Evaluating Engineering Students' Detection of Hallucinations in Large Language Model Solutions

Zachary Deutsch[1]; Siohban Oca, PhD[1]

[1]Department of Mechanical Engineering, Duke University

Introduction

- Engineering students increasingly use LLMs like ChatGPT for coursework and problem solving [1], [2].
- LLMs can produce confident but incorrect responses, known as **hallucinations** [1].
- This is especially concerning in engineering, where small errors in equations, assumptions, or diagrams can change the final answer [4]–[7].
- Prior studies show that students know LLMs can be unreliable, but they may still struggle to detect specific errors [2], [3].
- In engineering tasks, LLMs often make errors in force resolution, circuit reasoning, arithmetic, and multi-step problem solving [4]–[7].
- These issues are more common in diagram-based or open-ended problems [4], [6], [7].
- This study examines **whether engineering students can identify hallucinations in GPT-generated Mechanics and Electricity and Magnetism (E&M) solutions**.
- This study also explores how confidence, prior coursework, and AI-use frequency relate to hallucination detection accuracy.
- **This study has a two-part design: (1) model benchmarking and (2) student hallucination detection**.

Methods: Benchmarking

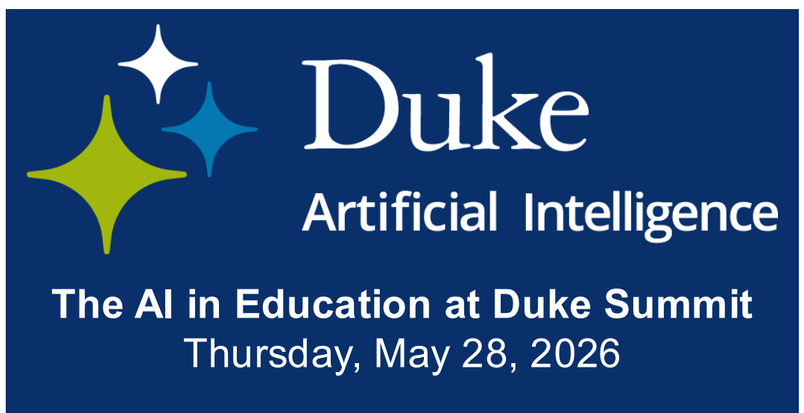
- GPT-5.1 was tested through the OpenAI API for reproducibility.
- Each question was shown to GPT-5.1 as a screenshot with 5 answer choices.
- GPT-5.1 completed 100 independent trials for each question
- Temperature was set to 1.0, same as ChatGPT
- GPT-5.1 was limited to selecting one multiple-choice answer lettered A–E.
- Accuracy was calculated by measuring how often GPT-5.1 selected the correct answer across trials.

Mechanics

- 20 questions were selected from the Force Concept Inventory, testing Newtonian physics.

Electricity and Magnetism

- 26 questions were selected from the Conceptual Survey of Electricity and Magnetism.



Results: Benchmarking

Mechanics

Questions (n)	Trials per Question	Total Trials	Per-Trial Accuracy
20	100	2,000	76.25%

- 11 questions were answered correctly 100% of the time, while 3 questions were never answered correctly.
- The results show three patterns: stable correct performance, stable confident errors, and variable responses with uncertainty across answer choices.

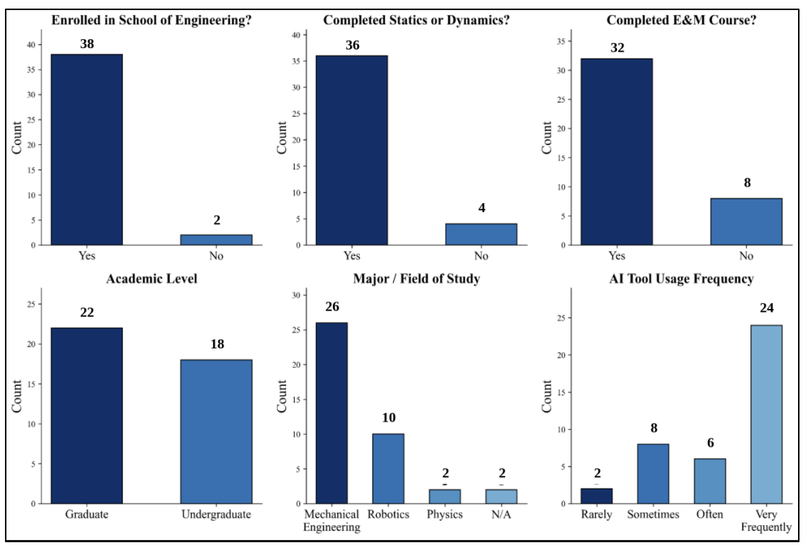
Electricity and Magnetism

Questions (n)	Trials per Question	Total Trials	Per-Trial Accuracy
26	100	2,600	57.12%

- 11 questions were always correct, while 2 questions were never correct.
- E&M accuracy was lower than mechanics, with more systematic errors and variable answer patterns.

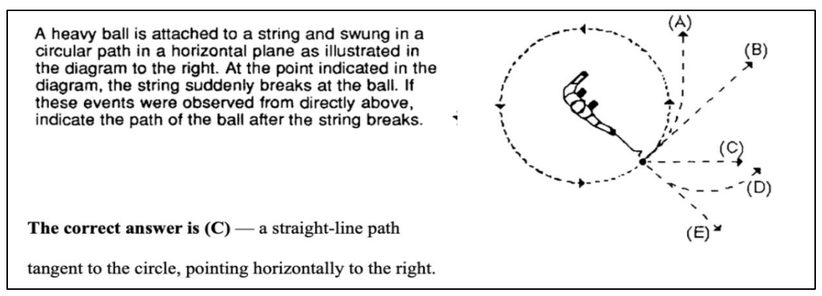
Methods: Detection

- Survey measured how engineering students evaluate LLM-generated solutions.



Survey Participant Demographics

- Students evaluated two mechanics and two electricity and magnetism problems with GPT-5.1-generated solutions.
- The four tasks included 2 correct solutions and 2 hallucinated solutions.
- For each task, students judged whether the solution was correct and rated their confidence.



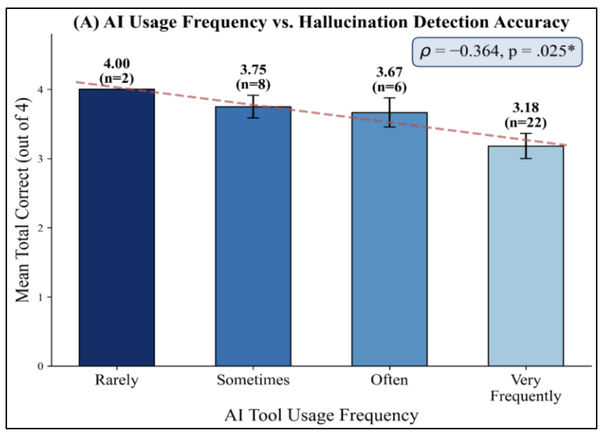
Example of hallucination solution: The GPT-5.1 solution incorrectly selected C, while the correct answer is B.

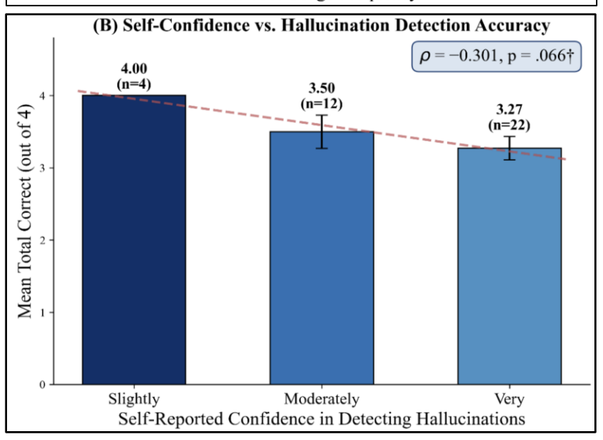
Results: Detection

- 38 engineering students completed the survey

Subject	Question	Ground Truth	Correct Responses	Incorrect Responses
Mechanics	Q1	Hallucination	36 (94.7%)	2 (5.3%)
Mechanics	Q2	Correct	34 (89.5%)	4 (10.5%)
E&M	Q1	Hallucination	34 (89.5%)	4 (10.5%)
E&M	Q2	Correct	26 (68.4%)	12 (31.6%)

- Higher AI-use frequency was significantly associated with **lower detection accuracy.**
- Higher self-reported confidence also showed a negative trend with accuracy.





Conclusion

- Students were more likely to **reject correct solutions** th an accept hallucinated ones
- Incorrect judgments were often made with **medium to high confidence**.
- GPT-5.1 showed systematic hallucination patterns on some conceptual questions, and some students still failed to identify those repeated error types.
- Higher AI-use frequency and higher self-reported confidence were associated with **lower hallucination detection accuracy**.

References

[1] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, Art. no. 102274, 2023.
[2] J. Lederman, "Students say generative AI has mixed effects on learning and critical thinking," Inside Higher Ed, Jan. 2025. [Online].
[3] S. Kabir, D. N. Udo-Imeh, B. Kou, and T. Zhang, "Is Stack Overflow obsolete? An empirical study of the characteristics of ChatGPT answers to Stack Overflow questions," in Proc. CHI Conf. Human Factors Comput. Syst. (CHI '24), Honolulu, HI, USA, May 2024, pp. 935:1–935:17.
[4] A. Patel, R. Singh, and M. Zhang, "Evaluating GPT-4 on undergraduate statics problems: Accuracy, hallucinations, and prompt sensitivity," arXiv:2502.00562, 2025.
[5] E. Reganova and P. Steinbach, "Testing uncertainty of large language models for physics knowledge and reasoning," arXiv:2411.14465, 2024.
[6] L. Chen, P. Morales, and S. Gupta, "Can ChatGPT pass an undergraduate control systems course? A performance and error analysis," arXiv:2503.05760, 2025.
[7] L. Skelić et al., "CIRCUIT: A benchmark for circuit interpretation and reasoning capabilities of LLMs," arXiv:2502.07980, Feb. 2025.

Duke | THOMAS LORD DEPARTMENT of MECHANICAL ENGINEERING & MATERIALS SCIENCE