explicit which bibliographic discrepancies reduce confidence in a generated reference. Although an additive alternative that accumulates evidence from zero could produce different classification behaviour, the present scheme was adopted as a conservative framework that emphasizes identifiable citation problems rather than partial agreement alone. This is also consistent with the broader pattern observed in the study: stricter output constraints often increased surface-level reference errors, similar to how highly specific prompting in other LLM tasks, such as code generation, can increase the likelihood of flawed outputs even when the response appears structured or plausible. These parallels further highlight the importance of verification and validation in accuracy-sensitive settings.

From a practical perspective, these results suggest that different prompting conditions produced different error types. Instead of obvious fabricated citations, they increasingly appear as invented frameworks, unsupported claims, or minor metadata inconsistencies. This highlights the need for context-aware deployment: accuracy-critical applications require stronger grounding and verification, while exploratory settings may tolerate limited speculative output. Making uncertainty visible, such as flagging unverifiable references or speculative claims, may further support more critical interpretation in academic and technical contexts [15].

From an ethical perspective, the use of Deep Research tools in academic-style writing raises concerns related to academic integrity and responsible disclosure. The generated papers often appeared credible and well-structured, even when they contained invented frameworks or unsupported claims, which could mislead users if the outputs are not critically verified. This highlights the importance of transparent use of AI-generated content, particularly in research and educational contexts. While Deep Research can support synthesis and exploration, its outputs should not be treated as authoritative sources without verification, especially when references and conceptual claims are involved.

Overall, reference hallucination was infrequent, context-sensitive, and most pronounced under explicit referencing pressure, indicating that strict formatting and bibliographic constraints influence how hallucinations manifest rather than simply increasing their frequency.

5. Conclusion and Future Work

This study shows that hallucinations in large language models depend strongly on task structure and citation requirements. Rather than simply increasing with topic difficulty, hallucinations changed forms across prompting conditions: constrained citation prompts produced more invalid references, while open prompts more often produced unsupported conceptual claims. These findings suggest that LLM reliability should be evaluated in relation to task design, particularly when models are used for research-style writing or literature assistance. Appropriate verification and uncertainty signaling, therefore, remain important when interpreting generated academic content.

Future work could extend this study by conducting larger-scale repeated sampling and controlled settings to better characterize variability across runs and models. While this study focused primarily on reference hallucinations, many errors occurred at deeper conceptual and reasoning levels, such as invented frameworks and unsupported claims, which remain difficult to systematically detect in long-form outputs. An important direction for future research is developing methods to distinguish between hallucinations that enable creative exploration and those that introduce misleading or harmful content across different interaction modes and research tasks.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Code Availability. The code used for reference verification in this study is publicly available at: https://github.com/Kimberly-D/CANAI-2026-reference-hallucinations-llms. The repository includes the final evaluation script (reference_verifier.py) and the materials needed to reproduce the reported verification results.

Declaration of AI Use. AI tools were used for limited editing, polishing, revision of author-written text, and assistance in organizing portions of the paper. They were not used to determine the study design, results, or conclusions. All critical thinking, analysis, and final content decisions were made by the authors.

References

[1] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans Inf Syst, vol. 43, no. 2, p. 42:1-42:55, Jan. 2025, doi: 10.1145/3703155.

[2] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why Language Models Hallucinate," Sep. 04, 2025, arXiv: arXiv:2509.04664. doi: 10.48550/arXiv.2509.04664.

[3] Y. Chen et al., "Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models," in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, in CIKM '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 245–255. doi: 10.1145/3583780.3614905.

[4] W. H. Walters and E. I. Wilder, "Fabrication and errors in the bibliographic citations generated by ChatGPT," Sci. Rep., vol. 13, no. 1, p. 14045, Sep. 2023, doi: 10.1038/s41598-023-41032-5.

[5] X. Jiang, Y. Tian, F. Hua, C. Xu, Y. Wang, and J. Guo, "A Survey on Large Language Model Hallucination via a Creativity Perspective," Feb. 02, 2024, arXiv: arXiv:2402.06647. doi: 10.48550/arXiv.2402.06647.

[6] S. Yuan, Z. Qu, A. Y. Kangen, and M. Färber, "Can Hallucinations Help? Boosting LLMs for Drug Discovery," Aug. 22, 2025, arXiv: arXiv:2501.13824. doi: 10.48550/arXiv.2501.13824.

[7] M. Cossio, "A comprehensive taxonomy of hallucinations in Large Language Models," Aug. 03, 2025, arXiv: arXiv:2508.01781. doi: 10.48550/arXiv.2508.01781.

[8] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," Sep. 14, 2025, arXiv: arXiv:2309.01219. doi: 10.48550/arXiv.2309.01219.

[9] Y. Bang et al., "HalluLens: LLM Hallucination Benchmark," Apr. 24, 2025, arXiv: arXiv:2504.17550. doi: 10.48550/arXiv.2504.17550.

[10] A. Agrawal, M. Suzgun, L. Mackey, and A. Kalai, "Do Language Models Know When They're Hallucinating References?," in Findings of the Association for Computational Linguistics: EACL 2024, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 912–928. Accessed: Dec. 03, 2025. [Online]. Available: https://aclanthology.org/2024.findings-eacl.62/

[11] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.

[12] A. Eghbali and M. Pradel, "De-Hallucinator: Mitigating LLM Hallucinations in Code Generation Tasks via Iterative Grounding," Jun. 19, 2024, arXiv: arXiv:2401.01701. doi: 10.48550/arXiv.2401.01701.

[13] J. Gravel, M. D'Amours-Gravel, and E. Osmanlliu, "Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions," Mayo Clin. Proc. Digit. Health, vol. 1, no. 3, pp. 226–234, Sep. 2023, doi: 10.1016/j.mcpdig.2023.05.004.

[14] Y. Zhang et al., "The Law of Knowledge Overshadowing: Towards Understanding, Predicting and Preventing LLM Hallucination," in Findings of the Association for Computational Linguistics: ACL 2025, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 23340–23358. doi: 10.18653/v1/2025.findings-acl.1199.

[15] Z. Li, W. Yi, and J. Chen, "Beyond Accuracy: Rethinking Hallucination and Regulatory Response in Generative AI," Oct. 23, 2025, arXiv: arXiv:2509.13345. doi: 10.48550/arXiv.2509.13345.