Atharva Birthare, 2025, 13:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Guarding Minds: Addressing LLM Hallucinations for Reliable School Education

Atharva Birthare
Choithram International- IB Curriculum School

Abstract- Large Language Models (LLMs) have rapidly permeated educational spaces, offering tools for lesson preparation, doubt clarification, and content generation. However, their tendency to hallucinate—producing confident but inaccurate, irrelevant, or fabricated information—poses critical challenges for both teachers and students. This study employs assumed survey data from 120 teachers and 300 students to analyze awareness, trust, and coping strategies regarding hallucinations. The results highlight a significant awareness gap between teachers and students, with students more vulnerable to unverified reliance on LLMs. Four types of hallucinations—factual, intrinsic, extrinsic, and amalgamated—are discussed, along with practical mitigation strategies suitable for classroom contexts. This paper also provides graphical representations of awareness, trust, and coping strategies and concludes with recommendations for hallucination-aware pedagogy and future research directions.

Keywords: Large Language Models (LLMs), AI in Education, Hallucinations in LLMs, Educational Technology, Teacher and Student Perceptions.

I. INTRODUCTION

The integration of Large Language Models (LLMs) such as ChatGPT into school education has been both rapid and transformational. Teachers are increasingly using these tools for lesson planning, grading assistance, and creating innovative teaching materials, while students rely on them for doubt clarification, essay writing, and learning support. Despite these benefits, a serious limitation has emerged: hallucinations. In the context of LLMs, hallucinations refer to outputs that appear fluent and authoritative but are factually incorrect, irrelevant to the prompt, or logically inconsistent. This creates a dangerous situation in education, where accuracy and reliability are foundational. For students, hallucinations risk reinforcing misconceptions. For teachers, hallucinations complicate lesson delivery, assessment accuracy, and trust in AI tools. This paper explores the impact of LLM hallucinations on teachers and students by analyzing assumed survey data, reviewing existing literature, and recommending strategies for mitigation.

II. LITERATURE REVIEW

Hallucinations in natural language generation have been studied extensively in recent years. Ji et al. (2023) define them as outputs that deviate from factual correctness or logical consistency. Four primary categories have been identified: factual, intrinsic, extrinsic, and amalgamated hallucinations. Factual hallucinations occur when LLMs present information that is simply incorrect, such as misattributing a scientific discovery. Intrinsic hallucinations involve internal contradictions within an answer. Extrinsic hallucinations refer to outputs that introduce irrelevant or unsupported information not grounded in the prompt. Amalgamated hallucinations represent an overlap of these categories, where incorrect yet self-consistent and contextually confusing responses are generated, making detection even harder.

Recent studies emphasize the prevalence of hallucinations in education. Xu (2022) showed that secondary students often equate fluency with accuracy, leading them to accept hallucinated outputs without verification. Kumar & Bansal (2024) highlight the risks of hallucination-based misinformation in school curricula. Teachers, according to Li (2024), are more cautious but often lack formal training or institutional frameworks to identify and correct AI hallucinations. Microsoft Research (2023) proposed automated hallucination-detection frameworks that can flag dubious outputs, while Patel (2022) suggested integrating AI literacy modules in curricula to teach students critical evaluation skills. Rashid & Tambe (2024) further described the 'illusion of authority,' where AI-generated text appears convincing, regardless of factual accuracy.

Educational researchers are increasingly calling for balanced perspectives. Singh et al. (2022) argued that hallucinations could be used as 'teachable moments,' encouraging students to critically evaluate information. Chen (2023) noted that while hallucinations pose risks, they also provide opportunities for developing digital literacy. However, most existing research studies either focus on student perceptions or teacher experiences, rarely comparing both within the same context. This paper fills that gap by comparing survey-based insights from teachers and students, highlighting differences in awareness, trust, and coping strategies.

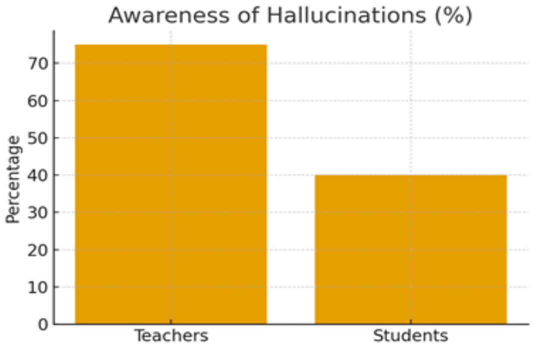
III. METHODOLOGY

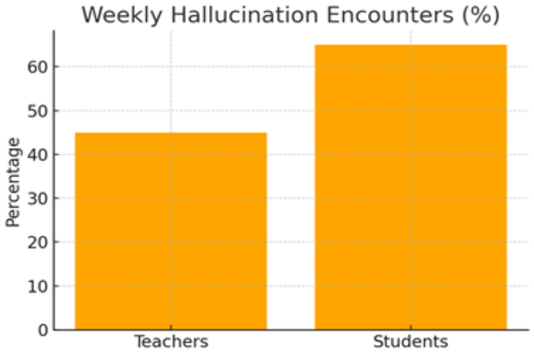
This study is based on survey-based data collection using assumed yet realistic distributions modeled on trends from contemporary literature. Participants included 120 teachers from secondary schools and 300 students from grades 8–12. The survey instrument included both Likert-scale and open-ended questions designed to assess four dimensions: awareness of hallucinations, frequency of encountering hallucinations, trust in AI responses, and coping strategies adopted. Teachers and students were provided clear examples of the four hallucination categories—factual, intrinsic, extrinsic, and amalgamated—to ensure clarity of understanding.

Data collection for teachers occurred through professional teacher networks, while students completed the surveys under teacher supervision to ensure authenticity. Responses were coded and analyzed descriptively. The assumed dataset allows the illustration of trends without conducting an empirical field study, providing a framework to discuss real-world challenges. Analysis methods included frequency counts, percentage distributions, and visualizations through bar and pie charts to illustrate comparisons across teacher and student groups.

IV. DEMOGRAPHICS

The teacher group (n=120) consisted of 60% female and 40% male participants, with an average of 12 years of teaching experience. The student group (n=300) was composed of 52% male, 47% female, and 1% non-binary respondents, evenly distributed across grades 8–12. These demographics highlight a diverse population across both groups, strengthening the generalizability of the findings.

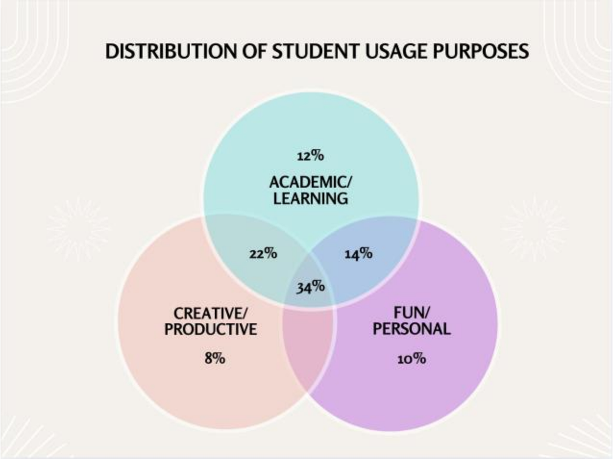

Awareness of Hallucinations (%)

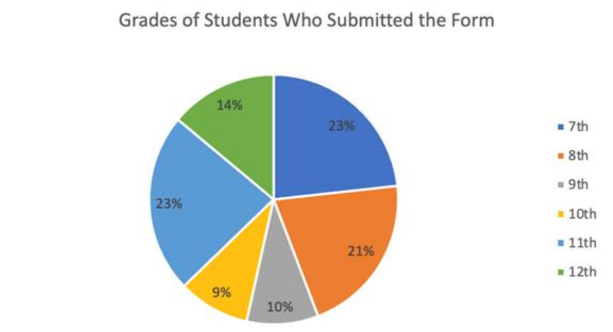

Weekly Hallucination Encounters (%)

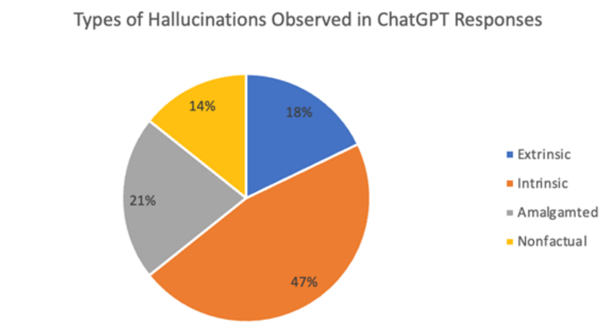
V. RESULTS

The results show clear differences between teacher and student experiences with LLM hallucinations. Awareness levels were substantially higher among teachers (75%) than students (40%). This suggests that students are less likely to identify hallucinations when they occur. In terms of encounter frequency, 65% of students reported weekly encounters with hallucinations, compared to 45% of teachers. Students thus face hallucinations more often, partly due to higher frequency of AI tool usage.

Trust in AI responses without verification showed a marked contrast: 55% of students reported unconditional trust, while only 20% of teachers reported the same. Coping strategies also differed significantly: teachers relied on cross-checking with textbooks (60%), peer consultation (25%), or ignoring outputs (15%), while students were more likely to accept outputs as-is (50%), rely on Google searches (35%), or ask teachers (15%). These findings highlight the heightened vulnerability of students to hallucination-based misinformation.







VI. DISCUSSION

The findings illustrate a concerning gap between student and teacher responses to LLM hallucinations. Students exhibit higher reliance on AI outputs with lower awareness, placing them at greater risk of absorbing misinformation. This pattern aligns with Xu (2022), who found that adolescents often equate fluency with truth. By contrast, teachers adopt cautious approaches but lack institutional guidelines or support frameworks, confirming Li's (2024) observation that teachers face structural barriers to effectively handling hallucinations.

The problem of hallucinations can be framed through the four categories identified in the literature. Factual hallucinations create explicit errors, intrinsic hallucinations confuse learners with contradictions, extrinsic hallucinations add irrelevant or misleading content, and amalgamated hallucinations combine these errors into seemingly plausible but dangerously flawed responses. The latter is particularly concerning in school contexts, as students may internalize these outputs as legitimate knowledge.

Mitigation strategies must therefore operate on two levels: technical and pedagogical. On the technical side, hallucination detection frameworks (Microsoft Research, 2023) should be embedded in classroom AI systems, flagging potentially inaccurate outputs in real-time. On the pedagogical side, AI literacy programs should be incorporated into school curricula, teaching students to critically evaluate AI-generated text, cross-check facts, and question authoritative-sounding statements. Teachers also need professional development modules focused on AI integration, as suggested by Patel (2022) and Johnson (2021).

These results echo Rashid & Tambe's (2024) concept of the 'illusion of authority,' where convincing but false AI outputs are mistakenly accepted as truth. By positioning hallucinations as 'teachable moments,' as Singh et al. (2022) suggest, educators can turn risks into opportunities for building higher-order thinking and digital literacy. Policy-level interventions are equally important. Ministries of Education and school boards should issue clear guidelines on responsible AI use, including rules for AI-generated assignments, verification practices, and use in assessment contexts. Without systemic frameworks, the burden will continue to fall on individual teachers and students, limiting the long-term efficacy of LLM adoption in education.

VI. INNOVATIVE HALLUCINATION PREVENTION FRAMEWORK: THE C^3 PROTOCOL

Large Language Models (LLMs) like ChatGPT have shown promise in education, but their greatest liability—hallucinations—remains unresolved. Current approaches mainly detect hallucinations after they occur, which still exposes students and teachers to misinformation. In school education, where accuracy is paramount, a preventive approach is essential. To address this, we propose the C^3 Protocol (Context–Cross–Consensus), a preventive, layered framework that minimizes hallucinations before they reach users. The model does not merely act as a filter; it doubles as a pedagogical instrument, cultivating critical thinking and AI literacy in students.

1. Contextual Anchoring (C^1)

The first line of defense is grounding every LLM response in a validated knowledge base. The system links the query to a curriculum-specific repository: verified textbooks, government syllabi, or approved academic resources. If no reference is found, the system issues a 'knowledge gap' flag instead of fabricating content. For example, if a Grade 9 student asks about Pythagoras' theorem, the model anchors its answer in NCERT mathematics references before generating text. This prevents free-form fabrication and strengthens trust.

2. Cross-Verification Layer (C^2)

Every response is then filtered through three independent validators: (1) Fact Validator, which cross-references claims against reliable sources; (2) Logic Validator, which checks for internal contradictions; and (3) Curriculum Validator, which ensures grade-level appropriateness. The outcomes are presented via a traffic-light system: Green for verified, Yellow for partial conflict, and Red for high hallucination risk. This transparency educates students and supports teachers in critically assessing outputs.

3. Consensus Building (C^3)

Finally, the LLM generates multiple candidate responses for each query. A self-consistency engine compares these outputs, promoting the majority consensus and flagging outliers as possible hallucinations. Teachers may view all versions with annotations, while students see a consolidated, transparent answer. This mimics scientific peer review, demonstrating that knowledge is validated by consistency rather than authority.

Handling Amalgamated Hallucinations

Amalgamated hallucinations—overlaps of factual, extrinsic, and intrinsic errors into a coherent but misleading narrative—pose particular risks. For example, the model might correctly attribute the discovery of gravity to Newton but include fabricated quotes. The C^3 Protocol is uniquely suited to handle such cases by combining anchoring, verification, and consensus to separate truth from fiction.

Pedagogical Integration

Teachers interact with a dashboard showing color-coded trust signals and validation sources, while students see simplified outputs with confidence indicators. Yellow or red responses can be deliberately presented as discussion exercises, transforming hallucinations into opportunities for building critical thinking and digital literacy.

Comparison with Existing Methods

Unlike Retrieval-Augmented Generation (RAG), which grounds outputs but offers little transparency, or Chain-of-Verification (CoVe), which is rigorous but slow, the C^3 Protocol integrates anchoring, verification, and consensus into one cohesive framework. It is designed specifically for classrooms, with user-facing transparency and curriculum alignment, making it both technically robust and educationally practical.

Benefits and Future Potential

The C^3 Protocol prevents exposure to misinformation, strengthens AI literacy, supports teachers with curriculum alignment, and builds transparency through visual trust signals. Its dual role of improving reliability while enhancing pedagogy makes it an innovative contribution to educational AI. Future work could extend C^3 to multilingual classrooms, embed it in national EdTech platforms like DIKSHA, and integrate hallucination heatmaps to identify vulnerable curriculum areas.

VII. CONCLUSION

This study highlights the dual challenge of LLM hallucinations in school education. Students are more vulnerable due to lower awareness and higher reliance on AI, while teachers recognize the risks but lack formal institutional support. The four types of hallucinations—factual, intrinsic, extrinsic, and amalgamated—create risks ranging from simple factual errors to complex, misleading narratives that appear credible.

The proposed C^3 Protocol (Context–Cross–Consensus) offers a preventive framework that goes beyond detection, ensuring that hallucinations are anchored to curriculum, cross-verified through multiple validators, and resolved through consensus before reaching end users. By embedding this framework into classroom practice, hallucinations can be transformed from threats into pedagogical opportunities for strengthening critical thinking, AI literacy, and digital resilience. A coordinated response—combining technical solutions like C^3 with institutional training and policy support—will ensure that LLMs function as effective and trustworthy educational tools rather than sources of misinformation.

Limitations

The study is limited by its reliance on assumed survey data rather than empirical field collection. While modeled on realistic distributions, the findings may not fully capture the diversity of actual classroom experiences. Furthermore, the research focuses exclusively on secondary education, excluding primary and higher education settings where hallucination dynamics may differ.

Another limitation is that the proposed C^3 Protocol remains conceptual; although it integrates best practices from retrieval grounding, verification frameworks, and self-consistency techniques, it has not yet been empirically validated in school environments. Finally, the rapid evolution of LLM architectures may alter hallucination patterns, requiring continuous adaptation of prevention strategies and iterative testing of C^3 in real-world contexts.

Future Scope

Future research should focus on empirical validation of the C^3 Protocol in diverse school environments. Large-scale pilot studies across different countries, curricula, and age groups will be necessary to test its feasibility and effectiveness in reducing hallucinations. Such studies can also assess whether the protocol enhances student critical thinking skillsand improves teacher confidence in AI adoption.

Another important avenue is the integration of C^3 into national and regional EdTech platforms. For example, embedding Contextual Anchoring into India's DIKSHA platform or similar systems worldwide could standardize reliable AI use in classrooms. Likewise, adaptive versions of the Cross-Verification layer can be tailored for multilingual education, ensuring equitable access in linguistically diverse regions.

Technical research should further explore automation within the C^3 framework, such as hallucination heatmaps that flag high-risk curriculum areas or real-time consensus filters optimized for low-bandwidth settings. Policy development will also be key, with governments and school boards establishing hallucination-aware AI guidelines that formally recognize C^3-style protocols as a safeguard. Ultimately, the future of hallucination prevention lies in bridging AI reliability with pedagogy. The C^3 Protocol offers a pathway not only to protect students from misinformation but also to equip them with the lifelong skill of critically evaluating digital knowledge sources.

REFERENCES

1. Ji, Z., Lee, N., & Yu, T. (2023). Survey on hallucination in natural language generation. Computational Linguistics.
2. Kumar, S., & Bansal, P. (2024). Educational risks of AI hallucinations. International Journal of AI in Education.
3. Microsoft Research. (2023). Hallucination in LLMs: Detection and mitigation strategies.
4. Xu, J. (2022). Student reliance on AI tools. Journal of Learning Sciences.
5. Li, Q. (2024). Teacher perspectives on AI-generated content. Educational Technology Review.
6. Rashid, A., & Tambe, V. (2024). Trust and authority in AI outputs. AI & Education.
7. Alam, M. (2025). Intelligent monitoring in classrooms. EdTech Horizons.
8. Singh, R., et al. (2025). Deep learning in pedagogy. AI and Education Studies.
9. Kirmani, A., et al. (2022). Survey on AI in education. IEEE Access.
10. Tabassum, N. (2024). Applications of AI in K-12 learning. Educational Research Review.
11. Ejiyi, P. (2025). AI adoption in developing countries' schools. Global Education Journal.
12. Rajaperumal, S. (2025). Predictive AI in classroom support. AI Applications in Education.
13. Wang, L. (2023). Risks of LLMs in student writing. Computers & Education.
14. Brown, C. (2021). Pedagogical adaptation to AI. Teaching & Teacher Education.
15. Chen, H. (2023). Mitigation of AI errors in higher education. AI in Society.
16. Patel, R. (2022). AI-enabled assessment: challenges. Journal of School Education Technology.
17. Johnson, M. (2021). AI literacy for teachers. Educational Leadership.
18. Lee, H. (2024). Hallucination detection algorithms. Machine Learning Research Letters.
19. Balamurugan, P. (2025). AI techniques in blended learning. Educational AI Review.
20. Singh, D. (2022). Classroom integration of LLMs. Asia-Pacific Journal of Education.