ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #53628

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

Zachary Deutsch, Duke University

Zachary Deutsch is an undergraduate student at Duke University in the Thomas Lord Department of Mechanical Engineering and Materials Science, studying mathematics, mechanical engineering, and philosophy. His research interests include artificial intelligence and ethics.

Siobhan Oca, Duke University

Siobhan Rigby Oca is an assistant professor of the practice in the Thomas Lord Department of Mechanical Engineering and Materials Science at Duke University, NC, USA. She received her B.Sc. from Massachusetts Institute of Technology and Master in Translational Medicine from the Universities of California Berkeley and San Francisco. She completed her Ph.D. in Mechanical Engineering in 2022 from Duke University. Her research interests include applied medical robotics, human robot interaction, and robotics education.

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

Abstract

Large language models (LLMs) such as ChatGPT are widely used by engineering students to complete coursework, but current literature lacks empirical evidence on whether engineering students can recognize when LLMs generate confident but incorrect outputs (hallucinations). While LLMs can provide explanations, guidance, and feedback, they can also provide incorrect solutions, raising concerns for engineering education. This study addresses this gap through a sequential design consisting of (1) measurement of GPT-5.1 hallucination rates and illustrative examples using undergraduate mechanics and electricity and magnetism (E&M) problems, and (2) a survey of engineering students' ability to detect these hallucinations in LLM-generated solutions. To our knowledge, this is the first study focused specifically on engineering students' detection and confidence in LLM hallucinations in problem solving.

The authors generate benchmark data from GPT-5.1 to map failure modes of LLMs on quantitative tasks, quantify error frequencies, and develop examples of errors for use in the student survey. The authors survey engineering students' ability to detect LLM hallucinations in engineering problem solving. Specifically, the survey presents GPT-5.1-generated solutions to mechanics and E&M problems—some correct, some containing hallucinations—and asks students to accept or reject each solution and rate their confidence. Primary outcomes include hallucination detection accuracy and corresponding confidence, while secondary analyses examine associations with self-reported artificial intelligence (AI) use, prior coursework, and background characteristics. Results indicate that higher AI usage frequency is significantly correlated with lower hallucination detection accuracy, and higher self-reported confidence is marginally correlated with lower hallucination detection accuracy.

By combining empirical LLM-generated benchmark data and a student survey, this work (1) quantifies how reliably engineering students detect LLM hallucinations, (2) identifies which error types are most frequent, and (3) offers practical insights for developing hallucination detection strategies in engineering education.

Keywords: Large Language Models, AI Hallucination, AI Trust

I. Introduction

Large language models (LLMs) are widely used by engineering students for learning and completing coursework. While existing research examines student adoption of LLM use, significantly less is known about students' ability to recognize when LLMs produce confident but incorrect outputs, commonly referred to as hallucinations. This gap is particularly important in engineering education, where correctness is essential to learning outcomes.

A. Student Use of and Trust in LLMs

Survey-based studies have recently shown that students hold neutral attitudes toward LLM reliability. A survey of over 1,000 college students conducted in 2023 found that although LLM use was widespread, most students reported incomplete trust in LLM-generated information and described engaging in at least some form of verification [1]. Similarly, a national poll reported that students perceived LLMs as having mixed effects on their critical thinking: supporting efficiency while increasing the risk of unquestioned acceptance of answers [2]. These findings suggest that while students are aware of potential hallucinations, this awareness does not necessarily translate into systematic detection of hallucinations.

B. Student Detection of LLM-Generated Errors in Related Domains

Related results have been observed in programming contexts, which share important similarities with engineering problem solving. A study of ChatGPT responses to Stack Overflow programming questions found that more than half of the LLM's answers were incorrect, yet study participants failed to detect errors in nearly 40% of cases [3]. Furthermore, participants often preferred LLM-generated explanations to correct human answers due to their clarity and professional tone. These stylistic features appeared to mask hallucinations and in turn increase users' acceptance of incorrect solutions. Together, these studies indicate that students' ability to detect hallucinations depends not only on correctness but also on presentation and domain expertise.

C. Performance and Failure Modes of LLMs in Engineering Tasks

Recently, research has started to evaluate LLM performance on engineering problems. In mechanical engineering statics, a recent study shows that OpenAI's GPT-4 can achieve exam-level performance comparable to student averages [4]. However, even with relatively high accuracy, systematic failures persist. Common errors include incorrect resolution of angled forces, misclassification of truss members as being in tension or compression, and the introduction of nonexistent forces in free-body diagrams. These errors are particularly prevalent in multi-step problems and when visual information is included.

Quantitative variability has also been shown in otherwise straightforward calculations. Repeated trials of basic Newtonian mechanics problems revealed nontrivial deviations from correct values, with average errors exceeding 10% in some cases [5]. Some mistakes stem from arithmetic miscalculations while other mistakes stem from incorrect intermediate assumptions that the LLM propagates through subsequent steps. Such behavior is consistent with hallucination-like reasoning, where internally plausible but externally invalid steps are treated as true.

Similar behaviors appear in electrical engineering contexts. In an undergraduate control systems course, GPT-4 achieved passing grades on structured assessments but performed significantly worse on open-ended design projects [6]. The model frequently introduced overly specific numerical values and invoked advanced terminology beyond the course scope. In circuit analysis tasks, LLMs were found to perform well on simple configurations but struggled with multi-loop or diagram-dependent problems, often misapplying circuit laws or assuming incorrect component behavior [7].

D. Research Questions

Building on previous research in LLM output accuracy and engineering problem solving, this study is guided by three research questions: (1) How accurately do students detect when LLMs produce hallucinated solutions to engineering problems? (2) How do students' confidence in LLM solutions to engineering problems correlate with their accuracy when evaluating the solutions? and (3) Are detection accuracy and confidence associated with background variables such as course exposure and AI-use frequency?

II. Benchmarking GPT-5.1 on the Force Concept Inventory

A. Methods: Mechanics

The LLM tested was OpenAI's GPT-5.1, accessed via the OpenAI API. GPT-5.1 is the model underlying ChatGPT, but was accessed directly via the API for reproducibility. To benchmark GPT-5.1's performance solving mechanics problems, the Force Concept Inventory (FCI) was used, a multiple-choice test designed to assess understanding of the most basic concepts in Newtonian physics [8]. A subset of 20 questions was selected from the 30-question test. Each question was presented to GPT-5.1 through the API as a screenshot containing the prompt and answer choices. To estimate the consistency of the model's responses under repeated sampling, the authors generated 100 independent trials per question. Each trial consisted of a separate API call with identical prompts, no conversation history was provided, and temperature fixed at 1.0 to allow for variability in model responses. Model outputs were restricted to a single answer choice (A–E). Performance was evaluated using per-trial accuracy, defined as the proportion of the 100 trials (across all questions) selecting the correct answer.

B. Results: Mechanics

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

Table 1. Summary of GPT-5.1 performance on FCI [8]

The mechanics subset comprised 20 FCI questions (2,000 total trials). Across this subset, GPT-5.1 achieved an overall per-trial accuracy of 76.25% (1525/2000). Question-level performance was heterogeneous. Eleven questions were answered correctly with 100% accuracy (100/100 trials, indicating both high accuracy and high consistency. Three questions exhibited near-deterministic behavior with 98% accuracy (98/100 trials). In contrast, three questions were never answered correctly (0% accuracy). Two of these failures were deterministic: the model selected the same incorrect answer in all 100 trials. Other questions showed partial accuracy but substantial variability, including near-equal splits between two answer choices and multimodal distributions over three choices. Collectively, the FCI results indicate that under screenshot-based prompting, GPT-5.1 alternated between (1) highly stable, correct performance on many questions, (2) stable, confident failure modes on a smaller subset, and (3) some cases had substantial uncertainty and multimodal response patterns.

III. Benchmarking GPT-5.1 on the Conceptual Survey of Electricity and Magnetism

A. Methods: Electricity and Magnetism

To benchmark GPT-5.1's performance in introductory electricity and magnetism, the authors used the Conceptual Survey of Electricity and Magnetism (CSEM) [9]. A subset of 26 questions was selected from the 32-question test. Each question was presented under the same experimental setup as in the mechanics benchmark, including screenshot-based prompting, 100 independent trials per question, no conversation history, and temperature fixed at 1.0. Per-trial accuracy was computed.

B. Results: Electricity and Magnetism

Subject	Test	Questions (n)	Trials per Question	Total Trials	Per-Trial Accuracy
E&M	CSEM	26	100	2,600	57.12%

Table 2. Summary of GPT-5.1 performance on CSEM [9]

The E&M subset comprised 26 items (2,600 total trials). GPT-5.1 achieved an overall per-trial accuracy of 57.12% (1485/2600). Eleven questions were answered correctly with 100% accuracy (100/100 trials). However, the E&M benchmark also showed several questions with systematic error. Two questions had 0% accuracy (0/100 trials), with responses heavily concentrated on a single incorrect option (e.g., one question had 87% of trials selecting the same incorrect answer while never selecting the correct option). Some questions showed very low accuracy (e.g., 2–10%) with high concentration on an incorrect answer, and several questions exhibited multimodal behavior across two or more choices, resulting in 30–50% accuracy. Overall, the CSEM results indicate that under identical screenshot-based prompting, GPT-5.1 had lower accuracy on E&M questions than mechanics questions.

IV. Student Detection of LLM Hallucinations in Engineering Problem Solving

A. Methods

The authors assessed engineering students' ability to evaluate the correctness of LLM-generated solutions using a structured online survey composed of five sections: (1) consent, (2) eligibility screening, (3) background characteristics, (4) four problem-evaluation tasks, and (5) reflection on hallucination detection strategies and trust. The survey was developed by the authors specifically for this study. The survey was administered online and distributed to engineering students through flyers and university mailing lists. Participation was voluntary and occurred outside of formal classroom instruction. Eligibility criteria required participants to be at least 18 years old, enrolled in a school of engineering, and to report prior or current enrollment in mechanics (statics or dynamics) and/or electricity and magnetism (E&M). The survey included an explicit instruction not to use AI tools while completing the evaluation tasks. Participant demographics and background characteristics are summarized in Figure 1.

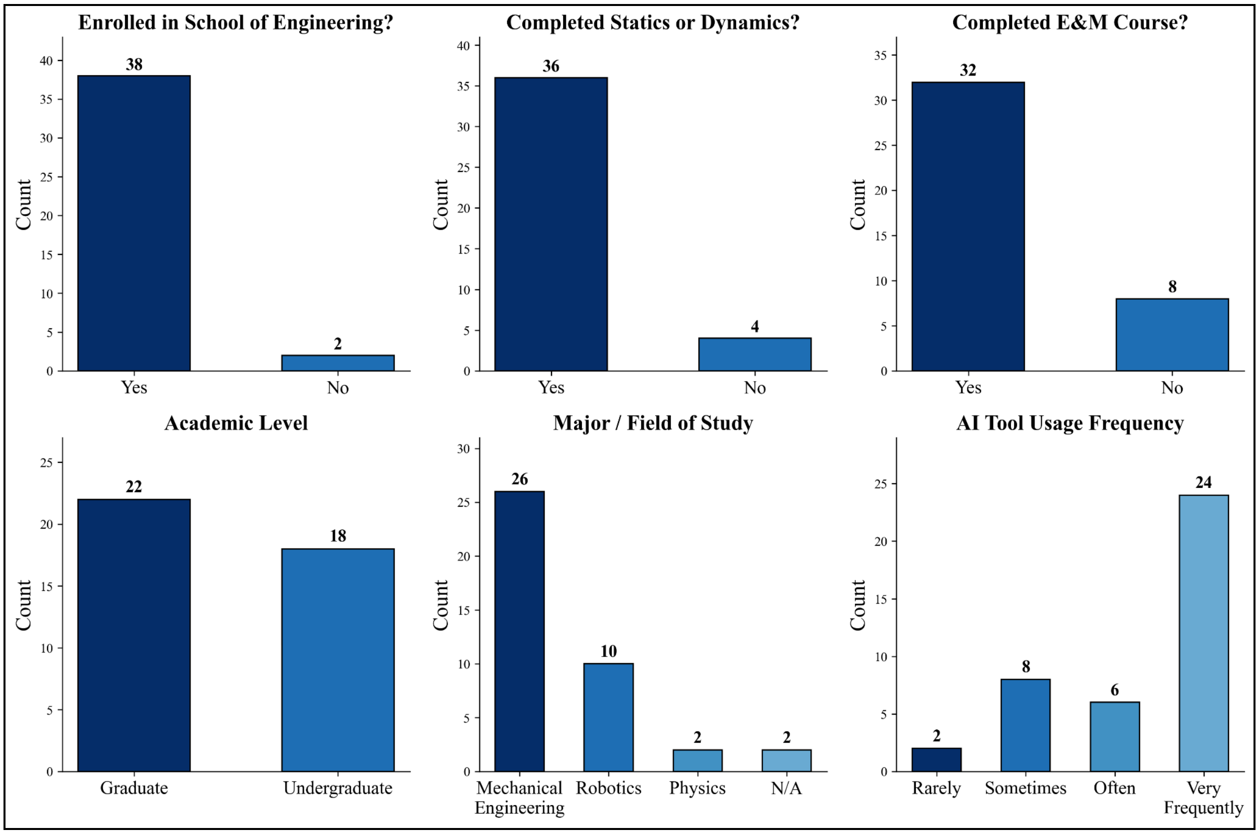
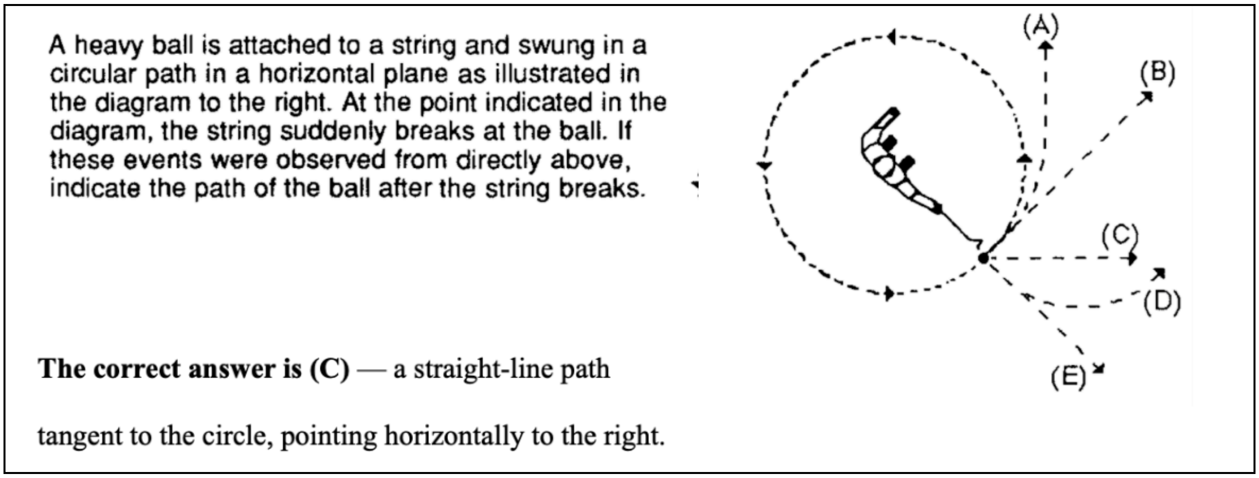


Figure 1. Participant demographics and background characteristics

To identify potential moderators of hallucination detection, participants reported academic level, major/field of study, AI tool use frequency, and self-confidence in detecting AI hallucinations. These items were included to support secondary analyses relating hallucination detection performance to academic background and prior AI-use.

Participants completed four problem-evaluation tasks: two mechanics and two E&M questions (see Figures 2–5). Each task consisted of (1) a mechanics or E&M question (presented as an image) and (2) a GPT-5.1-generated solution that selected an answer option and provided a brief rationale. Participants then indicated whether the solution was correct (Yes/No) and rated their confidence on a three-point scale. The four problem-evaluation tasks were selected to include both correct solutions and hallucinations.

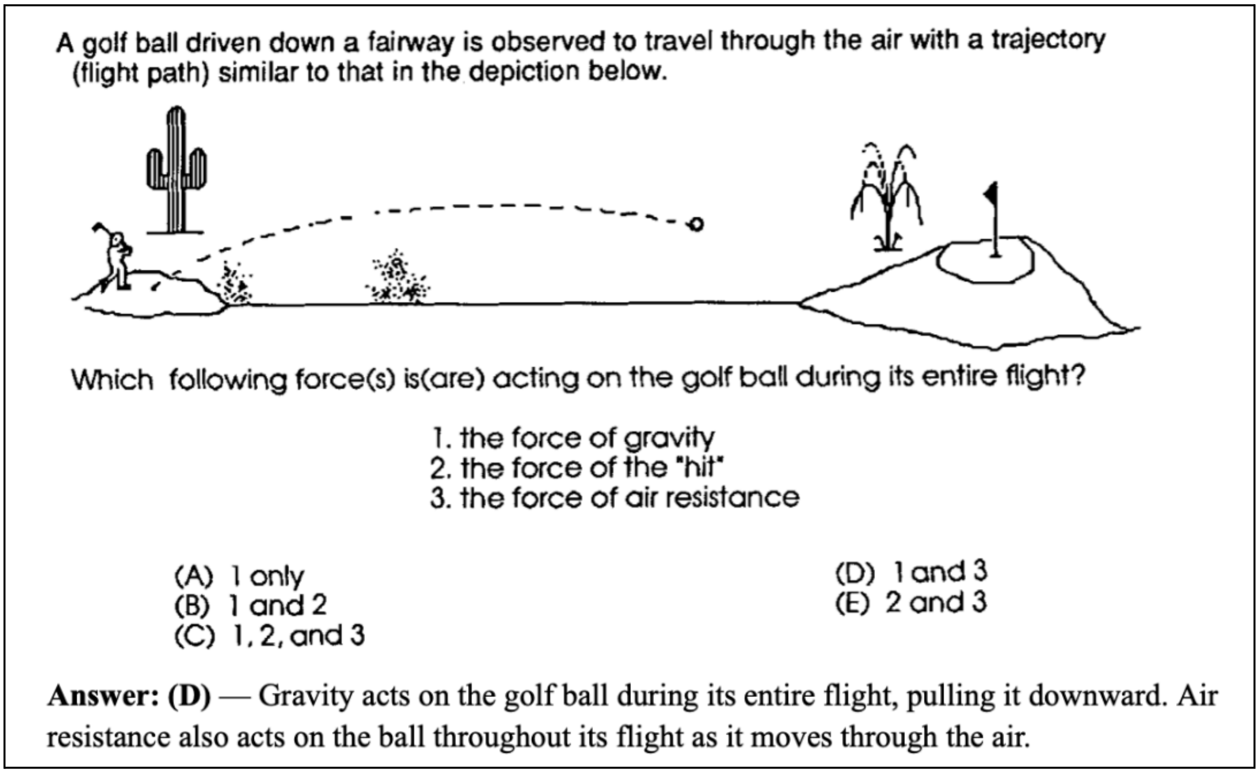


A heavy ball is attached to a string and swung in a circular path in a horizontal plane as illustrated in the diagram to the right. At the point indicated in the diagram, the string suddenly breaks at the ball. If these events were observed from directly above, indicate the path of the ball after the string breaks.

The correct answer is (C) — a straight-line path

tangent to the circle, pointing horizontally to the right.

Figure 2. Mechanics question 1 presented in the survey. The GPT-5.1 solution incorrectly selected C, while the correct answer is B.



A golf ball driven down a fairway is observed to travel through the air with a trajectory (flight path) similar to that in the depiction below.

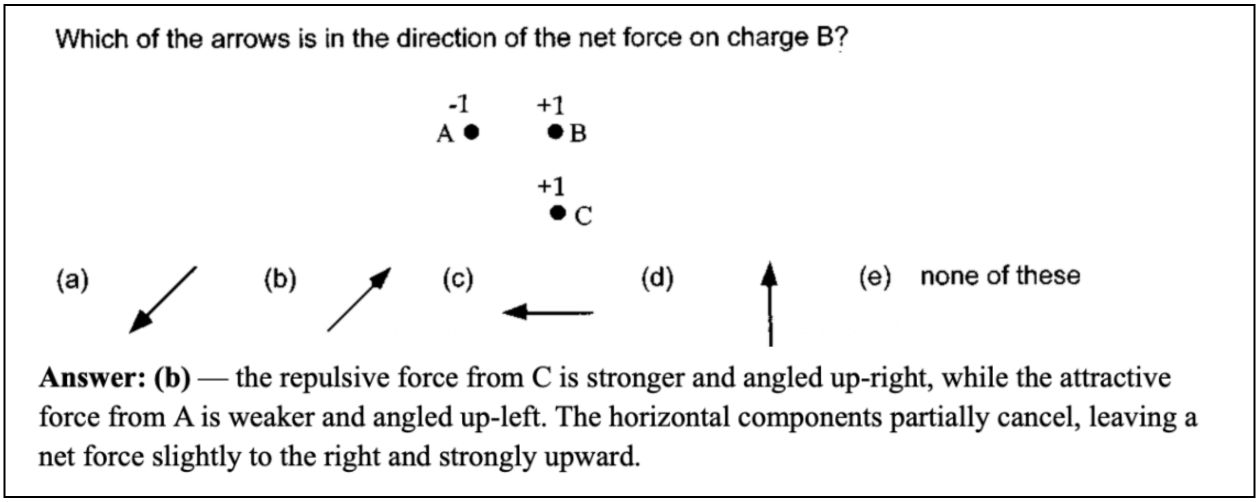
Which following force(s) is(are) acting on the golf ball during its entire flight?

1. the force of gravity
2. the force of the "hit"
3. the force of air resistance

(A) 1 only
(B) 1 and 2
(C) 1, 2, and 3
(D) 1 and 3
(E) 2 and 3

Answer: (D) — Gravity acts on the golf ball during its entire flight, pulling it downward. Air resistance also acts on the ball throughout its flight as it moves through the air.

Figure 3. Mechanics question 2 presented in the survey. The GPT-5.1 solution correctly selected D.

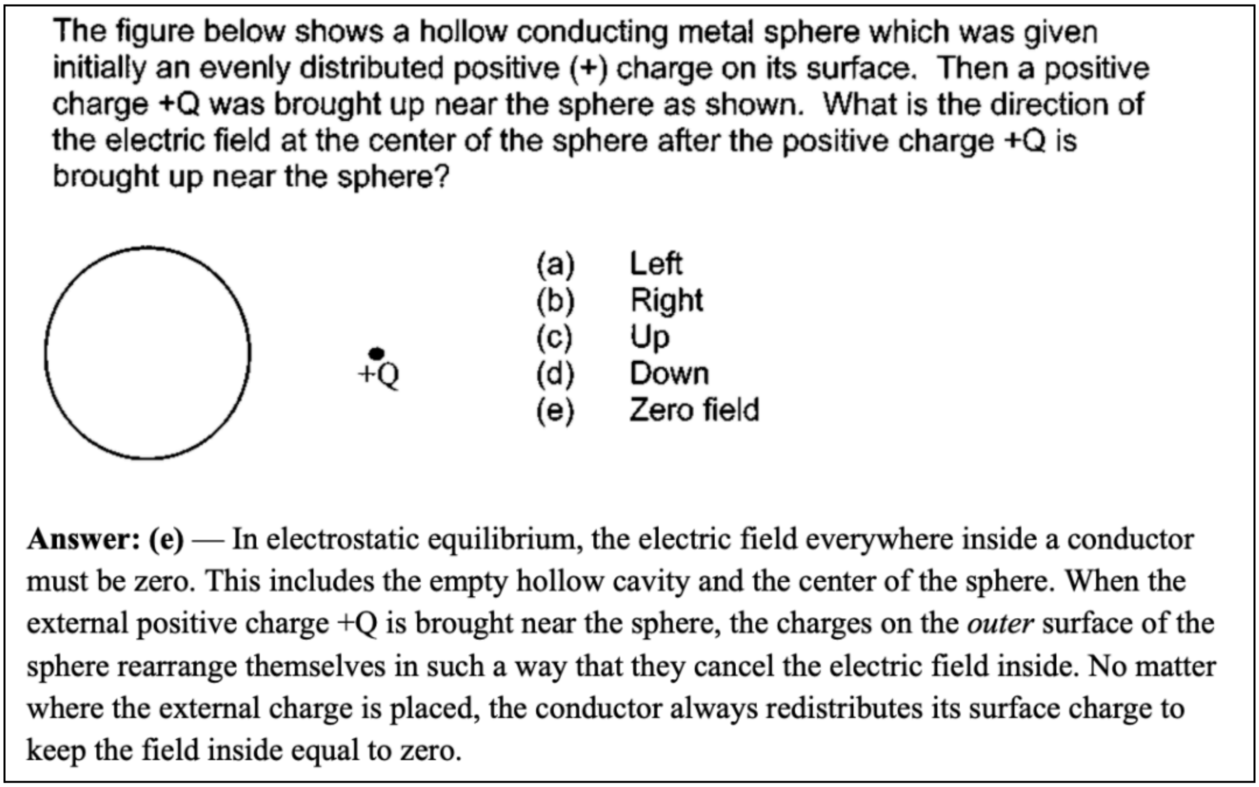


Which of the arrows is in the direction of the net force on charge B?

(a) ↙ (b) ↗ (c) ← (d) ↑ (e) none of these

Answer: (b) — the repulsive force from C is stronger and angled up-right, while the attractive force from A is weaker and angled up-left. The horizontal components partially cancel, leaving a net force slightly to the right and strongly upward.

Figure 4. E&M question 1 presented in the survey. The GPT-5.1 solution incorrectly selected B, while the correct answer is E.



The figure below shows a hollow conducting metal sphere which was given initially an evenly distributed positive (+) charge on its surface. Then a positive charge +Q was brought up near the sphere as shown. What is the direction of the electric field at the center of the sphere after the positive charge +Q is brought up near the sphere?

(a) Left
(b) Right
(c) Up
(d) Down
(e) Zero field

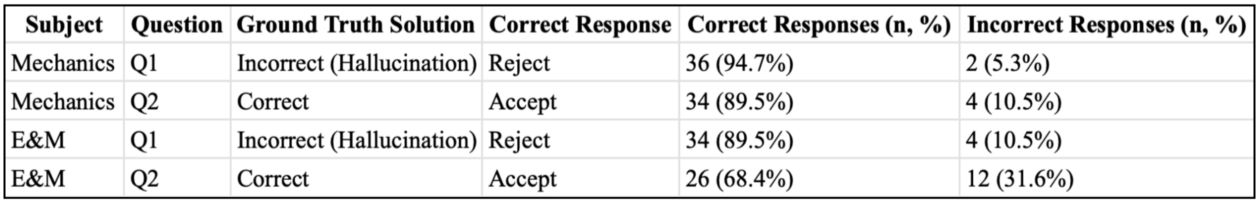
Answer: (e) — In electrostatic equilibrium, the electric field everywhere inside a conductor must be zero. This includes the empty hollow cavity and the center of the sphere. When the external positive charge +Q is brought near the sphere, the charges on the *outer* surface of the sphere rearrange themselves in such a way that they cancel the electric field inside. No matter where the external charge is placed, the conductor always redistributes its surface charge to keep the field inside equal to zero.

Figure 5. E&M question 2 presented in the survey. The GPT-5.1 solution correctly selected E.

The primary outcome was hallucination detection accuracy for each question. Secondary outcomes included participants' confidence ratings relative to whether their judgments were correct (e.g., overconfidence when accepting hallucinated solutions). Analyses also examined asymmetries in error type, separating (1) false acceptance of incorrect solutions from (2) false rejection of correct solutions, and examined associations with background variables, including AI-use frequency and self-reported hallucination detection confidence.

To contextualize quantitative outcomes, the survey concluded with an open-response item asking what cues or reasoning strategies participants used to evaluate LLM-generated solutions, followed by a question about general trust in AI tools for solving engineering problems.

B. Results



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

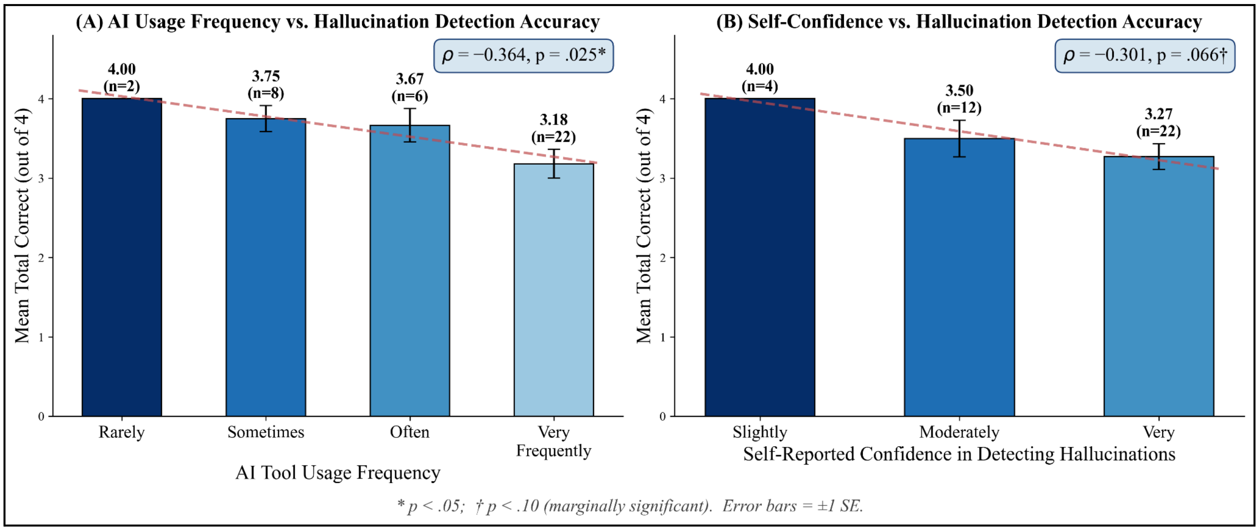
Table 3. Student evaluation of LLM solutions

A total of 38 eligible engineering students completed the entire survey. For each problem-evaluation task, participants indicated whether the GPT-5.1-generated solution was correct and rated their confidence. Two tasks presented hallucinated solutions (one mechanics, one E&M), and two presented correct solutions.

Performance differed depending on whether the solution was correct or hallucinated. For Mechanics Question 1, in which the GPT-5.1 solution incorrectly answered option C instead of the correct answer B, 36 of 38 participants (94.7%) correctly rejected the solution, while 2 participants (5.3%) incorrectly accepted it. A similar pattern was observed for E&M Question 1, where the GPT-5.1 solution selected B despite the correct answer being E: again 34 participants (89.5%) correctly rejected the solution, and 4 participants (10.5%) accepted the hallucinated solution. In contrast, detection accuracy was higher for tasks with correct solutions. For Mechanics Question 2, 34 of 38 participants (89.5%) correctly accepted the solution, with 4 participants (10.5%) rejecting it despite its correctness. For E&M Question 2, 26 participants (68.4%) correctly accepted the solution and 12 (31.6%) incorrectly rejected it. Taken together, these results indicate that participants were less accurate at rejecting hallucinated solutions than at accepting correct ones, with false acceptance rates of 7.9% for hallucinated answers compared to false rejection rates of 21.1% for correct answers.

Confidence ratings showed a difference between subjective confidence and objective correctness. Across all four questions, confidence distributions were similar: the majority of responses fell in the high confidence category, regardless of whether the judgment itself was correct. For hallucinated solutions (Mechanics Question 1 and E&M Question 1), some incorrect acceptances were accompanied by high confidence. In both cases, across the 6 instances in which participants accepted hallucinated solutions, several reported having high confidence, indicating overconfidence in hallucinated solutions. This pattern was also reflected in the confidence distribution of correct rejections, suggesting that confidence alone was not a reliable indicator of successful hallucination detection. Conversely, for correct solutions, incorrect rejections also occurred with medium to high confidence. Among the 4 participants who rejected the correct solution in Mechanics Question 2 and the 12 participants who rejected the correct solution in E&M Question 2, confidence ratings spanned from low to high confidence.

To analyze relationships between participants' background and hallucination detection performance, Spearman correlations were computed between AI usage frequency, self-reported confidence in detecting hallucinations, and total correct responses. AI usage frequency was significantly negatively associated with detection accuracy (ρ = -0.364, p = .025). Self-reported confidence was marginally negatively associated with detection accuracy (ρ = -0.301, p = .066). As shown in Figure 6, both relationships showed negative trends, indicating that higher AI usage and higher self-reported confidence are associated with lower hallucination detection accuracy.



** p < .05; † p < .10 (marginally significant). Error bars = ±1 SE.*

Figure 6. Relationship between AI usage, self-reported confidence in detecting hallucinations, and hallucination detection accuracy

V. Discussion

This study examined engineering students' ability to evaluate the correctness of LLM-generated solutions in mechanics and E&M. The results indicate that while students generally recognized hallucinated solutions, errors still occurred. Importantly, these errors were often accompanied by medium to high confidence. Participants were less accurate at accepting correct solutions than at rejecting hallucinated ones. The overall false acceptance rate for hallucinated items (7.9%) was lower than the false rejection rate for correct items (21.1%). This asymmetry suggests that, in this sample, students may have been somewhat cautious or skeptical when evaluating

LLM-generated responses, at times rejecting correct solutions. While this pattern differs from some prior findings in non-engineering domains, it nonetheless highlights that confidence and correctness were not consistently aligned.

From an educational standpoint, this pattern is concerning because false acceptance directly undermines learning: students who accept hallucinated reasoning may learn incorrect concepts. Confidence ratings were broadly similar across correct and incorrect judgments. In particular, the presence of highly confident incorrect acceptances suggests that LLM-generated explanations may produce an illusion of understanding.

In both the mechanics and E&M benchmarks, GPT-5.1 hallucinated on a subset of conceptual questions. Two hallucinated solutions were included in the student survey (one mechanics, one E&M). Both of these questions reflected a systematic failure mode observed under repeated sampling (mechanics Question 1 and E&M Question 1), where GPT-5.1 selected the same incorrect answer in the majority of trials. Despite this, a subset of students still failed to detect the error. Taken together, these findings suggest that current patterns of student interaction with LLMs pose nontrivial risks for engineering learning: the issue is not merely that LLMs make mistakes, but that students may fail to recognize them and do so with high confidence.

Several limitations should be acknowledged. First, the sample size was modest and drawn from a single institution, limiting generalizability. Second, the study relied on a small number of evaluation tasks. Third, while confidence ratings provide useful subjective insights, they cannot fully capture the reasoning processes underlying acceptance or rejection decisions.

VI. Conclusion

This study provides empirical evidence that engineering students do not reliably detect hallucinated solutions generated by LLMs in mechanics and E&M problems. Although a majority of participants correctly identified incorrect solutions, a nontrivial minority still accepted hallucinated solutions, and many of these acceptances were accompanied by medium to high confidence. In contrast, students were less accurate when evaluating correct solutions. Additionally, higher frequency of AI tool use and higher self-reported confidence were both associated with lower hallucination detection accuracy. This study reframes LLM hallucinations as not only a technical limitation of AI systems but also an educational challenge for students. As LLMs become increasingly embedded in learning and education, the ability to evaluate LLM-generated responses will be as important as the ability to generate solutions themselves. Future research should extend this work across institutions, problem types, and instructional contexts, and should investigate interventions that improve students' ability to detect and appropriately respond to LLM-generated errors.

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.

[8] D. Hestenes, M. Wells, and G. Swackhamer, “Force concept inventory,” The Physics Teacher, vol. 30, no. 3, pp. 141–158, Mar. 1992, doi: 10.1119/1.2343497.

[9] D. P. Maloney, T. L. O’Kuma, C. J. Hieggelke, and A. Van Heuvelen, “Surveying students’ conceptual knowledge of electricity and magnetism,” American Journal of Physics, vol. 69, no. S1, pp. S12–S23, 2001, doi: 10.1119/1.1371296.