JOBIG Journal of Business, Innovation and Governance
ISSN 2717-9559 dergipark.org.tr/tr/pub/jobig
İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ

HALLUCINATION IN LLM-BASED EDUCATIONAL TOOLS: RISKS AND SOLUTIONS FOR RELIABLE LEARNING

Elitsa Peltekova,
Dr., Sofia University, epeltekova@fmi.uni-sofia.bg, 0000-0001-8672-0612.
Dafinka Miteva,
Asst. Prof. Dr., Sofia University, dafinca@fmi.uni-sofia.bg, 0000-0002-4841-2245.
Ioannis Patias,
Assoc. Prof. Dr., Sofia University, patias@fmi.uni-sofia.bg, 0000-0003-1355-7433

Kabul Tarihi / Accepted: 27 Şubat 2026

İletişim / Correspondence: Elitsa Peltekova

Benzerlik Oranı/ Plagiriasim: %1

Makale Türü/Article Type: Araştırma Makalesi/ Research Article

ABSTRACT

Large Language Models (LLMs) have rapidly emerged as transformative tools in education, offering personalized tutoring, content generation, and intelligent feedback. However, their widespread adoption is often restricted by a critical limitation: hallucinations, sounding true but factually incorrect outputs. This paper proposes a structured, three-layered framework for mitigating hallucinations in LLM-based educational tools.

At the Input Level, hallucination risks are addressed by refining the quality and clarity of user prompts. Techniques such as prompt engineering, contextual grounding with curriculum-aligned materials, and input validation are explored to ensure that the model receives precise and relevant queries.

The Model Level focuses on enhancing the internal reasoning and factual grounding of the LLM itself. We examine the efficacy of retrieval-augmented generation (RAG), fine-tuning with curated educational datasets, and the application of symbolic constraints or logic overlays.

At the Output Level, we propose post-generation verification strategies to ensure factual accuracy before content is delivered to learners. This layer acts as a final safeguard, ensuring that only validated and pedagogically sound content reaches the end user.

By organizing hallucination mitigation strategies across these three layers, this framework provides a comprehensive roadmap for developers, educators, and researchers seeking to deploy LLMs responsibly in educational environments.

Key words: Large Language Models (LLMs), Educational Technology, AI Hallucinations, Retrieval-Augmented Generation, AI Safety in Education.

JEL Codes: D83, I21, I23, O33.

1. INTRODUCTION

The integration of LLMs into educational technology has initiated a new era of personalized, scalable, and interactive learning experiences (Alhafni et al., 2024: 1) (Patias, 2025: 4). From intelligent tutoring systems and automated grading assistants to curriculum-aligned content generators, LLMs are reshaping how learners engage with knowledge (Dennison et al., 2025: 4) (Shahzad et al., 2025: 4). Their ability to generate coherent, contextually relevant, and human-like responses makes them particularly attractive for educational applications. However, this promise is under the risk of a critical limitation: hallucinations (Cossio, 2025: 1). These are instances where the model produces information that is syntactically true but factually incorrect, misleading, or entirely fabricated.

1.1. Definition and impact of hallucinations

In educational contexts, hallucinations are not just technical issues, they are pedagogical liabilities (Jančařík & Dušek, 2024: 122). A hallucinated explanation in a science lesson, a fabricated historical date, or an incorrect mathematical derivation can misinform learners, reinforce misconceptions, and destroy the trust in AI-assisted learning environments. Unlike casual applications of LLMs, where minor inaccuracies may be tolerable, educational tools demand a higher standard of factual integrity and epistemic responsibility. As such, mitigating hallucinations is not just a technical challenge but a foundational requirement for the ethical deployment of LLMs in education (Patias et al., 2026: 163) (Wang et al., 2025: 1) (K. Z. Zhou et al., 2024: 3).

1.2. Motivation for a layered mitigation framework

This paper introduces a structured, three-layered framework for hallucination mitigation tailored specifically to educational use cases. The framework is organized into three interdependent layers: input-level, model-level, and output-level interventions. Each layer addresses hallucination risks at a different stage of the LLM pipeline, from the formulation of the user query to the final delivery of the model's response.

At the input level, the focus is on shaping the user's prompt and contextualizing it with relevant, high-quality information. This includes techniques such as prompt engineering, curriculum-aware grounding, and input validation. At the model level, the emphasis shifts to the internal mechanisms of the LLM, including retrieval-augmented generation (RAG), and fine-tuning with cross-checked educational datasets. Finally, the output level introduces post-generation safeguards such as fact-checking modules, confidence scoring, human-in-the-loop review, and explainability features like citation generation.

By decomposing hallucination mitigation into these three layers, the framework provides a comprehensive and modular approach that can be adapted to various educational platforms and pedagogical goals. It also facilitates clearer evaluation metrics and implementation strategies, enabling developers and educators to make informed decisions about the design and deployment of LLM-based tools. This paper aims to bridge the gap between the technical sophistication of LLMs and the pedagogical accuracy required in educational settings, offering a roadmap for building safer, more trustworthy AI-powered learning systems.

2. RELATED WORK

The phenomenon of hallucination in natural language generation has been widely documented across various domains, including machine translation, summarization, and open-domain question answering. In the context of LLMs, hallucinations are typically categorized as either intrinsic (arising from the model's internal representations) or extrinsic (resulting from a lack of grounding in external knowledge) (Cossio, 2025: 1). While much of the early research focused on hallucinations in general-purpose models, recent studies have begun to explore their implications in high-stakes domains such as healthcare, law, and education (Kim et al., 2025: 1) (Magesh et al., 2024: 1) (J. Zhou et al., 2025: 1) (Z. Z. Chen et al., 2024: 1).

2.1. Existing mitigation strategies

Fine-tuning LLMs on domain-specific datasets is another common strategy. In educational contexts, this involves training models on textbooks, academic papers, and curriculum-aligned materials (Patias et al., 2026: 160). While this can improve domain precision, it also introduces challenges related to data curation, bias, and overfitting. Moreover, fine-tuning alone does not guarantee hallucination-free outputs, especially in open-ended or generative tasks.

Post-hoc verification methods have also gained traction. These include automated fact-checking systems, confidence scoring mechanisms, and human-in-the-loop review pipelines.

2.2. Gaps in current educational applications

Despite these advances, few studies have proposed a unified framework that integrates mitigation strategies across the entire LLM lifecycle. Most existing approaches focus on isolated interventions, either at the model level (e.g., fine-tuning, RAG) or at the output level (e.g., fact-checking) (Cossio, 2025: 7) (Z. Z. Chen et al., 2024: 9) (Rahman et al., 2025: 1). This siloed approach limits the effectiveness of hallucination control, particularly in dynamic and learner-driven environments where inputs vary widely and pedagogical stakes are high.

The proposed three-layered framework builds on this body of work by offering a holistic perspective. It synthesizes insights from prompt engineering, model architecture, and post-processing to create a modular and adaptable system for hallucination mitigation. By aligning technical interventions with pedagogical goals, the framework addresses a critical gap in the literature and provides a foundation for future research in trustworthy educational AI.

3. THE THREE-LAYERED FRAMEWORK

This section outlines the methodology for mitigating hallucinations in LLM-based educational tools through a three-layered framework (**Figure 1.** The three-layered frameworkFigure 1). Each layer (input-level, model-level, and output-level) targets a distinct phase in the LLM pipeline, enabling modular and scalable intervention strategies.

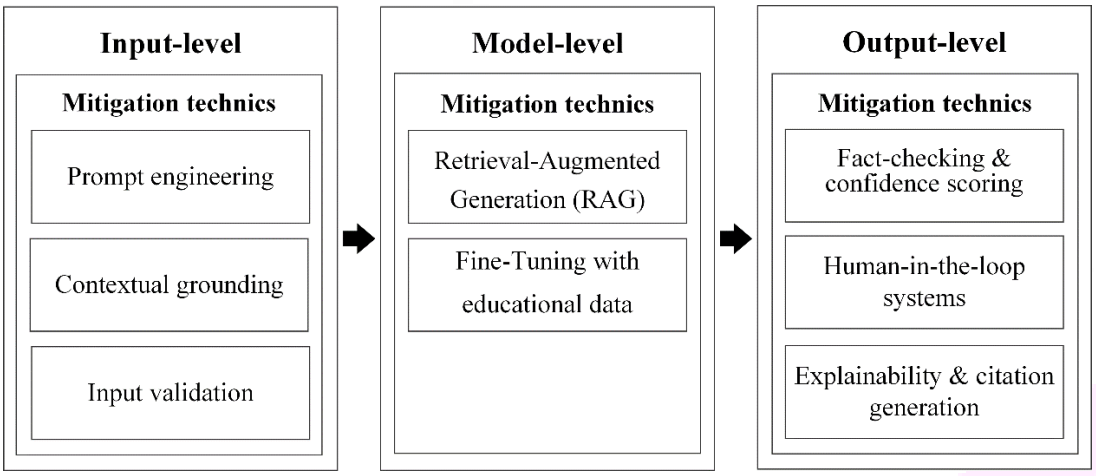

Figure 1. The three-layered framework

The framework is designed to be adaptable across educational platforms and learning contexts, ensuring both technical robustness and pedagogical integrity.

3.1. Input-level mitigation

Input-level strategies aim to reduce hallucination risk by optimizing the quality, clarity, and contextual relevance of the user's prompt before it reaches the model. This layer is critical because ambiguous or poorly structured inputs often lead to speculative or fabricated outputs.

3.1.1. Prompt engineering

Prompt engineering involves crafting inputs that guide the model toward accurate, relevant, and grounded responses (B. Chen et al., 2025: 1). In educational settings, this includes:

- Structured prompts: Using templates like "Explain [concept] using [source]" or "Summarize this passage in three bullet points" to constrain the model's generative space.
- Explicit task framing: Clarifying the user's intent (e.g., "define," "compare," "solve") to reduce ambiguity.
- Instructional framing: Embedding pedagogical signals such as "use examples from biology" or "align with 9th-grade curriculum" to tailor the response to the learner's level.

Effective prompt engineering not only improves factuality but also enhances pedagogical alignment, ensuring that outputs are age-appropriate and curriculum-relevant.

3.1.2. Contextual grounding

Contextual grounding supplements the prompt with relevant background information or reference materials (Lee et al., 2025: 17). This is particularly important in education, where domain specificity and curricular alignment are essential, following:

- Curriculum-aware grounding: Attaching textbook excerpts, lesson plans, or learning objectives to the prompt.

- Dynamic retrieval: Pulling in relevant documents or examples from a knowledge base or learning management system (LMS).

- Semantic linking: Embedding key terms or concepts from prior lessons to maintain continuity and coherence.

By linking the model's response to trusted educational content, contextual grounding reduces the likelihood of hallucinations and promotes conceptual accuracy.

3.1.3. Input validation

Input validation acts as a gatekeeper, screening prompts for ambiguity, incompleteness, or misleading phrasing before they are processed by the model having (Qamar et al., 2024: 3):

- Intent classification: Detecting whether the user is asking for factual information, creative writing, or opinion-based content, and routing accordingly.

- Ambiguity detection: Flagging vague prompts (e.g., "Tell me about cells") and prompting the user for clarification (e.g., "Do you mean biological cells or prison cells?").

- Pre-processing filters: Removing or correcting malformed inputs, such as typos or contradictory instructions.

This layer ensures that the model receives high-quality inputs, reducing the cognitive load on the LLM and minimizing the risk of speculative generation.

3.2. Model-level litigation

Model-level strategies focus on enhancing the internal mechanisms of the LLM to improve factual grounding and reduce the generation of hallucinated content. These interventions are typically architectural or training-based.

3.2.1. Retrieval-Augmented Generation (RAG)

RAG integrates external knowledge retrieval into the generation process, allowing the model to ground its responses in real-time data with concrete characteristics (Cheng et al., 2025: 1):

- Architecture: Combines a retriever with the generator, the transformer-based LLM.

- Workflow: The retriever fetches relevant documents based on the input query, and the generator conditions its output on these documents.

- Benefits: Reduces hallucinations by linking the responses in verifiable sources, especially useful for fact-heavy subjects like history or science.

In educational tools, RAG can be linked to curated repositories such as academic databases, digital textbooks, or institutional knowledge graphs.

3.2.2. Fine-tuning with educational data

Fine-tuning involves training the LLM on domain-specific datasets to improve its accuracy and relevance in educational contexts (Patias et al., 2026: 1), based on:

- Data sources: Textbooks, exam questions, teacher-generated content, and annotated student responses.
- Curriculum alignment: Ensures that the model's outputs reflect the standards and expectations of specific educational systems.
- Bias control: Allows for the removal of culturally or pedagogically inappropriate content during training.

While fine-tuning enhances domain accuracy, it must be carefully managed to avoid overfitting or the inclusion of outdated information.

3.3. Output-level mitigation

Output-level strategies serve as the final checkpoint before the model's response is delivered to the user. These interventions focus on verification, transparency, and human oversight.

3.3.1. Fact-checking and confidence scoring

Automated fact-checking modules and confidence scoring systems assess the factual integrity of the model's output (Rahman et al., 2025: 1) (Qamar et al., 2024: 1), by using:

- Fact-checking APIs: Compare generated claims against trusted databases (e.g., Wikipedia, academic journals).
- Claim decomposition: Break down complex responses into atomic claims for individual verification.
- Confidence scoring: Assign a reliability score to each output, allowing users to gauge the model's certainty.

These tools can be integrated into educational dashboards, alerting teachers or learners when a response may require further control.

3.3.2. Human-in-the-loop systems

Human oversight remains a critical component of hallucination mitigation, especially in high-stakes or formative learning environments (Z. Z. Chen et al., 2024: 34), with various means:

- Teacher review portals: Allow educators to approve, edit, or reject AI-generated content before it reaches students.

- Feedback loops: Enable users to flag hallucinated responses, which can be used to retrain or fine-tune the model.

- Hybrid workflows: Combine automated generation with human curation, particularly for assessments or instructional materials.

This approach balances the scalability of LLMs with the pedagogical judgment of human educators.

3.3.3. Explainability and citation generation

Transparency mechanisms help users understand the origin and rationale behind the model's responses, as they have different options (Papagiannopoulos et al., 2025: 1):

- Source assignment: Include citations or links to the documents used during generation.

- Rationale tracing: Provide step-by-step explanations of how the model arrived at a particular answer.

- Visual aids: Use diagrams, highlights, or annotations to clarify complex concepts.

Explain ability not only builds trust but also supports metacognitive learning, helping students understand not just what the answer is, but why it is correct.

The discussed techniques and their benefits for learning are summarized in Table 1.

Table 1: The three-layered framework – description and benefits

Layer	Technique	Description	Educational benefit
Input-level	Prompt engineering	Crafting structured, task-specific prompts to guide model behavior.	Reduces ambiguity and improves alignment with learning objectives.
	Contextual grounding	Supplementing prompts with curriculum-aligned documents or examples.	Anchors responses in trusted educational content.
	Input validation	Screening prompts for vagueness, errors, or misleading phrasing.	Ensures high-quality inputs and reduces speculative generation.
Model-level	Retrieval-Augmented Generation (RAG)	Integrating external knowledge retrieval into the generation process.	Grounds outputs in real-time, verifiable sources.
	Fine-Tuning with educational data	Training the model on vetted academic content and domain-specific datasets.	Enhances domain fidelity and curriculum relevance.
Output-level	Fact-checking & confidence scoring	Verifying claims post-generation and tagging responses with reliability scores.	Flags potentially hallucinated content before learner exposure.
	Human-in-the-loop systems	Allowing educators to review, approve, or correct AI-generated content.	Adds expert oversight and ensures pedagogical integrity.
	Explainability & citation generation	Providing source references and rationale for each response.	Builds trust and supports metacognitive learning.

This three-layered framework offers a holistic and modular approach to hallucination mitigation in LLM-based educational tools. By addressing risks at the input, model, and output levels, it enables developers and educators to build systems that are not only technically robust but also pedagogically sound. Each layer contributes uniquely to the overall reliability of the system, and together they form a resilient architecture capable of supporting trustworthy, AI-enhanced learning experiences.

3.3.4. Implementation considerations

Implementing the proposed three-layered framework for hallucination mitigation in LLM-based educational tools requires careful orchestration of technical components, pedagogical goals, and operational constraints. This section outlines key considerations for deploying the framework in real-world educational environments.

3.4. Integration into existing platforms

Educational platforms vary widely in architecture, ranging from lightweight mobile apps to institution-level learning management systems (LMS). Integrating LLMs into these platforms demands modular design (Patias et al., 2024: 1). Input-level components such as prompt engineering and validation can be embedded into user interfaces, while model-level interventions like RAG and fine-tuning require backend infrastructure. Output-level safeguards, including fact-checking and human review portals, can be layered into content delivery pipelines.

3.5. Trade-offs between automation and oversight

Educational data often includes sensitive information about students, teachers, and institutions. Implementing retrieval systems or fine-tuning models on proprietary datasets must comply with data protection regulations such as GDPR. Secure data handling, anonymization, and access control are essential to ensure ethical deployment.

3.6. Scalability and cost

While hallucination mitigation improves reliability, it introduces computational overhead. Retrieval-augmented generation, ensemble modeling, and post-processing modules can increase latency and resource consumption. Developers must balance accuracy with performance, especially in large-scale deployments. Techniques such as caching, model distillation, and selective verification can help optimize resource use.

3.7. Human oversight and workflow design

Human-in-the-loop systems require clear workflows for educators to review, approve, or correct AI-generated content. This includes designing intuitive dashboards, notification systems, and feedback loops. Training educators to interpret model confidence scores and rationale traces is also critical for effective oversight.

3.8. Curriculum alignment and localization

Educational content must align with regional curricula, language standards, and cultural norms. Fine-tuning and grounding strategies should reflect local pedagogical frameworks. Localization efforts may involve translating prompts, adapting examples, and incorporating region-specific knowledge bases.

3.9. Continuous monitoring and adaptation

Hallucination patterns may evolve over time as models are updated or as user behavior shifts. Implementing monitoring systems to track hallucination frequency, types, and contexts enables adaptive mitigation. Feedback from users, both learners and educators, should be systematically collected and used to refine prompts, retrievers, and verification modules.

In summary, successful implementation of the framework requires a multidisciplinary approach, combining AI engineering, instructional design, data governance, and user experience. By addressing these considerations, educational platforms can deploy LLMs that are not only powerful but also trustworthy and pedagogically sound.

4. EVALUATION METRICS

Evaluating the effectiveness of hallucination mitigation strategies in LLM-based educational tools demands a multi-dimensional approach (Xu et al., 2024: 3). This section outlines key metrics and methodologies for assessing the performance, reliability, and educational impact of the proposed framework.

4.1. Accuracy and factuality benchmarks

The primary metric for hallucination mitigation is factual accuracy. This can be measured using benchmark datasets, or custom curriculum-aligned corpora (Rahman et al., 2025: 1) (Qamar et al., 2024: 1). Metrics include:

- Precision and recall of factual claims
- Rate of hallucinated outputs per 100 responses
- Comparison against human-verified ground truth

Automated fact-checking tools can assist in large-scale evaluations, while manual annotation provides deeper insights into nuanced errors.

4.2. User trust and satisfaction

Trust is essential for adoption. Surveys, interviews, and usage analytics can measure (Brown, 2024: 1):

- Learner confidence in AI-generated content
- Educator satisfaction with oversight tools

- Frequency of flagged or corrected outputs

Trust metrics should be tracked periodically to assess how mitigation strategies influence user perception over time.

4.3. Educational outcomes

Beyond factual correctness, outputs must align with pedagogical goals. Educational validity assesses whether responses are (Benedetto et al., 2024: 11351):

- Curriculum-aligned
- Age-appropriate
- Conceptually accurate and clear

Tests developed by educators can be used to score AI-generated content on these dimensions. Peer review by subject-matter experts adds further accuracy.

4.4. Explainability and transparency

Explainability metrics assess how well the system communicates its reasoning (Seth & Sankarapu, 2025: 1). These include:

- Citation coverage (percentage of outputs with sources)
- Rationale clarity (user-rated quality of explanations)
- Educator feedback on traceability

Explainable outputs not only reduce hallucination risk but also support metacognitive learning.

Together, these metrics provide a comprehensive framework for evaluating hallucination mitigation in educational LLMs. They enable developers and educators to make data-driven decisions, ensuring that AI tools enhance learning without compromising trust or accuracy.

5. CASE STUDY – THE PROJECT RISK MANAGEMENT TRAINER

In this section, the application of the three-layer model in a practical case study is described, as applied by a project risk management instructor in practice (Patias et al., 2026: 1).

5.1. Context

We described an AI-trainer for project risk management, to achieve personalized learning, as LLM-powered assistant. The trainer is designed to answer learner queries, simulate risk scenarios, and provide feedback on risk assessment exercises.

5.2. Challenge

While the LLM demonstrates strong fluency and engagement, it occasionally hallucinates critical project management concepts, such as misclassifying risk types, fabricating mitigation strategies, or citing non-existent standards (e.g., ISO guidelines that don't exist). These hallucinations pose serious risks to learner understanding and professional credibility, especially in a domain where precision and compliance are essential.

5.3. Framework application

Table 2 provides a summary of the details related to the example of using the framework application in the LLM-based risk management trainer.

5.3.1. Input-level mitigation

- Prompt Engineering: Trainers design structured prompts like "Explain the difference between qualitative and quantitative risk analysis using PMBOK terminology" to guide the model toward accurate, standards-based responses.
- Contextual Grounding: Each prompt is paired with references to the PMBOK Guide, ensuring the model has access to authoritative sources.
- Input Validation: The system flags vague queries such as "What's a good risk strategy?" and prompts learners to specify context (e.g., industry, project phase).

5.3.2. Model-level mitigation

- RAG: The LLM is connected to a curated database of project management literature, including PMI publications, case studies, and regulatory documents.
- Fine-Tuning with Educational Data: The model is fine-tuned on past training materials, certified exam questions, and annotated trainer feedback to align with professional standards.

5.3.3. Output-level mitigation

- Fact-Checking and Confidence Scoring: Responses are automatically checked against the PMI database. Outputs with low confidence scores are flagged for review.
- Human-in-the-Loop Systems: Trainers receive dashboards showing flagged responses and can approve or correct them before learners see them.

• Explainability and Citation Generation: Each response includes citations (e.g., "PMBOK 6th Edition, Section 11.3") and rationale tracing to show how the model arrived at its conclusion.

Table 2: The three-layered framework – Case study: the project risk management trainer.

Layer	Technique	Application in educational tools	Example use case
Input-level	Prompt Engineering	Design structured prompts that guide the model toward factual and pedagogically aligned responses.	"Explain the difference between qualitative and quantitative risk analysis using PMBOK terminology."
	Contextual Grounding	Attach curriculum-aligned materials or reference documents to the prompt.	Include references to PMBOK when asking for risk identification techniques.
	Input Validation	Screen prompts for ambiguity or errors before processing.	Flag vague queries like "What's a good risk strategy?" and prompt for project phase or context.
Model-level	Retrieval-Augmented Generation (RAG)	Integrate external knowledge sources during generation to ground responses in real-time data.	Retrieve relevant PMI standards or case studies when asked about risk mitigation planning.
	Fine-Tuning with Educational Data	Train the model on vetted educational content to improve domain-specific accuracy.	Use annotated trainer feedback and past exam questions from risk management modules.
Output-level	Fact-Checking & Confidence Scoring	Verify generated claims and tag responses with reliability scores.	Flag low-confidence answers about risk prioritization or probability-impact matrices.
	Human-in-the-Loop Systems	Allow educators to review and approve AI-generated content before delivery.	Trainers validate simulated risk scenarios before learners interact with them.
	Explainability & Citation Generation	Provide source references and rationale for each response.	Include citations like "PMBOK Guide, Section 11.3" when explaining risk response strategies.

6. DISCUSSION

The proposed three-layered framework offers a structured and scalable approach to mitigating hallucinations in LLM-based educational tools. By addressing hallucination risks at the input, model, and output levels, the framework enables developers and educators to intervene at multiple stages of the LLM pipeline, enhancing both technical reliability and pedagogical integrity.

One of the key strengths of this framework is its modularity. Each layer can be implemented independently or in combination, allowing for flexible adaptation to different educational platforms, user demographics, and curricular goals. For instance, a lightweight tutoring app may prioritize input-level prompt engineering and output-level fact-checking, while a full-scale LMS might integrate retrieval-augmented generation and human-in-the-loop review systems.

The framework also bridges the gap between AI safety and educational design. Hallucinations are not merely technical errors; they are pedagogical risks that can destroy learning outcomes. By embedding mitigation strategies into the educational workflow, the framework promotes responsible AI use and supports epistemic trust among learners and educators.

However, several challenges remain. Implementing retrieval systems and ensemble models can be resource-intensive, especially in low-bandwidth or underfunded educational settings. Human oversight, while effective, may not scale easily across large user bases. Additionally, evaluation metrics must be continuously refined to capture the nuanced impact of hallucinations on learning.

Future research should explore adaptive systems that learn from hallucination patterns, as well as collaborative frameworks that combine AI with educator expertise. The integration of multimodal inputs (e.g., images, diagrams) and multilingual support also presents promising avenues for expanding the framework's applicability.

In sum, this layered approach provides a robust foundation for building trustworthy, AI-enhanced educational tools that prioritize accuracy, transparency, and learner safety.

7. CONCLUSION AND FUTURE WORK

Hallucinations in LLM-based educational tools pose a significant challenge to the reliability, safety, and pedagogical value of AI-assisted learning. As these tools become increasingly embedded in classrooms, tutoring platforms, and assessment systems, ensuring factual accuracy and epistemic trust is paramount. This paper presents a three-layered framework for hallucination mitigation, offering a comprehensive strategy that spans input-level, model-level, and output-level interventions.

At the input level, prompt engineering, contextual grounding, and input validation help shape high-quality queries that reduce ambiguity and guide the model toward accurate responses. Model-level strategies such as retrieval-augmented generation, and fine-tuning with educational data. Output-level safeguards, including fact-checking, human-in-the-loop review, and explain ability features, serve as the final checkpoint before content reaches learners.

Together, these layers form a resilient architecture that can be tailored to diverse educational contexts and technical infrastructures. The framework not only mitigates hallucinations but also supports broader goals of AI safety, transparency, and pedagogical alignment. It empowers developers to build systems that are both intelligent and trustworthy, and it equips educators with tools to oversee and refine AI-generated content.

While challenges remain in terms of scalability, resource demands, and evaluation, the framework provides a clear roadmap for future innovation. By integrating technical rigor with educational sensitivity, it lays the groundwork for a new generation of AI-powered learning environments that are accurate, adaptive, and learner-centered.

Ultimately, the success of LLMs in education will depend not only on their generative capabilities but on their ability to support meaningful, truthful, and equitable learning experiences. This framework is a step toward realizing that vision.

REFERENCES

Alhafni, B., Vajjala, S., Bannò, S., Maurya, K. K., and Kochmar, E. (2024, 18 September). *LLMs in Education: Novel Perspectives, Challenges, and Opportunities*. https://arxiv.org/abs/2409.11917 (Access Date, 30 November 2025)

Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., and Buttery, P. (2024). Using LLMs to simulate students' responses to exam questions. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 11351–11368). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.663

Brown, N. B. (2024, 4 June). *Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs*. https://arxiv.org/abs/2406.01943 (Access Date, 30 November 2025)

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2025, 13 June). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(6), 101260. https://doi.org/10.1016/j.patter.2025.101260 (Access Date, 30 November 2025)

Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., and Wang, W. Y. (2024, 2 May). *A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law*. https://arxiv.org/abs/2405.01769 (Access Date, 30 November 2025)

Cheng, M., Luo, Y., Ouyang, J., Liu, Q., Liu, H., Li, L., Yu, S., Zhang, B., Cao, J., Ma, J., Wang, D., and Chen, E. (2025, 11 March). *A Survey on Knowledge-Oriented Retrieval-Augmented Generation*. https://arxiv.org/abs/2503.10677 (Access Date, 30 November 2025)

Cossio, M. (2025, 3 August). *A comprehensive taxonomy of hallucinations in Large Language Models*. https://arxiv.org/abs/2508.01781 (Access Date, 30 November 2025)

Dennison, D. V., Ahtisham, B., Chourasia, K., Arora, N., Singh, R., Kizilcec, R. F., Nambi, A., Ganu, T., and Vashistha, A. (2025, 1 July). *Teacher-AI Collaboration for Curating and Customizing Lesson Plans in Low-Resource Schools*. https://arxiv.org/abs/2507.00456 (Access Date, 30 November 2025)

Jančařík, A., and Dušek, O. (2024, 23 October). The Problem of AI Hallucination and How to Solve It. *European Conference on E-Learning*, *23*, 122–128. https://doi.org/10.34190/ecel.23.1.2584 (Access Date, 30 November 2025)

Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., Gameiro, R., Fan, L., Park, E., Lin, T., Yoon, J., Yoon, W., Sap, M., Tsvetkov, Y., Liang, P., Xu, X., and Breazeal, C. (2025, 2 November). *Medical Hallucinations in Foundation Models and Their Impact on Healthcare*. https://arxiv.org/abs/2503.05777 (Access Date, 30 November 2025)

Lee, H., Yoon, S., Won, Y., Oh, H., Kim, G., Bui, T., Dernoncourt, F., Stengel-Eskin, E., Bansal, M., and Seo, M. (2025, 18 June). *Context-Informed Grounding Supervision*. https://arxiv.org/abs/2506.15480 (Access Date, 30 November 2025)

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2024). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. https://arxiv.org/abs/2405.20362

Papagiannopoulos, I., Koutalidis, H., Rempi, P., Ntanos, C., and Askounis, D. (2025, 23 July). Comparison of explainability methods for hallucination analysis in LLMs [version 1; peer review: 1 not approved]. *Open Research Europe*, *5*(191). https://doi.org/10.12688/openreseurope.20839.1 (Access Date, 30 November 2025)

Patias, I. (2025, 12 August). Foundational Models as General-Purpose Technology: A Guide to Corporate Transformation. *2025 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 1–6. https://doi.org/10.1109/ISAS66241.2025.11101869(Access Date, 30 November 2025)

Patias, I., Miteva, D., and Peltekova, E. (2026). Using Old Lessons for New AI – A Trainer for Project Risk Management. In G. De Tré, S. Sotirov, J. Kacprzyk, G. Psaila, G. Smits, T. Andreasen, G. Bordogna, & H. Legind Larsen (Eds.), *Flexible Query Answering Systems* (pp. 155–167). Springer Nature Switzerland.

Patias, I., Miteva, D., Peltekova, E., Wright, M., and Gasteiger-Klicpera, B. (2024, 6 December). Leveraging Large Language Models to Enhance Mental Health Literacy and Diversity Awareness in Adolescents: The me_HeLi-D Project. *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 1–5. https://doi.org/10.1109/ISAS64331.2024.10845582 (Access Date, 30 November 2025)

Qamar, M. T., Yasmeen, J., Pathak, S. K., Sohail, S. S., Madsen, D. Ø., and Rangarajan, M. (2024, 18 June). Big claims, low outcomes: Fact checking ChatGPT's efficacy in handling linguistic creativity and ambiguity. *Cogent Arts & Humanities*, *11*(1), 2353984. https://doi.org/10.1080/23311983.2024.2353984 (Access Date, 30 November 2025)

Rahman, S. S., Islam, M. A., Alam, M. M., Zeba, M., Rahman, M. A., Chowa, S. S., Raiaan, M. A. K., and Azam, S. (2025, 26 September). *Hallucination to Truth: A Review of Fact-Checking and Factuality Evaluation in Large Language Models*. https://arxiv.org/abs/2508.03860 (Access Date, 30 November 2025)

Seth, P., & Sankarapu, V. K. (2025, 20 November). *Bridging the Gap in XAI-Why Reliable Metrics Matter for Explainability and Compliance*. https://arxiv.org/abs/2502.04695 (Access Date, 30 November 2025)

Shahzad, T., Mazhar, T., Tariq, M. U., Ahmad, W., Ouahada, K., and Hamam, H. (2025, 14 January). A comprehensive review of large language models: Issues and solutions in learning environments. *Discover Sustainability*, *6*(1), 27. https://doi.org/10.1007/s43621-025-00815-8 (Access Date, 30 November 2025)

Wang, H., Fu, W., Tang, Y., Chen, Z., Huang, Y., Piao, J., Gao, C., Xu, F., Jiang, T., and Li, Y. (2025, 16 January). *A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy*. https://arxiv.org/abs/2501.09431 (Access Date, 30 November 2025)

Xu, H., Gan, W., Qi, Z., Wu, J., and Yu, P. S. (2024, 12 May). *Large Language Models for Education: A Survey*. https://arxiv.org/abs/2405.13001 (Access Date, 30 November 2025)

Zhou, J., Zhang, J., Wan, R., Cui, X., Liu, Q., Guo, H., Shi, X., Fu, B., Meng, J., Yue, B., Zhang, Y., and Zhang, Z. (2025, 19 March). Integrating AI into clinical education: Evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. *BMC Medical Education*, *25*(1), 406. https://doi.org/10.1186/s12909-025-06916-2 (Access Date, 30 November 2025)

Zhou, K. Z., Kilhoffer, Z., Sanfilippo, M. R., Underwood, T., Gumusel, E., Wei, M., Choudhry, A., and Xiong, J. (2024, 23 January). *"The teachers are confused as well": A Multiple-Stakeholder Ethics Discussion on Large Language Models in Computing Education*. https://arxiv.org/abs/2401.12453 (Access Date, 30 November 2025)