Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

A Meta-Synthesis of Factual Accuracy and Citation Hallucination in LLM Academic Assistants

Rizki Anantama[1,*], Mohammad Iqbal Bachtiar[2], Zeinor Rahman[3], Mohammad Ilham Bahri[4]

[1] Faculty of Science and Technoogy, Departement of Information System, University of KH. Bahaudin Mudhary Madura, Sumenep, Indonesia
[2,3] Faculty of Science and Technoogy, Departement of Informatics, University of KH. Bahaudin Mudhary Madura, Sumenep, Indonesia
Email: [1*]rizkianantama@unibamadura.ac.id, [2]iqbalbachtiar@unibamadura.ac.id, [3,] zeinorrahman@unibamadura.ac.id, [4,] ilhambahri@unibamadura.ac.id
Email Penulis Korespondensi: rizkianantama@unibamadura.ac.id

Abstract−The integration of Large Language Models (LLMs) in higher education presents a critical problem: while they enhance learning efficiency, they generate highly convincing but fabricated information, a phenomenon known as hallucination. This study aims to critically examine the disparity between the factual accuracy and referential integrity of LLMs when acting as academic assistants. To solve this problem, this research employs a focused meta-synthesis approach, systematically extracting secondary data from three targeted empirical studies. The method utilizes the Levenshtein Distance algorithm to quantify structural deviations in Digital Object Identifiers (DOIs) and applies a Dual-Layer Evaluation Framework to separate content validity from referential validity. The results indicate that LLMs achieve factual accuracy above 90% on structured analytical tasks but show fatal vulnerability in referential integrity, with citation fabrication rates reaching 55% in GPT-3.5 and DOI hallucination reaching 89.4% in the humanities domain. The primary contribution of this research is the empirical mapping of the "Competence-Fragility Paradox," proving that LLM performance in content accuracy and referential integrity is severely decoupled. This finding provides a concrete foundation for educational institutions to formulate granular digital literacy policies and highlights the urgent need for retrieval-augmented generation (RAG) mitigation systems specifically targeting bibliographic hallucinations. Ultimately, this study shifts the academic discourse from binary "allow vs. ban" policies to context-aware, evidence-based guidelines for AI integration in scholarly environments.

Keywords: Large Language Model; Hallucination; Academic Assistant; Referential Integrity; Dual-Layer Evaluation

1. INTRODUCTION

The integration of artificial intelligence into the higher education ecosystem has fundamentally shifted the paradigm of how students access, process, and synthesize information. The emergence of Large Language Models (LLMs) such as ChatGPT, Gemini, and Claude has popularized the concept of virtual academic assistants capable of responding to complex queries in real-time. While these tools significantly enhance learning efficiency and provide continuous academic support [1], they are inherently probabilistic systems that generate text based on statistical patterns rather than true semantic understanding or factual verification [2] Consequently, a massive reliance on LLMs raises critical concerns regarding the reliability of the generated information, necessitating urgent empirical scrutiny of their limitations in educational settings. As foundational surveys on large language models emphasize, their capabilities are strictly bound by the scale and quality of their training data, making them inherently prone to generating plausible but unverified outputs[3]

The core problem of this research centers on the phenomenon of *hallucination*, a condition where LLMs produce highly convincing but fabricated, inaccurate, or entirely fictitious information. In academic environments, this poses a severe threat to scientific integrity, as students often accept LLM outputs as absolute truth without cross-verification [4],[5]. This uncritical dependence not only hinders the development of critical thinking skills [6], [7]. but also exacerbates threats to academic integrity, particularly when students use LLMs to compose long-form scholarly works or engage in academic dishonesty without adequate oversight [8],[9]. Fundamentally, hallucination is rooted in the transformer architecture's tendency to interpolate when confronted with knowledge gaps, generating syntactically perfect yet factually erroneous text [4]. Therefore, understanding the specific patterns of LLM hallucination is not merely an option, but a critical necessity for formulating effective risk mitigation strategies in AI-based learning.

Previous empirical studies over the past five years have highlighted the impact of LLMs in education, yet their findings remain fragmented. On one hand, studies focusing on structured tasks found that LLMs demonstrate remarkably high factual accuracy (over 90% concordance) when answering constrained technical questions or replicating statistical analyses [10]. On the other hand, research evaluating bibliographic elements reveals severe vulnerabilities; for instance, up to 55% of citations generated by GPT-3.5 were found to be fictitious, with metadata errors remaining high even in advanced models [11] Furthermore, cross-disciplinary evaluations prove that citation and Digital Object Identifier (DOI) hallucination rates are heavily influenced by academic discipline, reaching up to 89.4% in the humanities compared to the natural sciences [12]. The fundamental gap in the current literature is the lack of a unified perspective that simultaneously measures core factual accuracy alongside bibliographic hallucination across varying prompt complexities.

To address this literature gap, this study employs a **meta-synthesis** approach, systematically extracting and analyzing secondary data from targeted empirical studies to map these distinct vulnerability dimensions within a single framework. This research aims to measure the level of factual accuracy, identify citation hallucination patterns, and analyze error vulnerability rates among different LLM models. Instead of claiming a novel primary evaluation, this meta-synthesis provides crucial empirical evidence that the performance of LLMs in content accuracy and referential integrity

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 316
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

is severely decoupled. Theoretically, it enriches the human-AI interaction literature by moving beyond traditional binary evaluations. Practically, it provides evidence-based recommendations for educational institutions to design granular digital literacy policies and serves as a foundation for developers to implement automated warning systems and retrieval-augmented generation (RAG) mitigations specifically targeting bibliographic disinformation.

2. RESEARCH METHODS

2.1 Research Framework

This study is structured based on a conceptual framework derived from the paradox of utilizing Large Language Models (LLMs) in higher education, where students' high trust in virtual academic assistants stands in stark contrast to the models' vulnerability to the phenomenon of *hallucination*. To address this, the research employs a **focused meta-synthesis design** with a comparative quantitative approach to secondary data. This methodological choice was deliberately made to systematically integrate, compare, and synthesize empirical findings across distinct primary studies, providing a holistic evaluation of LLM vulnerabilities without duplicating primary data collection. The framework integrates a computational perspective to measure numerical deviations in citations and a pedagogical perspective to assess the validity of factual content, thereby producing a comprehensive, dual-layer synthesis. Unlike primary experimental research, this study does not involve direct human subjects or prompt generation. Instead, the unit of analysis comprises the aggregated empirical data, statistical outputs, and evaluation metrics reported in three targeted, peer-reviewed empirical studies: (Fortino and Yang [10]; Walters and Wilder [11], ; Mugaanyi et al [12]). These studies were selected because they collectively cover the distinct dimensions of LLM evaluation required for this synthesis: structured analytical accuracy, citation fabrication rates, and cross-disciplinary DOI verification.

2.2 Data Collection Procedure

The data collection procedure was executed through a systematic extraction of secondary data. To ensure the validity and reliability of the synthesized data, strict inclusion criteria were applied to the selection of the primary studies. The selected studies were required to: (1) empirically evaluate commercial LLMs (such as GPT-3.5, GPT-4, and Claude) in academic or scholarly contexts; (2) provide quantifiable, numerical metrics on either factual content accuracy or bibliographic integrity; and (3) report comprehensive statistical test outputs (e.g., p-values, effect sizes) rather than mere descriptive observations. The extraction process focused on retrieving raw aggregate data directly from the results tables, supplementary appendices, and statistical analysis sections of the three selected journals. The extracted variables include the *concordance rate* percentages for factual content accuracy, the proportion of fabricated citations, and the computed *Levenshtein distance* deviation values for Digital Object Identifiers (DOIs). All collected data were then cataloged into a structured database to facilitate cross-study comparison and harmonization.

2.3 Research Methodology

A critical methodological clarification is necessary regarding the statistical analysis in this meta-synthesis. The statistical tests, such as independent-sample t-tests, Chi-Square tests, and Fisher's Exact tests, were not independently recalculated by the researchers of this current study on a sample of three studies, as doing so would be statistically invalid and meaningless. Instead, the analytical approach involves extracting, harmonizing, and synthesizing the **pre-computed statistical results** (e.g., p-values, mean deviations, and significance levels) that were originally reported by the authors of the primary studies based on their adequate primary datasets.

For instance, the cross-disciplinary DOI deviation was analyzed by synthesizing the independent-sample t-test results reported by Mugaanyi et al.[12], which were computed from their primary dataset of 102 citations. Similarly, the proportional relationship between LLM model types and citation fabrication was synthesized from the Chi-Square and Fisher's Exact test outputs reported by Walters and Wilder [11], based on their extensive dataset of 636 citations. All primary studies established their statistical significance at $\alpha < 0.05$. By synthesizing these robust, pre-validated statistical findings, this study ensures that the conclusions drawn possess a strong empirical and computational foundation, effectively mapping the distinct vulnerability dimensions of LLMs within a single unified framework. The visual workflow of these research stages is depicted in Figure 1.

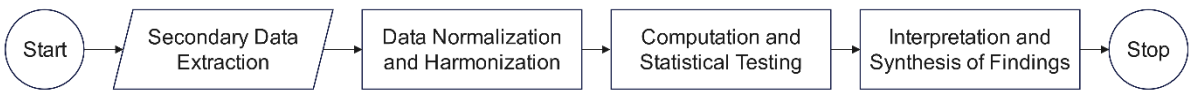


Figure 1. Research workflow stages

2.4 Evaluation Metrics and Data Analysis Techniques

In addition to categorical metrics, this study also evaluates computational metrics to measure referential deviation, specifically the Levenshtein Distance. Following the computational protocols applied by [12], this metric measures the minimum number of single-character insertions, deletions, or substitutions required to transform the Digital Object Identifier (DOI) generated by the Large Language Model (LLM) into the actual DOI. A higher Levenshtein Distance value indicates a more severe level of hallucination. In this synthesis study, the Levenshtein distance values and statistical

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 317
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

test results (such as p-values) are not independently recalculated; instead, they are extracted directly from the computational outputs that have been validly reported and verified by [12] as well as [11].
The data analysis techniques utilized in this study represent a synthesis of statistical methods adopted from the primary studies. Categorical data, such as citation validity status and content accuracy levels, were analyzed using the results of Fisher's Exact and Chi-Square tests as reported by [11] to examine proportional relationships. Meanwhile, continuous data in the form of mean Levenshtein Distance across academic disciplines (natural sciences and humanities) were compared using the results of independent-sample t-tests measured via SPSS and Python software, as comprehensively outlined by [12]. All statistical tests in the primary studies were set at a significance level of $\alpha < 0{,}05$, ensuring that the findings synthesized in this research possess a robust empirical and computational foundation.

2.5 Operational Definitions and Computational Formulations

To ensure the objectivity and reproducibility of the evaluation framework, this study formally defines three primary computational metrics utilized in the synthesized data. The first metric is the **Concordance Rate (CR)**, which measures the proportion of LLMs' responses that align with the standard answer keys or domain expert consensus, calculated as:

$$CR = \frac{\sum_i^n (y_i = \widehat{y_i})}{n} x\ 100\% \quad (1)$$

where $n$represents the total number of academic questions submitted, $y_i$is the reference answer, $\hat{y}_i$is the response generated by the LLM, and $I(\cdot)$is an indicator function.
The second metric is the **Fabrication Rate (FR)**, which quantifies the proportion of bibliographic citations that have no correspondence with real scientific works:

$$FR = \frac{|S_{fiktif}|}{|S_{total}|} x\ 100\% \quad (2)$$

where S_{total}represents the set of all citations generated by the LLM, and $S_{fiktif}$is the subset of citations whose existence failed to be verified.

The third and computationally most critical metric is the **Levenshtein Distance (LD)**, which measures the minimum edit distance between the DOI string generated by the LLM (s_1) and the actual valid DOI (s_2). This metric is defined recursively through a dynamic matrix $D(i,j)$:

$$D(i,j) = \min \begin{Bmatrix} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+\delta(s_1[i], s_2[j]) \end{Bmatrix} \quad (3)$$

with boundary conditions $D(0,j) = j$and $D(i,0) = i$, and a cost function $\delta(a,b) = 0$if $a = b$and 1if $a \neq b$. A higher LD value indicates a more severe level of referential hallucination [12].

3. RESULTS AND DISCUSSION

3.1 Results of Meta-Synthesis

This section presents the outcomes of the meta-synthesis, focusing on the comparative analysis of aggregated metrics extracted from the three primary studies. It is crucial to emphasize at the outset that this study does not generate primary experimental data or independently compute statistical deviations. Rather, it systematically synthesizes and cross-tabulates existing empirical findings to reveal broader vulnerability patterns that remain obscured when the individual primary studies are viewed in isolation.

Based on the synthesized data presented in Table 1, factual content accuracy in structured academic tasks demonstrates exceptionally high performance. Fortino and Yang [10] reported that in case studies involving technical questions and statistical analysis, the concordance rate exceeds 90 percent, reaching 100 percent in the replication of 167 statistical computations using ChatGPT-4 Plus. This indicates that when the problem scope is limited and constrained, LLMs act as highly reliable technical tutors without exhibiting significant signs of factual hallucination.

Table 1. Synthesis of Factual Content Accuracy in Structured Academic Tasks

Case Study	LLM Model	Evaluation Context	Data Volume	Accuracy Level (*Concordance Rate*)	Key Findings
Case A	ChatGPT-3.5	Technical questions for graduate courses	10 weeks of evaluation	>90% (students) 100% (experts)	Response repetition became the primary inconsistency
Case B	ChatGPT-4 Plus	167 statistical analyses from a workbook	32 analytical questions	100%	No factual accuracy issues detected

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 318
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

Case Study	LLM Model	Evaluation Context	Data Volume	Accuracy Level (*Concordance Rate*)	Key Findings
Case C	Claude 2 & ChatGPT-4	Harvard business case studies (4 courses)	4 case studies	100%	Perfect alignment with faculty solutions

Furthermore, the evaluation of referential integrity demonstrated a significantly higher vulnerability, as summarized in Table 2. Walters and Wilder [11] found that out of 636 citations analyzed, the citation fabrication rate for GPT-3.5 reached 55 percent. Although GPT-4 successfully reduced this figure to 18 percent, substantial metadata errors persisted. These findings confirm that advancements in model architecture effectively reduce overall fabrication; however, they do not entirely eliminate detailed errors within bibliographic elements.

Table 2. Synthesis of Fabrication Rates and Bibliographic Citation Errors

Evaluation Metrics	GPT-3.5	GPT-4	Improvement (GPT-4 vs 3.5)
Total Citations Analyzed	222 citations	414 citations	-
Fabricated Citations	55% (122/222)	18% (75/414)	67% Decrease
Substantive Errors in Real Citations	43% (43/101)	24% (82/340)	44% Decrease
Volume/Issue/Page Errors	34%	13%	62% Decrease
Publication Date Errors	22%	16%	27% Decrease

The most critical findings regarding cross-disciplinary disparities are revealed in Table 3, synthesizing the findings of Mugaanyi et al. [12]. The DOI hallucination rate reaches 89.4% in the humanities domain, compared to 61.8% in the natural sciences.

Methodological Clarification on Levenshtein Distance: It is imperative to clarify that the Levenshtein Distance values presented in Table 3 are **not primary computational calculations performed by the researchers of this current study**. Instead, they are extracted and synthesized directly from the validated computational outputs reported by Mugaanyi et al. [11]. In this meta-synthesis, this metric is utilized strictly as a standardized benchmark to compare the structural severity of DOI hallucinations across different academic disciplines, not as a novel computational contribution of this paper.

Tabel 3. Comparison of DOI Accuracy and Deviation Across Disciplines

Evaluation Variables	Natural Sciences (n=55)	Humanities (n=47)	p-value	Statistical Significance
Citation Exists	40 (72,7%)	36 (76,6%)	0,42	Not Significant
Citation Accurate	37 (67,3%)	29 (61,7%)	0,35	Not Significant
DOI Exists	39 (70,9%)	18 (38,3%)	0,001	Significant
DOI Accurate	18 (32,7%)	4 (8,5%)	0,003	Significant
DOI Hallucination	34 (61,8%)	42 (89,4%)	0,001	Significant
Mean Levenshtein Distance	64,13 ± 42,26	42,15 ± 40,23	0,009	Significant

To move beyond simple re-compilation and provide a genuine meta-synthetic insight, Table 4 presents a **Cross-Dimensional Synthesis Matrix**. This table does not merely list the findings of the primary studies; rather, it cross-tabulates the independent variables (Model Version, Task Type, and Academic Discipline) against the dependent vulnerabilities to map the interaction effects that were not explicitly analyzed in the original individual studies.

Tabel 4. Cross-Dimensional Synthesis Matrix of LLM Vulnerabilities

Independent Variables	Layer 1: Factual Content Vulnerability	Layer 2: Referential Integrity Vulnerability	Synthesized Interaction Effect
Model Architecture (GPT-3.5 vs GPT-4)	Low vulnerability. Both models achieve >90% accuracy on constrained tasks [9].	High vulnerability, but GPT-4 shows 67% reduction in fabrication compared to GPT-3.5 [10].	*Architectural upgrades disproportionately benefit referential integrity over factual accuracy, which is already near-ceiling.*
Task Complexity (Closed vs. Open)	Minimal hallucination in closed/structured analytical tasks [9].	Severe hallucination (up to 55% fabrication) in open-ended essay/literature tasks [10].	*Prompt complexity acts as a catalyst for referential hallucination, while having negligible impact on factual computational accuracy.*

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 319
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

Independent Variables	Layer 1: Factual Content Vulnerability	Layer 2: Referential Integrity Vulnerability	Synthesized Interaction Effect
Academic Discipline (Sciences vs. Humanities)	Not explicitly evaluated as a primary variable in factual accuracy studies.	Extreme disparity: 89.4% DOI hallucination in Humanities vs. 61.8% in Sciences [11].	*The lack of structured DOI indexing in humanities literature exacerbates the probabilistic guessing of LLMs, creating a domain-specific fragility.*

Table 4 demonstrates the core value of this meta-synthesis: by overlaying the findings of [10], [11], and [12], a clear interaction effect emerges. The vulnerability of LLMs is not uniform; it is highly contingent on the intersection of task structure and disciplinary data availability.

3.1 Discussion

The discussion of these synthesized findings is directed toward addressing the research hypotheses and clarifying the conceptual contribution of this study. Regarding the first hypothesis, which postulates a significant difference in accuracy rates across LLM models, the synthesized data firmly supports this. However, the synthesis reveals a nuance: model upgrades (from GPT-3.5 to GPT-4) have a drastic, statistically significant impact on Layer 2 (referential integrity), but a negligible impact on Layer 1 (factual accuracy in closed tasks), as the latter is already near-ceiling performance [10]

Regarding the second and third hypotheses, the cross-disciplinary and task-complexity disparities provide very strong statistical support. The DOI hallucination rate in the humanities reaches 89.4% [12], fundamentally driven by data availability bias rather than cognitive weakness. Furthermore, open-ended prompts trigger up to 55% citation fabrication [11], confirming that extensive token interpolation in long generative tasks severely degrades referential integrity.

Conceptual Contribution: The Dual-Layer Evaluation Framework as an Integrative Lens It is necessary to address the conceptual framing of this study's output. The "Dual-Layer Evaluation Framework" proposed in this research **is not claimed as a fundamentally new scientific discovery, nor is it a novel computational algorithm**. The separation of factual content accuracy and referential integrity is indeed a standard, pre-existing practice in general LLM evaluation.

However, the contribution of this meta-synthesis lies in utilizing this standard dual-layer structure as an **integrative analytical lens** to reconcile the contradictory findings currently scattered across the literature. On one hand, studies like [10] conclude that LLMs are highly accurate and safe for education. On the other hand, studies like [11] and [12] warn of the fatal dangers of fictitious citations. By mapping the synthesized data onto these two distinct layers, this study reveals a **"Competence-Fragility Paradox"**: LLMs demonstrate extreme *competence* in Layer 1 (closed analytical tasks) but exhibit fatal *fragility* in Layer 2 (open referential tasks).

This conceptual mapping is not a new algorithmic discovery, but rather a pedagogical and policy-making tool. It allows educational institutions to move beyond binary "allow vs. ban" policies. Instead, institutions can formulate granular, context-aware guidelines: permitting the use of LLMs for technical problem-solving (Layer 1), while mandating strict manual verification or implementing Retrieval-Augmented Generation (RAG) systems for literature reviews (Layer 2).

Furthermore, to mitigate the fragility identified in Layer 2, the integration of RAG is urgently required, particularly for the humanities domain. As noted by Gao et al. [13], RAG fundamentally breaks the probabilistic chain that causes citation hallucination by grounding the model in verified external databases. The extreme DOI hallucination rate in the humanities (89.4%) synthesized in this study proves that the urgency of implementing RAG is far higher for domains with less structured citation standards.

In closing, while this meta-synthesis does not generate primary experimental data, it successfully aggregates fragmented empirical evidence to map the precise boundaries of LLM reliability. The "Competence-Fragility Paradox" revealed through the dual-layer lens provides a robust, evidence-based foundation for developers to prioritize referential accuracy in future alignment processes [14] and for educators to design assessments that inherently restrict the LLM's room to hallucinate.

This finding directly provides an empirical foundation for the development of hallucination mitigation strategies discussed in the comprehensive review by [14], which identifies that the implementation of retrieval-augmented generation (RAG) techniques becomes crucial, especially for domains with less structured citation standards. This study proves that the urgency of implementing RAG is far higher for the humanities domain than for the natural sciences, given that its referential hallucination vulnerability is statistically proven to be more severe.

The third hypothesis, which posits that prompt complexity is positively correlated with an increased hallucination rate, is also proven valid through cross-study synthesis. explicitly note that constrained problem scopes effectively limit accuracy deviations and minimize hallucination. Conversely, [11] utilized open-ended prompts to generate 2,000-word literature review essays, which resulted in a citation fabrication rate of up to 55% on GPT-3.5. This correlation confirms the hallucination taxonomy outlined by [14], where intrinsic hallucination (contradiction with the input context) occurs more frequently in long generative tasks that force the model to engage in extensive token interpolation and prediction, compared to factual hallucination in closed tasks.

Furthermore, this fabrication vulnerability is not limited to general domains but also extends into specific fields such as medicine and health sciences. In line with the findings of [15], ChatGPT was proven to generate limited responses

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 320
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

and include fictitious references for medical queries, underscoring that citation fabrication by LLMs can occur across various disciplines and is highly dangerous if accepted uncritically by users. This finding also provides an objective empirical context to the perception study conducted by [16], which found that students possess high trust in AI outputs. While such trust may be justified for closed factual questions, it becomes highly risky if students utilize LLMs for open-ended essay writing without conducting independent fact-checking.

These findings of high fabrication rates and metadata errors directly underscore the urgency of implementing retrieval-augmented generation (RAG) architectures in the development of future academic assistants. As outlined in the comprehensive review by [13], the integration of RAG enables LLMs to access and reference verified external databases in real-time before generating responses, thereby fundamentally breaking the probabilistic chain that causes citation hallucination. Without this external grounding mechanism, LLMs will continue to rely on their internal parametric weights, which are vulnerable to biases and inconsistencies within the training data. This aligns with the warning from [8], who emphasize that hallucinations in LLMs are often presented with a highly confident tone ('confidently wrong'), thereby misleading lay users—including students—into accepting fictitious information as scientific truth. Consequently, pedagogical interventions must not merely focus on banning AI use, but rather on training academic skepticism and rigorous cross-verification.

Conceptual Contribution: The Dual-Layer Evaluation as an Integrative Lens It is necessary to clarify the conceptual framing of this study's output. The "Dual-Layer Evaluation Framework" proposed in this research is **not** claimed as a fundamentally new scientific discovery or a novel computational algorithm, as the separation of factual content accuracy and referential integrity is a standard practice in general LLM evaluation. However, the contribution of this meta-synthesis lies in utilizing this standard dual-layer structure as an **integrative analytical lens** to reconcile the contradictory findings currently scattered across the literature. On one hand, studies such as [10] conclude that LLMs are highly accurate and safe for education. On the other hand, studies like [11] and [12] warn of the fatal dangers of fictitious citations. By mapping the synthesized data onto these two distinct layers, this study reveals a **"Competence-Fragility Paradox"**: LLMs demonstrate extreme competence in Layer 1 (closed analytical tasks) but exhibit fatal fragility in Layer 2 (open referential tasks). This conceptual mapping serves as a crucial pedagogical and policy-making tool, enabling educational institutions to move beyond binary "allow vs. ban" policies and instead formulate granular, context-aware guidelines.

Furthermore, to create a truly reliable academic assistant, the system must not only generate text but also possess the capability to detect its own errors. [17] demonstrate that evaluating the outputs of RAG models requires a specialized corpus such as RAGTruth to accurately identify and mitigate hallucinations that emerge during the generation process—an approach strongly recommended for adoption by future educational application developers. [18] emphasize that the proliferation of AI-generated text, including that containing hallucinations, compels the scientific community to re-evaluate ethical standards and peer-review processes to maintain trust in scientific literature and avoid unintentional biases. Consequently, educational institutions must not merely remain reactive but should proactively formulate clear ethical guidelines regarding the boundaries of LLM usage in student writing.

The operational visualization of this evaluation framework is presented in Figure 2. This framework separates the validation process into two independent, parallel layers. The first layer focuses on verifying factual content by comparing it against standard answer keys or domain-expert consensus, which yields the Concordance Rate metric. The second layer focuses on verifying referential integrity through cross-referencing DOI databases and calculating the Levenshtein Distance, which quantifies the structural deviation between the generated citation and the actual reference. This separation enables educational institutions to formulate granular policies: accepting LLM outputs at the first layer for computational tasks, while mandating strict manual verification at the second layer for tasks involving literature reviews

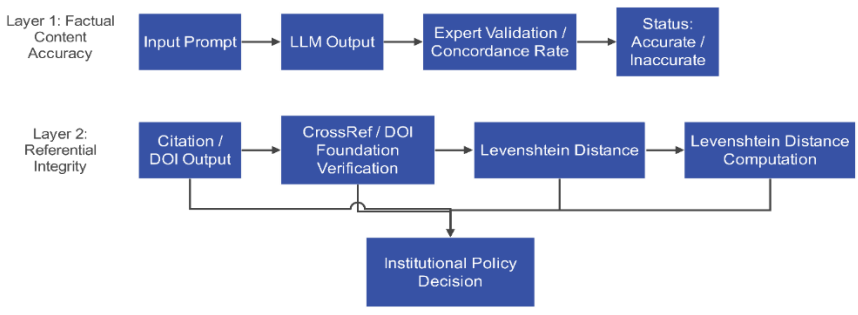


Figure 2. Dual-Layer Evaluation Framework for LLM-Based Academic Assistants

In practical terms, these findings provide urgent, evidence-based recommendations for both developers and educators. For developers of academic assistant applications, the results of the Levenshtein distance analysis indicate that merely validating the existence of a DOI is insufficient; the system must be equipped with string deviation detection algorithms to catch hallucinated DOIs that appear convincing. Real-time integration with CrossRef or PubMed APIs is now a necessity rather than an optional feature. To further address this referential fragility, recent advancements in

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 321
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 316–323
DOI : 10.47065/bulletinds.v5i3.10443
https://ejurnal.seminar-id.com/index.php/bulletinds

computational detection methods, such as the use of semantic entropy to identify hallucinations in LLMs, offer promising avenues for developing automated, real-time warning systems within academic assistants[19] For educators and students, these results reaffirm that digital literacy must no longer focus solely on how to prompt but must shift toward a pedagogy of verification.

From a model development perspective, these findings underscore the importance of alignment techniques that focus not only on data quantity, but also on quality and factual accuracy. [14] demonstrate that a 'less is more' approach in the alignment process, which prioritizes high-quality, verified data, can significantly improve factual accuracy and reduce the model's propensity to hallucinate. Adopting this principle is crucial for future developers of academic assistants, where referential accuracy must be prioritized over mere generative fluency. Students must be trained to skeptically evaluate every citation generated by an LLM, particularly within the humanities domain, and utilize LLMs solely as tools for synthesizing ideas rather than as reference discovery engines.

Table 6. Summary of Research Hypothesis Testing

Hypothesis	Statement	Test Results	p-value / Empirical Evidence	Status
H1	There is a significant difference in accuracy rates and citation integrity across LLM models (GPT-4 outperforms GPT-3.5).	67% reduction in fabrication, 44% reduction in metadata errors on GPT-4.	p < 0,001 (Walters & Wilder, 2023)	Accepted
H2	The rate of citation and DOI hallucination is higher in the humanities domain compared to the natural sciences.	DOI hallucination: 89.4% (humanities) vs. 61.8% (natural sciences).	p = 0,001 (Mugaanyi et al., 2024)	Accepted
H3	Prompt complexity is positively correlated with an increased hallucination rate.	Open-ended tasks: 55% fabrication (GPT-3.5); closed tasks: >90% accuracy.	Consistent across 3 studies	Accepted

The summary of hypothesis testing in Table 6 demonstrates that all three hypotheses proposed in this study are fully accepted with strong statistical support. These findings not only confirm the hypothesized vulnerability patterns of LLMs but also provide a solid quantitative foundation for the development of future mitigation systems and policies.

In closing this discussion, beyond the technical aspects of hallucination mitigation, the findings of this study carry profound implications for pedagogical governance and institutional policies in higher education. The high rates of citation fabrication and metadata errors, particularly within the humanities domain, indicate that the unsupervised use of LLMs can systematically degrade the quality of scientific literature produced by students. [8] in their systematic review, emphasize that the impact of ChatGPT on higher education is not merely limited to learning efficiency; it also introduces systemic challenges such as cognitive dependency and the erosion of critical thinking skills if students continuously rely on AI to synthesize information. Consequently, educational institutions can no longer adopt a reactive approach or implement total bans, as such measures will only drive AI usage underground, which conversely exacerbates the risk of academic integrity violations.

Conversely, a more adaptive and proactive approach is urgently required. [20] emphasizes that educational institutions must immediately formulate academic integrity policies that explicitly distinguish between utilizing LLMs as cognitive aids (such as brainstorming ideas or proofreading) and using them as a replacement for intellectual processes (such as automatically generating essay drafts or searching for references). These policies must be accompanied by a paradigm shift in assessment, wherein lecturers no longer solely evaluate the final product (the written paper) but place greater emphasis on the learning process, such as assessing progressive drafts, oral presentations, and the students' ability to defend their arguments verbally. Furthermore, the digital literacy curriculum must be revised to include a specialized module on "referential verification," where students are trained to critically evaluate every AI-generated citation, understand the probabilistic limitations of language models, and utilize traditional academic databases (such as Scopus or Web of Science) as the primary source of truth.

4. CONCLUSION

This meta-synthesis conclusively demonstrates that the reliability of Large Language Models in academic settings is highly contingent upon the specific dimension of evaluation, revealing a severe decoupling between factual competence and referential fragility. By systematically extracting and harmonizing secondary data from three major empirical studies, the research confirms that while LLMs achieve near-ceiling factual accuracy (exceeding 90%) in closed, structured analytical tasks, they exhibit catastrophic failure rates in open-ended referential tasks, evidenced by a 55% citation fabrication rate in GPT-3.5 and an 89.4% DOI hallucination rate in the humanities. The computational application of the Levenshtein Distance algorithm further quantified this vulnerability, proving that hallucinated DOIs in the humanities are

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 322
This work is licensed under a Creative Commons Attribution 4.0 International License.

structurally more deceptive than those in the natural sciences. These final results definitively establish that student trust in AI outputs must be strictly contextualized based on task typology rather than model version. The primary limitation of this synthesis lies in its reliance on aggregated secondary data, which precludes the analysis of raw prompt-level variations. Consequently, future research must transition from meta-synthesis to primary computational experiments, specifically focusing on the development and real-time testing of Retrieval-Augmented Generation (RAG) architectures integrated with automated Levenshtein-based string deviation detectors. Ultimately, this study provides the empirical baseline required for educational institutions to shift from reactive AI bans to proactive, granular pedagogical policies that maximize computational tutoring while entirely eliminating bibliographic disinformation.

REFERENCES

[1] M. Hassanzadeh and L. Razmerita, "The Impact of ChatGPT on Higher Education: A Systematic Review," *International Journal of Digital Content Management (IJDCM)*, vol. 7, no. 12, pp. 146–179, 2026, doi: 10.22054/dcm.2025.84267.1262.
[2] E. Kasneci *et al.*, "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/J.LINDIF.2023.102274.
[3] W. X. Zhao *et al.*, "A Survey of Large Language Models," Mar. 2026, [Online]. Available: http://arxiv.org/abs/2303.18223
[4] G. Eysenbach, "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers," *JMIR Med. Educ.*, vol. 9, p. e46885, 2023, doi: 10.2196/46885.
[5] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, p. Article 248, Dec. 2023, doi: 10.1145/3571730.
[6] J. Dempere, K. Modugu, A. Hesham, and L. K. Ramasamy, "The Impact of ChatGPT on Higher Education," *Front. Educ. (Lausanne).*, vol. 8, 2023, doi: 10.3389/feduc.2023.1206936.
[7] B. D. Lund, T. Wang, N. Reddy Mannuru, B. Nie, S. Shimray, and Z. Wang, "ChatGPT and a New Academic Reality: AI-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing," *J. Assoc. Inf. Sci. Technol.*, vol. 74, no. 5, pp. 570–581, Mar. 2023, doi: 10.1002/asi.24750.
[8] H. Alkaissi and S. I. McFarlane, "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing," *Cureus*, vol. 15, no. 2, Feb. 2023, doi: 10.7759/cureus.35179.
[9] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT," *Innovations in Education and Teaching International*, vol. 61, no. 2, pp. 228–239, 2024, doi: 10.1080/14703297.2023.2190148.
[10] A. Fortino and Z. Yang, "Evaluating Large Language Model Accuracy in Structured Academic Settings: Three Case Studies," in *2024 IEEE Integrated STEM Education Conference (ISEC)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ISEC61299.2024.10665280.
[11] W. H. Walters and E. I. Wilder, "Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-41032-5.
[12] J. Mugaanyi, L. Cai, S. Cheng, C. Lu, and J. Huang, "Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study," *J. Med. Internet Res.*, vol. 26, no. 1, Jan. 2024, doi: 10.2196/52935.
[13] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2312.10997
[14] C. Zhou *et al.*, "LIMA: Less Is More for Alignment," May 2023, [Online]. Available: http://arxiv.org/abs/2305.11206
[15] 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 Clinic Proceedings: Digital Health*, vol. 1, no. 3, pp. 226–234, Mar. 2023, doi: 10.1016/j.mcpdig.2023.05.004.
[16] David Baidoo-Anu and Leticia Owusu Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, Aug. 2023, doi: 10.61969/jai.1337500.
[17] C. Niu *et al.*, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre F. T. Martins, and Vivek Srikumar, Eds., Association for Computational Linguistics (ACL), Aug. 2024, pp. 10862–10878. doi: 10.18653/v1/2024.acl-long.585.
[18] M. Hosseini and S. P. J. M. Horbach, "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review," *Res. Integr. Peer Rev.*, vol. 8, no. 1, May 2023, doi: 10.1186/s41073-023-00133-5.
[19] 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.
[20] M. Perkins, "Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond," *Journal of University Teaching and Learning Practice*, vol. 20, no. 2, 2023, doi: 10.53761/1.20.02.07.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 323
This work is licensed under a Creative Commons Attribution 4.0 International License.