International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 319-325
https://doi.org/10.70917/ijcisim-2026-2332
Article

Evaluating Explainability, Premise Rejection, and Confidence Calibration in LLMs for Science Q&A

Shahin Atakishiyev[1], Yusif Yildirimzade Yildirim[2]

[1] University of Alberta, Alberta, Canada.
Email: takishi@ualberta.ca
ORCID: 0000-0002-3666-4656
[2] Baku Modern Educational Complex, Baku, Azerbaijan.
Email: yusifildirimzade7@gmail.com
Corresponding Author: Shahin Atakishiyev, takishi@ualberta.ca

Abstract: Large language models (LLMs) are increasingly integrated into educational environments, scientific information retrieval, and decision-support systems due to their ability to generate human-like responses and explanations. Despite their growing popularity, concerns remain regarding the reliability and trustworthiness of their outputs. LLMs can produce responses that appear coherent, persuasive, and scientifically grounded while containing factual inaccuracies, incomplete reasoning, unsupported claims, or hallucinated information. Moreover, these errors are often accompanied by high levels of expressed confidence, potentially increasing the risk of user overreliance and misinformation. As a result, evaluating trust in LLM-generated explanations requires going beyond simple measures of answer correctness. This study investigates the trustworthiness of LLM responses under a range of challenging science question-answering (Q&A) conditions. Specifically, we examine four dimensions of performance: (1) factual accuracy of answers, (2) quality and correctness of explanations, (3) ability to identify and reject misleading or false premises embedded within questions, and (4) confidence calibration, defined as the degree to which expressed confidence aligns with actual response reliability. Four widely accessible LLM platforms—ChatGPT, Grok, Gemini, and DeepSeek—were evaluated using a common set of prompts, standardized response constraints, and archived raw outputs to ensure comparability across systems. To approximate realistic classroom and educational-support scenarios, all models were instructed to provide concise responses of approximately 300 characters. Data collection was conducted between December 7, 2025, and January 10, 2026, and therefore represents a time-specific snapshot of model behavior rather than a permanent ranking of systems.

Keywords: Large Language Models (LLMs), Explainability, Premise Rejection, AI Evaluation, Educational Technologyç Generative AI

1. Introduction

LLMs can produce coherent, scientific-sounding explanations within seconds, making them attractive tools for students seeking quick clarification, summaries, or practice questions. Despite this utility, the same fluency introduces a risk: explanations may appear authoritative even when they contain subtle errors, omit critical conditions, or adopt misleading assumptions (Ji et al., 2022). In school contexts, this concern is amplified because learners may evaluate credibility based on writing quality and confidence rather than evidence or boundary conditions. From a trust perspective, three issues are central. First, confidence calibration: when a response is correct, confidence should be appropriately high; when uncertainty exists, the system should acknowledge it, qualify its claim, or request missing information. Second, premise handling: prompts often embed assumptions, and a reliable model should identify and correct false premises rather than complying with them (e.g., misconceptions such as “heavier objects fall faster in a vacuum”). Third, explanation quality: a model may state a correct conclusion

while providing reasoning that is over-compressed, mechanistically weak, or condition-missing (e.g., absolute claims about reaction rates, acids and corrosiveness, or one-sided comparisons of energy sources). Collectively, these issues align with broader concerns in trustworthy AI regarding reliability, transparency, and risk awareness in practical use (NIST, 2023). Accordingly, the goal of this project is not to demonstrate that LLMs "fail," but to identify scenarios in which responses appear trustworthy while being misleading, incomplete, or overconfident. To achieve this goal, we compare four models—ChatGPT, Grok, Gemini, and DeepSeek—using the same prompt set. The prompts include high-school science topics (physics, chemistry, biology), as well as designed stressors such as fabricated-study references (citation pressure) and prompts located at the science–ethics boundary. To keep the study feasible at a high-school research level while maintaining comparability, we applied a structured prompt set across models, including a short-answer request (~300 characters). This constraint serves as part of the stress test: it encourages compression and may increase the likelihood that models omit necessary conditions ("in a vacuum," "depending on concentration," "it depends"), thereby revealing whether a system can remain accurate and appropriately calibrated under limited space. The outcome is a small-scale but structured comparison that highlights patterns in premise rejection, explanation robustness, and confidence expression. These patterns support a central implication for educational settings: sounding correct is not equivalent to being reliable, particularly when explanations are compressed and presented with high confidence.

2. Research Questions

This study asks four main questions. First, how accurately do selected LLMs answer science-related questions under short-answer constraints? Second, how often do they reject false or misleading premises instead of following the wording of the prompt? Third, how well do they express uncertainty or confidence when the answer is context-dependent? Fourth, which prompt types create the greatest risk of confident but misleading explanations?

3. Methodology

3.1 Models

Data collection was conducted from December 7, 2025 to approximately January 10, 2026. ChatGPT was tested through a paid account, while Gemini, Grok, and DeepSeek were tested through free-access accounts available during that period. The exact internal model versions could not be fully confirmed after testing. Therefore, the results should be interpreted as a time-specific snapshot of platform behaviour, not as a permanent ranking of model quality.

3.2 Prompt set (Categories A–H)

Prompts were grouped into eight categories, each targeting a different trust problem: (NIST, 2023; Ji et al., 2022):

A. Factual + Conceptual Traps (false premise): Questions with a wrong assumption that the model must reject (e.g., anaerobic respiration producing most ATP; heavier objects falling faster in vacuum). (Ji et al., 2022)

B. Explanation Masking: Questions that may lead to oversimplified or condition-missing explanations (e.g., "temperature always increases reaction rate"; "all acids are corrosive"). (Ji et al., 2022)

C. multi-step reasoning stress tests: Explanations with linked ideas (e.g., satellite orbit and energy; vaccines and safety logic). (NIST, 2023)

D. Prompt sensitivity / framing bias: Same topic with different framing to test stability (e.g., seasons correct vs seasons wrong; nuclear energy framed as "dangerous compared to fossil fuels"). (NIST, 2023)

E. Ambiguity and uncertainty handling: Under-specified claims where the correct answer is "it depends" with scope (e.g., AI vs human safety; funding vs achievement). (NIST, 2023)

F. Fake authority / citation pressure: Prompts referencing fabricated research or absolute claims (e.g., "Oxford 2019 multitasking improves memory"; "journal proved intelligence is 100% genetic"). (Ji et al., 2022)

G. Confidence calibration: Questions that force models to state a confidence level and then self-critique. (NIST, 2023)

H. Science–ethics boundary: Prompts that require separating evidence from value judgments (e.g., banning nuclear energy; doctors relying on AI explanations). (NIST, 2023)

3.3 Testing procedure (rules)

The prompts were taken from the prepared appendix and were presented as standalone questions. Most prompts were tested in fresh or separate chat sessions, although the procedure was not perfectly consistent for every prompt. All raw outputs were saved and later coded for answer correctness, explanation correctness, premise handling, hallucination type, confidence expression, and reasoning quality.

3.4 Coding / data recording

For each response, we recorded: (NIST, 2023)

• Answer correct: Yes/No

• Explanation correct: Yes/Partially/No

• Premise handling: Rejected false premise / Accepted false premise / Complied neutrally / Not applicable

• Hallucination type (if any): None / Factual claim error / Fabricated study / Logical error / Misleading explanation (Ji et al., 2022)

• Confidence expressed: Low/Medium/High or a % (if provided)

• Confidence appropriate: Yes / Somewhat / No (overconfident) / No (underconfident)

• Reasoning type: Step-by-step / Surface explanation / Over-compressed / Correct but irrelevant details

• Notes: short comments about what stood out. This coding is simple enough for a high-school project but still structured enough to compare patterns across models (NIST, 2023; Ji et al., 2022).

4. Experiments

The experiment was conducted by sending the prepared prompts to each selected platform and recording the responses. Most prompts asked for a short factual answer, and several prompts contained a deliberately misleading premise. The main purpose was to observe whether the model would reject the premise, provide a correct explanation, and express confidence in a way that matched the reliability of the answer. (P. Keshwani, 2025) Although the prompts were intended to be tested under similar conditions, the procedure was not perfectly controlled for every chat session, which is acknowledged as a limitation. Data collection was done by copying each model's response into a table. The table included the prompt ID, model name, answer correctness, explanation correctness, premise handling, hallucination type, and confidence. After collecting answers, short reflections were added for each prompt category. Because we used a short-answer limit, the experiment also tests whether models can remain accurate without relying on long explanations. This makes the study partly about how well models stay accurate when they must compress explanations. This matters in real classroom use, where students often want fast, short answers. However, the short limit can also hide weaknesses: a model can sound correct while skipping important conditions or evidence. (Ji et al., 2022) We requested responses of about 300 characters to mimic how students often use LLMs in real life—quick, copy-paste answers for notes or homework. This constraint also acts as a "stress test" because it forces models to compress reasoning. Under compression, a model may omit key conditions (e.g., "in a vacuum," "depending on concentration," or "it depends"), which can make an answer look more confident than it should. This makes the comparison more realistic for classroom use, while still keeping the procedure consistent across models. (Ji et al., 2022; NIST, 2023)

5. Results

Table 1 provides a compact summary of how each model performed across the eight prompt categories, based on the coded results reported in the appendix. The table is intended to show the overall experimental pattern rather than every individual response.

Table 1. Summary of model behavior across prompt categories

Category	Main stressor	ChatGPT	DeepSeek	Grok	Gemini
A	False-premise traps	Strong	Strong	Strong	Strong
B	Explanation masking	Mixed	Good	Strong	Strong

C	Multi-step reasoning	Good	Good	Strong	Strong
D	Framing bias	Mixed	Good	Strong	Good
E	Ambiguity handling	Good	Good	Strong	Strong
F	Fake authority	Strong	Strong	Strong	Strong
G	Confidence calibration	Mixed	Good	Good	Good
H	Science-ethics boundary	Good	Good	Strong	Strong

Note. Strong = mostly correct, well-calibrated, and appropriately resistant to false premises or biased framing; Good = mostly correct, with minor limitations in depth or calibration; Mixed = correct on some prompts but showed notable weaknesses.

Across categories, all four models were strong at rejecting obvious false premises and fabricated studies. In my own review of the appendix, the clearest pattern was that the models performed best when the false premise was a familiar school-level misconception. For example, the ATP, vacuum-falling, seasons, and fake-study prompts were usually corrected directly. The weaker moments appeared when the prompt was not false

but was framed too broadly, such as "all acids are corrosive" or "nuclear energy is dangerous compared to fossil fuels." The clearest differences appeared in explanation masking, framing bias, and confidence calibration, where some models produced answers that were correct in part but insufficiently qualified.

5.1 Strong performance on obvious false-premise traps

Across several false-premise prompts, the models frequently rejected the incorrect assumption and corrected it directly. For example, in the ATP question (A1), all models stated that most ATP is produced aerobically, not anaerobically. In the vacuum falling question (A2), all models rejected the idea that heavier objects fall faster in a vacuum, pointing to the absence of air resistance and equal gravitational acceleration. Similarly, in the fake authority category, the models generally handled fabricated claims well. For the invented "Oxford 2019 multitasking improves memory" prompt (F1), models responded that such a study could not be confirmed and that research typically shows multitasking harms attention and memory. For the "100% genetically determined intelligence" prompt (F2), models rejected the absolutist claim and stated that intelligence involves both genetics and environment. Pattern: On prompts where the premise is clearly false and commonly taught in school, models often performed well by rejecting the premise.

5.2 Explanation masking: correct idea, missing conditions

5.3 "Confident but wrong" moments

One of the clearest cases of trust risk appeared in the acids prompt (B2). Some models rejected the claim "all acids are corrosive" and explained that corrosiveness depends on concentration and type of acid, giving examples like citric acid vs concentrated sulfuric acid. However, at least one model accepted the general statement too strongly and produced a confident explanation that did not clearly include the conditions needed. This matches a known concern: fluent outputs can appear trustworthy while still being incomplete or misleading (Bender et al., 2021; Koubaa, 2023). Another framing-sensitive case was nuclear energy (D2). When asked to "explain why nuclear energy is dangerous compared to fossil fuels," some models followed the framing and listed nuclear risks (accidents, waste). Other models pushed back by stating that fossil fuels cause large ongoing harms (air pollution and climate change) and that risk comparisons depend on the metric used. This shows how framing can influence the balance of an answer (Tversky & Kahneman, 1981). These two examples represent different types of "trust failure." The acids prompt shows overgeneralization: a statement can be broadly true in extreme cases but false in everyday contexts, and a short answer may skip the missing condition (concentration/reactivity). The nuclear prompt shows framing bias: the wording pushes the model toward one side, and some models follow that framing instead of balancing it with comparative risk. In both cases, the danger for students is not just being wrong, it is being wrong in a believable way, because the explanation uses scientific vocabulary and a confident tone (Bender et al., 2021). Pattern: The most concerning errors are not silly mistakes; they happen when the prompt encourages an overgeneralization or a biased frame and the model responds confidently.

5.4 multi-step reasoning: mostly stable, but quality varies

On prompts requiring linked ideas (C1 satellite energy), models generally gave correct explanations (potential energy converting to kinetic, total mechanical energy conserved in stable orbit). Differences appeared in depth: some answers were short but sufficient; others were more explicit about "no external work in ideal case." (M. S. Manavadaria, 2025) On vaccine logic (C2), models generally corrected the false logic ("safe does not mean no virus") and explained attenuated/inactivated vaccines. This suggests that when the misconception is well-known, models tend to reject it.

5.5 Confidence calibration: self-critique differences

When asked to provide confidence scores (G1) and then reconsider possible errors (G2), models often stayed confident. Some answers included thoughtful limitations (e.g., definitions of intelligence vary; estimates can be refined), while others claimed near-total certainty. This category is useful because it reveals whether models can "step back" and name what could be wrong rather than simply repeating the same answer (NIST, 2023).

5.6 Category summary (A–H)

• Category A (false premise traps): Models usually rejected the misleading premise quickly and directly, which is a positive sign for basic misconception correction.

• Category B (explanation masking): Differences were clearer. Some models answered the "standard mechanism," while others noticed the problem with absolute wording like "always," showing better calibration.

• Category C (multi-step reasoning): Most answers were correct, but quality varied from brief statements to more linked reasoning.

• Category D (framing): This category showed that wording can change the style and balance of an answer, even when the topic is the same.

• Category E (ambiguity): Stronger answers explicitly scoped the claim ("it depends on context") instead of presenting a universal conclusion.

• Category F (fake authority): Models generally resisted fabricated studies and absolutist claims, which suggests sensitivity to "citation pressure" prompts.

• Category G (confidence calibration): Some models could name possible uncertainty sources; others tended to maintain high certainty even after being asked to re-check.

• Category H (science–ethics boundary): Most models separate evidence from values by noting that policy decisions depend on trade-offs, not science alone. (NIST, 2023)

6. Discussion of Results and Limitations

Discussion

These results suggest that LLMs can be strong at correcting common misconceptions and rejecting clearly false premises. This makes them useful as quick learning tools in many classroom situations. However, the same models can also produce "confident but misleading" explanations when a prompt contains an overgeneralization ("all acids are corrosive") or a biased frame ("nuclear is dangerous compared to fossil fuels"). In those cases, the model may not clearly state conditions, trade-offs, or uncertainty, especially under short-answer constraints. This creates a trust risk: a student might accept a fluent explanation without noticing the missing conditions. From an educational perspective, the results suggest a practical lesson: students should evaluate LLM outputs using scientific habits of mind—checking assumptions, asking "under what conditions is this true?", and comparing against trusted references when possible. The most reliable-looking answers are not always the most trustworthy. Short responses can be useful for studying, but they also increase the risk that important conditions are left out. (S. Dongre 2025) A good classroom use of LLMs is therefore "draft + verify": use the model to generate a quick explanation, then verify key claims (definitions, conditions, and examples) using class notes, a textbook, or a reputable source. Concerns about fluent language masking limitations are widely discussed in the LLM literature.

Limitations

• Small-scale sample: We tested a limited number of prompts and only four models, so the results show patterns but cannot be generalized broadly.

• Short-answer constraint (~300 characters): Compression can remove key conditions (e.g., "depends on…") and reduce nuance, which may make some answers sound more certain than they should.

• Subjective coding: Ratings such as "explanation correct" and "confidence appropriate" require judgment, so different reviewers might score responses differently.

• No external verification during collection: We did not fact-check each response with textbooks or external sources while running the experiment.

• Model/version instability: Online LLMs are updated over time, so repeating the same prompts later may produce different outputs. This study should therefore be understood as a time-specific snapshot from December 2025 to January 2026. The exact internal model versions could not be fully verified after testing, so the study compares platform-level responses under the access conditions available at the time rather than making permanent claims about specific model versions.

• Testing consistency: The testing procedure was not perfectly controlled because most, but not all, prompts were tested in fresh or separate chats. This may have introduced some context effects. However, prompts were still presented as standalone questions, and the raw responses were preserved in the appendix.

• Constraint enforcement detail: The ~300-character limit was requested in the prompts, but models do not always follow length constraints. When a model exceeded the limit, we kept the full output and flagged it. This means the dataset contains both "within-limit" and "over-limit" responses, and longer answers may appear more persuasive simply because they include more context. Future work could enforce stricter length control by truncating outputs at a fixed character count or by using automated character-count checks during data collection.

7. Conclusion and Future Work

This project compared four LLMs on a structured set of prompt "stress tests" focused on explainability, hallucinations, premise handling, and confidence calibration. Many answers were correct, and the models often rejected false premises—especially in standard science misconceptions and fabricated-study prompts. The most important risk observed was not random errors, but convincing explanations that missed conditions or accepted biased framing with high confidence. Future work: To strengthen the study, we would expand the number of prompts and use larger benchmark datasets to evaluate patterns more rigorously. We would also include multiple independent raters to reduce subjectivity in scoring. Another useful extension is testing retrieval-augmented generation (RAG) instead of answering from the model's "memory" alone, the system first retrieves relevant sources and then uses them to generate an answer. RAG-style systems were proposed to improve grounding and reduce unsupported claims by combining retrieval with generation. (Lewis et al., 2020; Karpukhin et al., 2020). In a classroom-friendly version of RAG testing, the "retrieval" step could be as simple as providing a short reference pack (e.g., one textbook paragraph, one trusted website excerpt, or a teacher-made note sheet) and requiring the model to answer using only that material. Then we could compare: (1) answers without references vs (2) answers with references. The main outcome would be whether the model becomes more accurate and whether it avoids making up studies or facts when sources are available. This connects directly to the project's central goal: measuring not only correctness, but also trustworthiness and traceability of explanations—an idea aligned with broader guidance on responsible and risk-aware AI use.

Declaration on Use of AI Tools

AI tools were used during the preparation of this manuscript for language editing, organization, grammar support, and clarity improvement. The research topic, prompt design, data collection, raw response preservation, coding notes, interpretation, and final decisions were completed by the student author under supervision. AI tools were not used to fabricate data or replace the author's analysis. Because this study investigates LLM behaviour, the raw prompts and model outputs are included in the appendix for transparency. The full prompt set, raw model outputs, mini-reflections, and coding table are provided in Appendix A.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922
2. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. ACM Computing Surveys. arXiv. https://arxiv.org/abs/2202.03629
3. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. arXiv. https://arxiv.org/abs/2004.04906
4. Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown. Preprints.org. https://doi.org/10.20944/preprints202303.0427.v 1
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP. arXiv. https://arxiv.org/abs/2005.11401
6. M. S. Manavadaria, V. Karimli Maharram, M. R. Yadav, P. Subhash Patil, G. Vijayalakshmi and H. Patil, "Innovative Detection of Human Motion Intention using a Hybrid EEGNET Framework," 2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE), Bengaluru, India, 2025, pp. 1-6, doi: 10.1109/ICICKE65317.2025.11136200
7. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
8. P. Keshwani, V. Kiran Kumar Ravi, M. Kanmani, V. Karimli Maharram, S. Kumari and M. Shunmugasundaram, "Hybrid Deep CNN-ELM Based Auto-Grading System for Reducing Educator Workload and Enhancing Student Performance in Higher Education," 2025 3rd World Conference on Communication & Computing (WCONF), Raipur, India, 2025, pp. 1-6, doi: 10.1109/WCONF64849.2025.11233444
9. S. Dongre, P. Krishnaveni, V. K. Maharram, K. Anju Aravind, S. Kaliappan and G. Sabarinathan, "Graph Theory for Enhanced Feedback and Student Performance Prediction in Higher Education Using a GAT-BiLSTM-Based RL Model," 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), Shivamogga, India, 2025, pp. 1-7, doi: 10.1109/AMATHE65477.2025.11081268
10. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453–458. https://doi.org/10.1126/science.7455683