IOCES 2026 Conference
The 1st International Online Conference on Education Sciences
15–17 June 2026 | Online
MDPI

An Evaluation Of Artificial Hallucination Rates In Large Language Models For Educational Use

Cheng Keat TAN[1], Yin Ni Annie NG[1], Qing Hao NG[2], Seh Yi Joseph TAN[3]
[1]School of Applied Science, Nanyang Polytechnic [2]Medical Affairs, Lucence Diagnostics Pte. Ltd. [3]GSK Asia House

INTRODUCTION & AIM

Artificial hallucination, termed as the "*generation of plausible but factually incorrect information*" is a complication in the use of Large Language Models (LLMs). Inappropriate reliance on LLMs, when used as an educational tool, may undermine critical thinking and risk distorting academic understanding, particularly among foundational learners.

The study **systematically evaluate and compare the extent of hallucination rates** in four publicly available LLMs.- ChatGPT 4o, Microsoft Copilot, Google Gemini 1.5, and Claude Sonnet 3.5 - when used for educational purposes.

METHOD

1. A set of 50 multiple-choice questions was developed and validated by subject experts to ensure alignment with Core Concepts in Pharmacology and sound question design.
2. Four LLMs were queried using a standardized, direct prompting protocol that required them to provide the correct MCQ answers, detailed reasoning, and supporting evidence.
3. Two independent subject experts evaluated each model's response for hallucination in the following areas:
   - **Accuracy** The correctness of LLM's MCQ responses.
   - **Reasoning** The logical integrity of the feedback.
   - **Evidence** The relevance of the provided citations.
4. Chi-square and Fisher-Freeman-Halton tests were employed to measure performance differences and identify statistically significant trends across LLMs.

RESULTS

Accuracy
The correctness of LLM's MCQ responses.

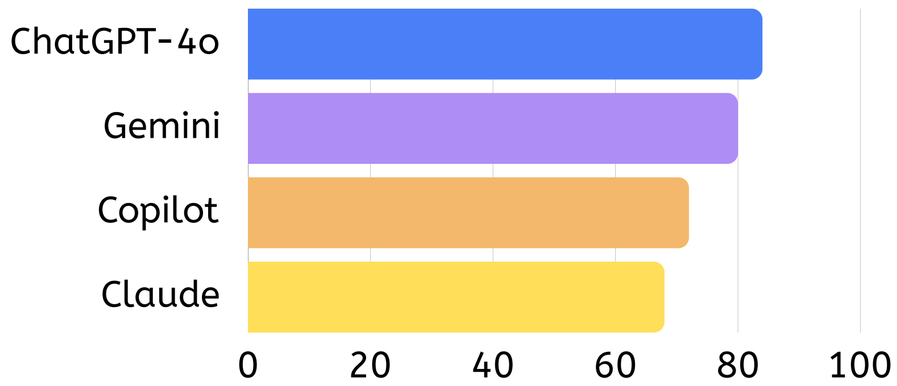

Figure 1: ChatGPT-4o recorded the highest accuracy rate (84%) in answering MCQs, with Gemini (80%), Copilot (72%) and Claude (68%) following in descending order. There is no statistical significance across the accuracy of the LLMs (p = 0.223).

Reasoning
The logical integrity of the feedback

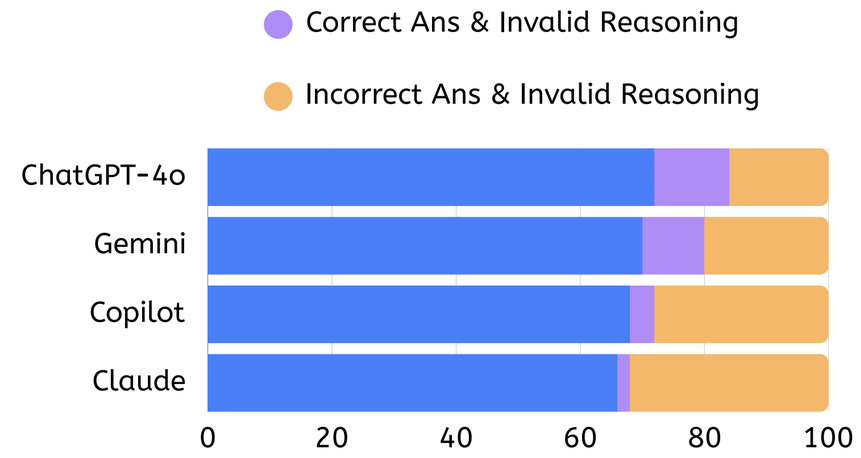

Figure 2: ChatGPT-4o attained the highest validity score of 72%. In contrast, hallucination rates - mainfesting as both incorrect answer and invalid reasoning - were descriptively higher in Gemini, Copilot and Claude. There is no statistical significant difference (p = 0.220).

Evidence
The reliability of the provided citations.

	ChatGPT-4o	Copilot	Gemini	Claude
Valid Reference [n (%)]	3 (3.0)	38 (44.7)	1 (3.1)	0 (0)
Invalid Reference [n (%)]	96 (97.0)	47 (55.3)	31 (96.9)	69 (100)
Total	99	85	32	69

Table 1: Copilot cited the largest proportion of citations that is reliable (i.e.: accessible and supports the explanations). The remaining demonstrated a pervasive susceptibility to citation-based hallucinations, with invalid references constituting between 96.9 - 100% of the total bibliographic output.

DISCUSSIONS

- Artificial hallucination is an **inherent and pervasive limitation shared across all evaluated LLMs**, regardless of the specific model.
- The generation of "synthetic knowledge" via inaccurate answers, invalid feedbacks and citations **threatens foundational understanding and evidence-based academic standards**.
- These systemic risks **necessitate a "human-in-the-loop" strategy** to ensure the responsible and verified integration of AI in higher education.

CONCLUSION

Artificial hallucination, pervasive across all LLMs studied, creates misconceptions and weakens critical thinking. Manifesting as incorrect answers or unreliable factual supports, a human-in-the-loop oversight is necessary to minimise the hallucination rates in the education use of LLMs.

REFERENCES

Choi, W. (2023). Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: A study using MCQs. BMC Medical Education, 23(1), 864. https://doi.org/10.1186/s12909-023-04832-x
Guilding, C., White, P. J., Cunningham, M., et al. (2024). Defining and unpacking the core concepts of pharmacology: A global initiative [Published correction appears in British Journal of Pharmacology, 181(7), 1150]. British Journal of Pharmacology, 181(3), 375-392. https://doi.org/10.1111/bph.16222
Pelánek, R (2025). Adaptive Learning is Hard: Challenges, Nuances, and Trade-offs in Modeling. International Journal of Artificial Intelligence in Education, 35, 304–329. https://doi.org/10.1007/s40593-024-00400-
Rahman, S. S., Islam, M. A., Alam, M. M., Zeba, M., Rahman, M. A., Chowa, S. S., Raiaan, M. A. K., & Azam, S. (2026). Hallucination to truth: A review of fact-checking and factuality evaluation in large language models. Artificial Intelligence Review, 59(2), Article 70. https://doi.org/10.1007/s10462-025-11454-w
Tacettin, C. M. (2021). Competency-based education: Theory and practice. Psycho-Educational Research Reviews, 10(3), 67-95. https://doi.org/10.52963/PERR_Biruni_V10.N3.06
Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv. https://doi.org/10.48550/arXiv.2401.11817