International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-5, Issue-6, November-December 2026
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Hallucinations in Educational LLMS: A course-grounded Framework for Reliability and Pedagogical Risk

Shamsu Abdullahi[1], Kamaluddeen Usman Danyaro[1], Abdallah Saleh Abdallah[2], Mazlina Mohamad[3], and Rasheed Abubakar Rasheed[4]

[1]Department of Computing, Faculty of Management, Science, and Computing, Universiti Teknologi PETRONAS, Perak 32610, Malaysia
[2]Integrated Islamic School Shah Alam, Seksyen 20, Shah Alam, Selangor, Malaysia
[3]Faculty of Civil Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia
[4]School of Computer Science, Taylor's University, Subang Jaya, Malaysia

Received: 14 Oct 2026, Received in revised form: 10 Nov 2026, Accepted: 15 Nov 2026, Available online: 20 Nov 2026
©2026 The Author(s). Published by IJTLE. This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Abstract

Educational Large Language Models (LLMs) outputs may contain factual errors, claims unsupported by course materials, citation defects, or cross-turn contradictions. These properties are related but not interchangeable: a correct claim may be absent from course materials, and course materials may themselves be mistaken. This paper presents a selective descriptive synthesis and proposes a course-grounded reporting framework. It organises five review dimensions into claim-level judgments (factual status, course support, and evidence integrity), a dialogue-level judgment (consistency), and a contextual interpretation of defect-related pedagogical risk. Learner outcomes are separate evidence requiring direct measures; causal attribution requires designs linking exposure to tutor outputs with measured outcomes. An eight-step procedure operationalises the review, and a cross-classification table makes factual status and course support explicit. The illustrative risk rubric is unvalidated and does not cover every source of instructional harm. Retrieval, abstention, and human oversight are safeguards to evaluate, not guarantees.

***Keywords*— large language models, educational AI, hallucination, learning outcomes, retrieval-augmented generation, reliability**

I. INTRODUCTION

Large language models (LLMs) explain concepts, generate practice, and provide feedback in education [2]. Fluent responses may conceal factual errors, unsupported claims, or mismatches to a learner's level and instructional goal; because hallucination is defined and measured in different ways, these failures should be specified rather than conflated [1], [2]. Factuality, course alignment, and instructional value are related but distinct properties. LLM applications have also been reviewed in specialised industrial settings where domain-specific adaptation and hallucination risks are noted [3]. Such findings do not establish the reliability of educational tutors but illustrate why evaluation should be situated in its intended domain. Existing studies examine different units: randomised tutor-package comparisons measure learner outcomes; feedback audits measure faithfulness to supplied student work; retrieval pilots distinguish incorrect answers from answers outside course context; and benchmarks

assess misconception handling or feedback diagnosis [4]–[10]. These results are not interchangeable: students were not randomized to exposure to specific errors, and benchmark performance does not establish learner effects. Section 4 synthesizes the selected evidence and its interpretive limits. This paper integrates five review dimensions with three evidence levels in an operational reporting workflow. It is a descriptive synthesis, not a new detector, retrieval method, or validated scale. The workflow makes these distinctions visible when researchers describe outputs and compare tutor systems. It separates judgments available from response text or dialogue from claims requiring direct learner measures in a specific evaluation context. The paper's organisation is summarized in Fig. 1.

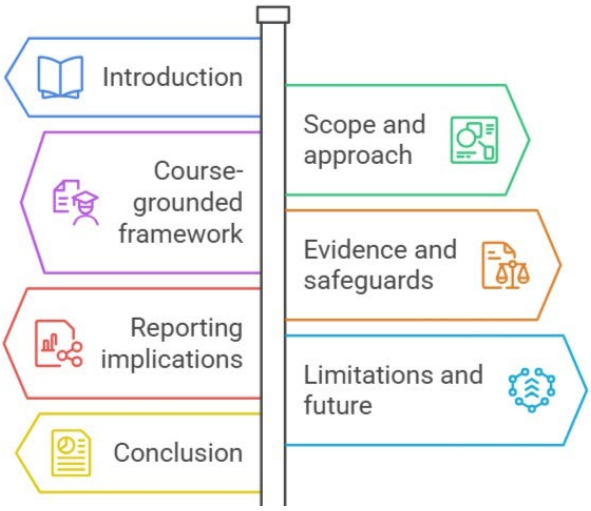


Fig. 1. Organisation and progression of the paper

Table 1: Course-Grounded Review Dimensions for Educational LLM Outputs

Dimension	Operational question	Evidence basis
Factual status	Is the proposition correct, incorrect, or unresolved against suitable references?	Claim-level review
Course support	Do the available course materials support, contradict, or not address it? Absence is not falsity	Claim-level review
Evidence integrity	Does each cited source exist, is it accurately represented, and does it support the claim it accompanies?	Claim-level review
Dialogue consistency	Does the dialogue contradict verified information or earlier turns, and is a contradiction corrected or repeated?	Dialogue-level review
Pedagogical risk	Given the objective, learner, and dialogue, what instructional consequence could an identified defect plausibly have?	Interpretation of claim, dialogue, and context; not direct learner evidence

II. SCOPE AND DESCRIPTIVE APPROACH

This is a focused, illustrative narrative synthesis, not a systematic review or formal critical appraisal. Unlike systematic reviews that document database searches and screening, such as the review by Nura and Abdullahi [11], selection here was purposive rather than exhaustive, with studies representing distinct evidence types: hallucination definitions, learner-outcome experiments, feedback-faithfulness and retrieval audits, and misconception/feedback benchmarks. Their populations, tasks, models, annotation units, and outcomes differ; reported rates and effects are therefore not pooled or treated as comparable. In this paper, a factual error is a claim judged incorrect against suitable references; unsupported means not substantiated by the course materials available in the interaction, not necessarily false; and pedagogical risk is a proposed judgment of plausible instructional consequence, not an observed learner outcome.

Table 2: Cross-Classification of Factual Status and Course Support

Factual status	Supported by course materials	Contradicted by course materials	Not addressed by course materials
Correct	The proposition is correct and supported by the course pack.	Suitable external references support the proposition, but the course pack conflicts, perhaps because it is mistaken or simplified.	A factually correct extension is absent from the course pack; absence does not imply falsity.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.3

Incorrect	The course pack itself supports a proposition contradicted by suitable references.	The proposition is incorrect and the course pack also rejects it.	Suitable references show the proposition is incorrect; the course pack is silent.
Unresolved	The course pack supports the proposition, but external evidence is insufficient or conflicting.	The course pack rejects the proposition, but external evidence is insufficient or conflicting	Neither the available course materials nor suitable references resolve the proposition.

III. A COURSE-GROUNDED FRAMEWORK

The cited approaches address different targets: hallucination research examines factuality and source support, feedback audits assess faithfulness to student work, and retrieval-grounded tutoring tests the use of supplied course materials [1], [6], [7]. The framework does not replace these methods; it brings their complementary perspectives together with cross-turn consistency, contextual interpretation of risk, and separate learner-level evidence. Table 1 summarizes the five dimensions and their evidence bases. Learner-level outcomes are not an additional review dimension; they require separate measures, as shown in Table 4. The two axes in Table 2 use different standards: factual status is judged against suitable references, whereas course support is judged against the materials available in the interaction. Nemeth et al. report that 16.5% of sampled answers were ´ outside supplied context; that label alone does not establish whether those answers were factually correct [7].

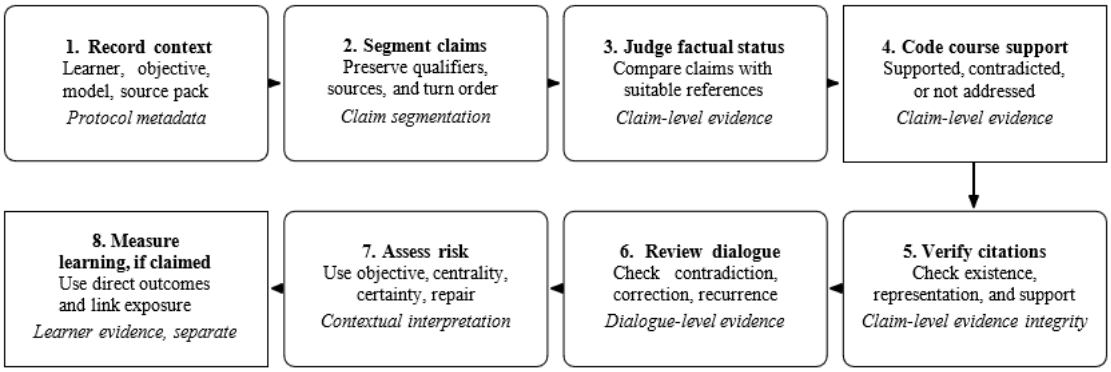


Fig. 2. Eight-step review workflow with evidence levels and the separate contextual risk interpretation.

3.1 Operational Review Procedure

Apply the dimensions in sequence and retain the transcript and source records so that each judgment is auditable. Figure 2 maps each step to its evidence level or interpretive layer; learner-level measurement is separate from output review.

A. Record the subject, learner level, learning objective, model/version, prompt, course source pack, and whether the exchange is single- or multi-turn.
B. Segment tutor responses into atomic, checkable claims; preserve qualifiers, uncertainty, cited sources, and turn order. Define units by task: equations and procedural transformations may require linked subclaims; essay or language feedback can be segmented into claims about the learner's text, diagnosis, and proposed revision; code feedback can be segmented into behavioral claims, tests, and suggested changes. Preserve dependencies rather than splitting content in ways that change its meaning.
C. Judge each claim's factual status against suitable references, recording the reference and an unresolved label when evidence conflicts or is insufficient.
D. Independently code course support as supported, contradicted, or not addressed by the materials available in that interaction.
E. Verify every citation against its source and the specific claim: existence, accurate representation, and support.
F. Review the full dialogue for contradictions with prior verified information, correction, recurrence, and persistence; retain the initial error and any repair as separate events.

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.3

G. Apply the proposed risk rubric in Table 3 using the objective, learner, error centrality, expressed certainty, repetition, and repair. Report the rationale, not just the category.
H. If making claims about learning, analyse direct learner level measures separately and link exposure to tutor outputs only when the study design permits it.

3.2 Factual Status and Course Support

Judge factual status against appropriate references, not solely against the course pack. Course materials may be incomplete or locally simplified; record disagreement with authoritative references rather than treating the course pack as a universal truth standard.

3.3 Evidence and Dialogue

A citation's presence does not establish source existence, accurate representation, or claim support. Retrieval makes materials available but does not ensure faithful use [6], [12]. Dialogue review captures reversal, repeated error after correction, and disregard of verified prior information—patterns missed by isolated answer scoring.

3.4 Pedagogical Risk

The rubric rates plausible instructional consequences of identifiable output defects relative to the learning objective; it is not a measure of overall tutor safety or an observed learner outcome. Assign high risk only when the defect is central to the objective, could materially misdirect learning or assessment, and is left uncorrected or persists after correction. Assign low risk when a defect is peripheral, unlikely to obstruct the objective, and promptly corrected or not repeated. Use moderate when the defect is relevant but bounded or repaired, or when indicators conflict. Certainty and repetition inform the judgment but do not determine it alone; record the rationale. Table 3 gives provisional examples. Boundary examples illustrate the thresholds: a confidently stated but peripheral error made once remains low when it does not obstruct the objective; an incorrect step in a required solution that is repaired before the learner acts is moderate rather than high; and a false prerequisite repeated after correction is high because it is central and persistent. These examples are prompts for consistent coding, not validated thresholds.

Table 3: Proposed Illustrative Rubric for Potential Pedagogical Risk

Level	Illustrative Criteria	Hypothetical Example
Low	Peripheral claim; unlikely to obstruct the objective; readily corrected and not repeated.	A minor imprecision in an optional essay comment, promptly clarified.
Moderate	Relevant to the objective or a plausible misconception; could misdirect practice but is bounded or corrected in the exchange.	One incorrect grammar correction in a target sentence, amended before the learner revises.
High	Central prerequisite or procedure; could materially misdirect learning or assessment; expressed with high certainty or repeated, and left uncorrected.	A false core language rule repeated as the basis for required exercises.

IV. EVIDENCE, INTERPRETATION, AND SAFEGUARDS

This section brings together the framework's evidence levels, selected educational findings, and the limits of common safeguards.

4.1 Evidence Levels and Causal Inference

The five review dimensions are assessed at two output analysis units: factual status, course support, and evidence integrity are claim-level judgments; dialogue consistency is assessed from the transcript. Pedagogical risk is a contextual interpretation of those findings, not an evidence level. Learner outcomes are separate from output review, are not a review dimension, and require direct measures such as objective-aligned assessments, transfer, delayed retention, confidence, or error detection. Table 4 distinguishes these layers. A faulty tutor claim does not show that a learner noticed or adopted it, and a score difference alone does not identify which output caused it. Random assignment to tutor packages can estimate the effect of assignment to each tested package relative to its control condition in that setting [4]; it cannot isolate the effects of individual package features or any particular erroneous claim. In observational analyses, error exposure may be associated with learner-selected questions and other factors related to performance. Such associations should not be described as causal unless the study design supports that inference.

4.2 Selected Educational Evidence

Table 5 contrasts selected studies by task, outcome, and evidence type. Educational reviews also place hallucination among wider concerns such as instructional quality, equity, AI literacy, and governance [2].

Table 5: Selected Studies Comparison

Study	Design and task	Principal finding	Relevance and limits
Bastani et al. [4]	Field experiment with nearly 1,000 students at a Turkish high school; GPT Base, teacher-designed GPT Tutor, and a no-AI control	Practice performance was 48% higher with GPT Base and 127% higher with GPT Tutor than control. On the unaided exam, GPT Base scored 17% lower than control; GPT Tutor did not differ significantly from control.	Learner-level outcomes and instructional context. Tutor packages were randomized; performance associations do not establish error causation.
Pardos and Bhandari [5]	Randomized mathematics-help study (274 participants) across four topics; AI help, human-authored help, or no help.	ChatGPT help produced statistically significant gains versus no help; its gains were not statistically distinguishable from human-authored help. ChatGPT-generated help failed quality checks on 32% of problems.	Claim quality and learner outcomes. A nonsignificant AI–human contrast is not a formal equivalence result.
Jia et al. [6]	Faithfulness audit of feedback on student project reports, plus a separate annotated detection dataset.	Input-faithfulness hallucination rates were 27.1% for the data-driven BART system and 23.5% for the prompt-driven GPT-4 system.	Evidence integrity and support relative to student input. Input faithfulness is not general factuality.
Nemeth et al. [7]	Pilot retrieval-augmented tutor across ´ four courses in three programs at two universities; expert coding of sampled answers.	Of sampled answers, 1.5% were coded incorrect and 16.5% were outside supplied context; hallucinations were less frequent on topics sufficiently covered by the materials.	Factual status and course support. Pilot/system-specific findings; not evidence of learning effects or broad generalizability.
Misconception Bench [8]	Science-misconception matching and generation, plus response-distribution prediction; 761 items, 1,474 distractors, and responses from more than 150,000 students.	Top generation models achieved Hit@1 of 38.5–39.6%; no model exceeded F1=0.21 when predicting grade-band differences in student responses.	Misconception handling. Benchmark tasks do not test course-pack faithfulness in live tutoring or post-tutoring learning.
XYBench [9]	8,115 misconception-bearing queries testing pragmatic recognition across technical and everyday contexts.	Models answered literal requests more often than intended needs (0.75–0.92 vs. 0.33–0.71); misconception identification reached at most 63%, compared with 79–90% for human responses.	Claim interpretation and response strategy. Not a course-specific tutor, multi-turn intervention, or learner-outcome study
Yasir et al. [10]	Seven feedback agents evaluated on 10,836 propositional-logic solution–feedback pairs across three conditions.	Agents often rejected valid but suboptimal reasoning or endorsed incorrect solutions; accurate diagnosis did not always yield actionable feedback.	Factual diagnosis and instructional usefulness. Benchmark performance does not establish classroom learning effects

Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.3

4.3 Safeguards and Their Limits

Retrieval-augmented generation supplies course context but does not ensure that materials cover the question, relevant evidence is retrieved, or generated claims reflect that evidence [7], [12]. Evaluate claim-source support rather than the presence of retrieval or citations alone [6]. When the source pack is silent, how the tutor handles that gap is part of its reliability. Evidence-based abstention can signal insufficient source coverage, express uncertainty, or request clarification. Its value depends on calibration: refusing answerable questions can reduce usefulness, while answering unanswerable ones can introduce unsupported claims. In non-educational search-augmented QA, search generally improved accuracy on answerable queries but harmed abstention on unanswerable ones; noisy retrieval exacerbated over-searching, which compounded across turns. This is not evidence of educational efficacy, but it supports testing both answerable and unanswerable course questions when evaluating retrieval and abstention [13]. Human oversight and pedagogical design may address concerns that retrieval alone cannot resolve. As proposed practices, instructors can review source packs, clarify objectives, and provide a way to flag or correct confusing output; a tutor can acknowledge a learner's reasoning, explain a relevant distinction, and check understanding. These recommendations align with human-centred guidance for educational AI but are not evidence that any specific practice improves learning; evaluate them in the implementation context [2], [15].

V. LIMITATIONS AND FUTURE DIRECTIONS

The dimensions and risk rubric are proposed and remain unvalidated; local adaptation and reliability testing are needed across curricula, ages, languages, and tasks. The defect-related rubric is narrower than pedagogical risk in general. Excessive answer-giving, answer leakage, reduced productive effort, or missing scaffolding may impair learning even when individual claims are factually correct. Tutor-package experiments such as Bastani et al. measure effects of a tested package, not the causal effect of a particular factual defect [4]. Future work should evaluate these broader interaction-design risks separately and link audited exposure to direct learner outcomes. The framework is intended to transfer to essay feedback, language learning, and code tutoring only after adapting the reference standards, course-support criteria, and annotation units. For example, essay feedback may be segmented into claims about the submitted text and proposed revisions; language tutoring into usage claims and corrections; and code tutoring into behavioral claims, test results, and suggested changes. Equations and procedural solutions should be segmented into linked steps when their dependencies affect correctness. The selected studies illustrate evidence types rather than provide an exhaustive map or comparable prevalence estimates. Candidate identification was purposive and is not reproducible from a documented search log; future syntheses should record search sources, dates, and selection decisions. Further research should test whether the framework improves annotation comparability, report null findings and safeguard trade-offs, and protect student data through ethics review and applicable privacy standards [15].

VI. CONCLUSION

Educational LLM reliability requires separate reporting of factual status, course support, evidence integrity, dialogue consistency, contextual risk from output defects, and direct learner outcomes. The proposed workflow and cross-classification make these distinctions operational while keeping broader instructional-design harms outside the defect rubric. The next step is to test annotation reliability and connect specific tutor exposures to learning measures.

REFERENCES

[1] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Art. no. 248, pp. 1–38, 2023, doi: 10.1145/3571730.

[2] K. U. Danyaro, S. Abdullahi, A. S. Abdallah, and H. Chiroma, "Hallucinations in Large Language Models for Education: Challenges and Mitigation," International Journal of Teaching, Learning and Education, vol. 4, no. 6, pp. 13–19, 2025, doi: 10.22161/ijtle.4.6.2.

[3] S. Abdullahi, Umama, K. U. Danyaro, H. Chiroma, A. Zakari, M. Abdullahi, S. H. Muhammad, and N. A. W. A. Zawawi, "Large Language Models Applications in Oil and Gas: Case Studies and Insights," in 2025 International Conference on Artificial Intelligence for Sustainable Innovation (AI-SI), pp. 1–6, 2025, doi: 10.1109/AISI66213.2025.11341041.

[4] H. Bastani, O. Bastani, A. Sungu, H. Ge, O. Kabakcı, and R. Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics," Proceedings of the National Academy of Sciences, vol. 122, no. 26, Art. no. e2422633122, 2025, doi: 10.1073/pnas.2422633122.

[5] Z. A. Pardos and S. Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-

authored help on mathematics skills," PLOS ONE, vol. 19, no. 5, Art. no. e0304013, 2024, doi: 10.1371/journal.pone.0304013.

[6] Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer, "On assessing the faithfulness of LLM-generated feedback on student assignments," in Proceedings of the 17th International Conference on Educational Data Mining, pp. 491–499, Atlanta, GA, USA, 2024, doi: 10.5281/zenodo.12729868.

[7] R. Nemeth, A. T ´ atrai, M. Szab ´ o, P. T. Zaletnyik, and ´ A. Tam ´ asi, ´ "Exploring the use of retrieval-augmented generation models in higher education: A pilot study on artificial intelligence-based tutoring," Social Sciences & Humanities Open, vol. 12, Art. no. 101751, 2025, doi: 10.1016/j.ssaho.2025.101751.

[8] Y. Zhang, Z. Wang, C. Zhang, Y. Gao, W. Ma, and X. Zhai, "MisconceptionBench: Evaluating LLMs on Science Misconceptions and Student Response Distributions," in Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pp. 16525–16539, Budapest, Hungary, 2026, doi: 10.18653/2026.emnlp-main.712.

[9] A. Yerukola, J. D. Hwang, M. Zheng, J. Godsey, H. Kim, V. Pyatkin, J. Hu, and M. Sap, "XYBench: Can LLMs Respond Pragmatically to Queries with Misconceptions?" in Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pp. 60203–60234, Budapest, Hungary, 2026, doi: 10.18653/2026.emnlp-main.2646.

[10] T. Yasir, W. Li, S. Gilson, S. Tithi, X. Tian, and T. Barnes, "Confirming correct, missing the rest: LLM tutoring agents struggle where feedback matters most," in Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pp. 819–840, San Diego, CA, USA, 2026, doi: 10.18653/v1/2026.bea-1.56.

[11] A. Nura and S. Abdullahi, "A systematic review of multi-depot vehicle routing problems," Systematic Literature Review and Meta-Analysis Journal, vol. 3, no. 2, pp. 51–60, 2022, doi: 10.54480/slrm.v3i2.37.

[12] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[13] R. Xie, D. Gopinath, D. Qiu, D. Lin, H. Sun, S. Potdar, and B. Dhingra, "Over-Searching in Search-Augmented Large Language Models," in Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7714–7739, Rabat, Morocco, 2026, doi: 10.18653/v1/2026.eacllong.361.

[14] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," Communication Methods and Measures, vol. 1, no. 1, pp. 77–89, 2007, doi: 10.1080/19312450709336664.

[15] UNESCO, Guidance for generative AI in education and research. Paris, France: UNESCO, 2023.