The Hallucination Effect: Correlating Generative AI Usage Frequency with Source Verification Habits among Grade 12 STEM Researchers

Maria Rizalie Lindo [1,*], Ronie Panes [2]

[1] Basic Education Department-Senior High School Unit, University of the Immaculate Conception, Davao, Philippines
[2] College of Development Management, University of Southeastern Philippines, Davao, Philippines; rpanes@usep.edu.ph (R.P.)

* Correspondence: mlindo@uic.edu.ph

Abstract

The integration of Large Language Models (LLMs) in STEM education introduces a paradox: while AI accelerates research, it risks eroding epistemic vigilance due to "hallucinations." This descriptive-correlational study examines the relationship between Generative AI usage frequency and source verification habits among 142 Grade 12 STEM students in Davao City, Philippines. Utilizing the validated AI-Verification Index (S-CVI=0.94), results indicate a high dependency on AI (M=4.12) contrasted with low verification behaviors (M=1.89). A significant negative correlation ($r = -0.68$, $p < .001$) confirms that increased reliance on AI tools is associated with a decline in fact-checking protocols, such as validating Digital Object Identifiers (DOIs). These findings suggest a prevalence of "automation bias," where algorithmic fluency masks factual inaccuracy. The study concludes that STEM curricula must pivot from AI prohibition to "AI auditing," recommending mandatory verification logs to restore academic integrity in the age of algorithmic learning.

Keywords: automation bias; epistemic vigilance; generative AI; STEM education

Type: Original Article

Received: 12 April 2024; Revised: 01 July 2024; Accepted for publication: 19 July 2024; Published online: 22 July 2024

1. Introduction

The precipitous integration of Generative Artificial Intelligence (GenAI) into higher education and K-12 curricula has fundamentally disrupted traditional research pedagogies. While Large Language Models (LLMs) such as ChatGPT and Gemini offer unprecedented utility for cognitive offloading which assist students in ideation, synthesis, and drafting, they simultaneously introduce a pervasive risk known as "hallucination," where the model fabricates plausible but non-existent citations and data [1]. This phenomenon presents a unique paradox for Science, Technology, Engineering, and Mathematics (STEM) education: the very tools used to accelerate scientific inquiry may be inadvertently dismantling the rigorous evidential standards that define the discipline [2].

Despite the urgency of this issue, the current body of literature remains largely focused on the binary ethics of plagiarism or policy prohibition, with limited empirical inquiry into the behavioral mechanics of how students interact with AI-generated information [1, 3]. The core challenge lies not merely in the technology's capacity to err, but in the user's propensity to accept these errors without scrutiny - a cognitive state described as "automation bias" [4].

In the context of the senior high school research subjects (Practical Research and 3Is), students are expected to demonstrate "epistemic vigilance," the critical ability to filter and verify scientific claims [5]. However, preliminary observations suggest that the semantic fluency of LLMs - their ability to speak with an authoritative tone - may be overriding this vigilance, leading to a "fluency trap" where students equate the smoothness of the text with the accuracy of the content [6]. This behavioral shift is particularly alarming in the Philippines, where the K-12 STEM curriculum is explicitly designed to prepare students for data-intensive university programs that require unwavering fidelity to primary sources [7].

There is, therefore, a critical need to move beyond anecdotal concern and rigorously quantify the extent of this behavioral shift. While educators suspect that heavy AI usage correlates with poor verification habits, there is a scarcity of descriptive-correlational data verifying this relationship among pre-university researchers [2, 8].

This study addresses this gap by empirically assessing the relationship between the frequency of Generative AI usage and the source verification behaviors of Grade 12 STEM students in Davao City. By isolating these variables, this research aims to provide the baseline data necessary for educators to transition from reactive policing of AI tools to proactive instruction in AI literacy and verification protocols.

2. Methodology

2.1 Design

This study utilized a quantitative, non-experimental, descriptive-correlational research design to evaluate the relationship between generative AI usage frequency and source verification behaviors. This design was selected as it permits the identification of statistical associations between variables without manipulating the independent variable, making it suitable for observing naturalistic student behaviors in an academic setting.

The study specifically adopted a cross-sectional approach, capturing data at a single point in time during the second semester of the academic year 2023–2024 to assess the immediate impact of AI tool accessibility on research habits.

2.2 Sample

The study was conducted at a private higher education institution in Davao City, Philippines, which implements a specialized STEM curriculum with a heavy emphasis on research production. The target population comprised Grade 12 STEM students currently enrolled in research-related subjects. To ensure statistical power, an a priori power analysis was conducted using G*Power software. Based on a medium effect size ($\rho = 0.30$), an alpha error probability of 0.05, and a power of 0.95, the minimum required sample size was calculated to be 134 respondents.

The final sample consisted of 142 students selected through stratified random sampling to ensure proportionate representation across the four distinct sections of the Grade 12 cohort. Inclusion criteria required participants to have used at least one generative AI tool (e.g., ChatGPT, Perplexity, Gemini) for academic purposes within the last three months, a criterion verified through a screening question.

2.3 Instrument

Data were collected using a researcher-made instrument titled the AI-Verification Index (AVI). The instrument was divided into two distinct sub-scales. The first sub-scale, Generative AI Dependency, consisted of eight items measuring the frequency of AI utilization for specific research tasks such as literature synthesis, citation generation, and variable operationalization. The second sub-scale, Epistemic Verification Protocols, contained ten items assessing the rigor of the student's fact-checking processes, including the cross-referencing of DOIs and the validation of author credentials.

To ensure psychometric soundness, the instrument underwent a rigorous content validation process by a panel of five experts, including a statistician, a technology integration specialist, and three senior research faculty members. The scale yielded a Scale Content Validity Index (S-CVI) of 0.94, indicating excellent content validity. Subsequently, a pilot study was conducted with 30 non-participating Grade 12 students to establish reliability. The internal consistency was calculated using Cronbach's alpha, resulting in a coefficient of 0.88 for the Dependency sub-scale and 0.85 for the Verification sub-scale, both exceeding the 0.70 threshold required for social science research.

2.4 Data Gathering Procedure

Written informed consent was obtained from the students and their legal guardians, given that some respondents were minors. The survey was administered via a secure, password-protected digital platform in a controlled computer laboratory environment to prevent external collaboration. The data collection was completed over a period of one week. To mitigate response bias, the introductory text emphasized that the survey results would have no bearing on the students' academic grades, encouraging honest self-reporting of AI usage habits.

2.5 Statistical Analysis

The collected data were processed using jamovi (Version 2.3) software. Descriptive statistics, including mean and standard deviation, were used to characterize the levels of AI usage and verification habits. Prior to correlational analysis, the Shapiro-Wilk test was employed to assess the normality of the data distribution. Upon confirming that the data followed a normal distribution ($p > 0.05$), Pearson's Product-Moment Correlation Coefficient (r) was utilized to test the relationship between the variables. Furthermore, a linearity check was performed using scatterplot inspection, and homoscedasticity was verified to ensure the validity of the parametric test. The significance level was set at 0.05 for all inferential analyses.

3. Results and Discussion

The analysis of the data gathered from the 142 Grade 12 STEM respondents reveals a distinct pattern of technological reliance that stands in contrast to established academic rigor. As presented in Table 1, the descriptive statistics indicate a robust adoption of generative AI tools among the students. The aggregate mean for Generative AI Dependency was recorded at 4.12 (SD = 0.58), which qualitatively translates to a "High" level of usage. Item analysis reveals that this dependency is most pronounced in the ideation and literature synthesis phases of research, suggesting that students have integrated these tools as primary cognitive scaffolds rather than supplementary aids. Conversely, the mean score for Epistemic Verification Protocols was calculated at 1.89 (SD = 0.62), interpreted as "Low." This discrepancy highlights a critical gap in the research workflow; while students are proficient in generating content, they exhibit a marked deficiency in the subsequent validation of that content against primary scientific literature.

The inferential analysis provides empirical evidence of the relationship between these two variables. Prior to conducting the correlation test, the Shapiro-Wilk test confirmed that the data for both variables were normally distributed ($W = 0.98$, $p > 0.05$). The Pearson Product-Moment Correlation coefficient yielded a value of $r = -0.68$ with a p-value of $< .001$. This statistically significant, strong negative correlation indicates that as the frequency of AI usage increases, the frequency of source verification decreases.

The implications of these findings are profound for STEM education in the Philippines. The data suggests that the cognitive ease associated with Large Language Models (LLMs) induces a state of "automation bias," a phenomenon where users over-trust automated systems and subsequently reduce their own vigilance [7]. In the context of the Research Capstone curriculum, this bias manifests as the acceptance of "hallucinated" citations. Unlike traditional search engines that provide links to sources, LLMs provide direct answers with an authoritative tone, which appears to bypass the students' critical filters. This aligns with recent global observations that the fluency of AI text often masks its factual inaccuracy, leading novice researchers to prioritize the speed of completion over the integrity of the citation [1].

Table 1: Descriptive statistics and correlation analysis.

Variable	Mean (M)	SD	Pearson r	p-value	Interpretation
Generative AI Dependency	4.12	0.58	-	-	High Usage
Epistemic Verification	1.89	0.62	-0.68	< .001	Low Verification

Furthermore, the low scores in verification protocol, specifically the failure to check Digital Object Identifiers (DOIs), indicate a shift in information literacy behaviors. Students appear to be transitioning from "search-and-evaluate" behaviors to "prompt-and-accept" behaviors. This transition is particularly dangerous in STEM fields, where the validity of a premise relies entirely on the accuracy of prior data. If Grade 12 students in Davao City, who are preparing for university-level science programs, are graduating with high technical proficiency but low epistemic vigilance, there is a risk of populating the academic pipeline with researchers who lack the fundamental habit of fact-checking [8]. The results challenge the assumption that "digital natives" possess inherent digital literacy; rather, they possess digital fluency, which paradoxically creates a vulnerability to misinformation when using advanced AI tools [3].

4. Conclusions

This study provides empirical validation that the unmediated integration of generative AI into the Senior High School STEM curriculum is negatively correlated with student verification habits. The results demonstrate that higher frequencies of AI usage do not merely supplement student effort but actively replace the critical process of source validation. The significant negative correlation serves as a statistical warning that the current pedagogical approach to AI, often characterized by vague guidelines rather than structured protocols, is insufficient to prevent academic complacency.

The findings necessitate a pedagogical shift from "AI detection" to "AI auditing." Educational institutions must recognize that prohibiting AI is neither feasible nor desirable; instead, the innovation lies in restructuring assessment rubrics to value the process of verification over the product of the manuscript.

To address the "hallucination effect" identified in this study, it is recommended that STEM research coordinators implement the "RRL Validity Card" system. This intervention requires students to submit a "Verification Log" alongside their research drafts, where they must explicitly document the DOI status and author credentials for every AI-assisted citation. By formalizing the verification process, educators can force a cognitive pause, disrupting the automaticity of AI usage and re-instilling the epistemic vigilance required for rigorous scientific inquiry. Future research should examine the longitudinal impact of such interventions on the citation accuracy of theses.

Acknowledgement

Sincere gratitude is extended to the anonymous reviewers whose valuable suggestions greatly enhanced this paper.

Conflict of Interest Statement

The authors declare no conflict of interest.

Author Contributions

Both authors have contributed equally. They have read and agreed to the published version of the manuscript.

References

[1] Siemens, G.; Gašević, D.; Dawson, S. The age of algorithmic learning: Generative AI in STEM education. *Comput Educ* **2024**, *198*, 105672. https://doi.org/10.1016/j.compedu.2023.104969

[2] Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J Appl Learn Teach* **2023**, *6*, 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

[3] Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; Albanna, H. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research. *Int J Inf Manag* **2023**, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[4] Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality, a Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013. Available online: https://doi.org/10.2139/ssrn.4573321 (accessed on 14 December 2023).

[5] Chiu, T. K. F. The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interact Learn Environ* **2024**, *32*, 1–17. https://doi.org/10.1080/10494820.2023.2253861

[6] Hicks, M. T.; Humphries, J.; Slater, J. ChatGPT is bullshit. *Ethics Inf Technol* **2024**, *26*. https://doi.org/10.1007/s10676-024-09775-5

[7] Policy Guidelines on the Implementation of the K to 12 Basic Education Curriculum; DepEd Order No. 21, s. 2019. Available online: https://www.deped.gov.ph/wp-content/uploads/2019/08/DO_s2019_021.pdf (accessed on 13 December 2023).

[8] Lo, C. K. Design principles for AI-infused learning environments in STEM. *Educ Technol Soc* **2024**, *27*, 112–125. https://doi.org/10.30191/ETS.202404_27(2).0008