When LLMs Hallucinate: Examining the Effects of Erroneous Feedback in Math Tutoring Systems

Marlene Steinbach
Leibniz-Institute for Science and Mathematics Education, Kiel
steinbach@leibniz-ipn.de

Shreya Bhandari
University of California, Berkeley
School of Education
shreya.bhandari@berkeley.edu

Jennifer Meyer
Leibniz-Institute for Science and Mathematics Education, Kiel
University of Vienna
Centre for Teacher Education
jmeyer@leibniz-ipn.de

Zachary A. Pardos
University of California, Berkeley
School of Education
pardos@berkeley.edu

ABSTRACT

This extended abstract summarizes a full paper published in the proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25).

Feedback can be a powerful tool to support the learning process by providing information on how to solve the task, and Large Language Models (LLMs) offer great potential to produce elaborated feedback content for Intelligent Tutoring Systems (ITS) effectively and efficiently. However, LLMs are prone to producing incorrect information ("hallucinations") that introduce a challenge when integrating LLM-generated feedback into ITS because they may discourage or even mislead learners. To examine the trade-offs between LLM-generated feedback that contains errors due to hallucinations and no feedback, we investigated the effects of erroneous LLM-generated feedback on learning gain, confusion, time on task, perception of usefulness, and perception of feedback accuracy on twelve mathematics problem-solving tasks within an open-source, ITS-based adaptive learning platform. In our preregistered randomized controlled study, we assigned $N = 252$ learners to one of four conditions, receiving feedback on twelve math problems covering three learning objectives. Participants received LLM-generated feedback containing errors due to hallucinations in 1) 0% (i.e., fully accurate; "0% hallucinations"), 2) 50% (half of the feedback instances contained errors; "50% hallucinations"), 3) 100% of the instances (all feedback instances contained errors; "100% hallucinations"), or 4) no feedback (control condition). Prior knowledge was included as a moderating variable. Per lesson, two identical questions were used for a pre- and post-test to measure learning gain as the difference between average post-test and pre-test scores. The pre-test scores were used as the indicator of prior knowledge. Statistically significant learning gains from pre- to post-test were observed for learners receiving LLM-generated feedback with 0% and 100% hallucinations. The two-way ANOVAs indicated that, compared to receiving no feedback, learners showed significantly higher learning gain after receiving feedback with 0% or 100% hallucinations and significantly higher time on task and confusion after receiving feedback with 100% hallucinations. Learners in the 100% hallucinations condition perceived the feedback as more useful than the 0% and the 50% hallucinations conditions. Learners receiving more hallucinations perceived the feedback as less accurate. Prior knowledge moderated the effects only for perceived accuracy, with high-prior knowledge learners being better at identifying erroneous feedback as inaccurate. The pattern of results suggests that learning still occurred in the 100% hallucination condition because learners detected inaccuracies and spent more time on the feedback, potentially leading to deeper cognitive engagement with the feedback. Unlike the observed learning gain, perceived feedback usefulness declined with more hallucinations, emphasizing the benefit of empirical studies over subjective impressions. Future research should explore long-term effects, differentiate error types, and examine impacts across various learners and domains to guide responsible LLM feedback implementation in education.

Keywords

LLM hallucination, learning gain, feedback, adaptive tutoring, RCT

For citation, please refer to the full paper:
Marlene Steinbach, Shreya Bhandari, Jennifer Meyer, and Zachary A. Pardos. 2025. When LLMs Hallucinate: Examining the Effects of Erroneous Feedback in Math Tutoring Systems. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25), July 21–23, 2025, Palermo, Italy.* ACM, New York, NY, USA, 11 pages.
https://doi.org/10.1145/3698205.3729555

Marlene Steinbach, Shreya Bhandari, Jennifer Meyer, and Zach Pardos. When LLMs Hallucinate: Examining the Effects of Erroneous Feedback in Math Tutoring Systems. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 663–663. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870127