Mitigating Hallucination in Large Language Models: Techniques, Applications, and Implications

Mahadev Dhanaji Limbuche[1]

[1]PDEA's Mamasaheb Mohol College, Paud Road, Pune

Publication Date: 2026/01/30

Abstract: Large Language Models (LLMs) such as GPT, LLaMA, and PaLM have transformed the field of Natural Language Processing (NLP) by achieving remarkable results in text generation, summarization, translation, question answering, and dialogue systems. Their wide adoption across industries highlights their usefulness but also exposes a critical limitation—hallucination. Hallucination occurs when models generate information that is false, misleading, or fabricated. These errors can vary from small factual mistakes, like incorrect dates or figures, to serious inaccuracies that may cause harm in sensitive areas such as healthcare, education, and software development. This paper explores the concept and classification of hallucinations in LLMs, examines techniques to reduce them—including prompt engineering, fine-tuning, and Retrieval-Augmented Generation (RAG)—and discusses ethical implications and real-world applications. By comparing multiple strategies, the study aims to contribute to developing more reliable and trustworthy AI systems.

Keywords: *Large Language Models, Hallucination, Prompt Engineering, Fine-Tuning, Retrieval-Augmented Generation, NLP, AI Ethics, Factual Consistency.*

How to Cite: Mahadev Dhanaji Limbuche (2026) Mitigating Hallucination in Large Language Models: Techniques, Applications, and Implications. *International Journal of Innovative Science and Research Technology,* 11(1), 2271-2276. https://doi.org/10.38124/ijisrt/26jan1060

I. INTRODUCTION

Large Language Models (LLMs) have brought a noticeable improvement in how artificial intelligence systems understand and generate human language. Built on deep learning architectures, especially transformers, these models can understand and generate human-like text by learning from massive amounts of data. Models like GPT-4, LLaMA, and PaLM have set new standards in NLP tasks such as summarization, translation, question answering, and conversational AI.

However, despite their impressive abilities, LLMs are not always accurate. They sometimes produce hallucinations—responses that sound plausible but are factually incorrect or entirely made up. These errors often occur due to the probabilistic nature of text generation, gaps in the training data, or the lack of real-time access to verified information.

➢ *Examples of Hallucinations:*

- Education: An AI tutor might give a wrong historical date or explain a scientific concept incorrectly.
- Healthcare: A chatbot could provide inaccurate medical advice, posing serious risks to patients.
- Software Development: Code assistants may suggest syntactically incorrect or insecure code.

These examples show the urgent need to understand and control hallucinations to ensure that AI systems are safe, ethical, and reliable.

II. LITERATURE REVIEW

As LLMs continue to evolve, research on hallucination and its mitigation has grown rapidly. Several methods have been proposed to reduce these issues.

Table 1 Types of Hallucinations in Large Language Models

Hallucination Type	Definition	Key Problem
Intrinsic Hallucination	The generated text directly contradicts the source information or context given in the input.	The model ignores or twists the facts it was just given.
Extrinsic Hallucination	The generated text is factually incorrect according to real-world knowledge, even if the input prompt did not provide the correct information.	The model makes up information that is not in the source and is false in reality.

➢ *Additionally, Hallucinations can be Classified by the Type of Error:*

- Factuality: The error involves a concrete, verifiable fact (like a date, name, or figure) that is wrong in the real world.
- Faithfulness: The error involves being unfaithful to the source material, often happening in tasks like summarization where the model invents details not present in the original text.

➢ *Prompt Engineering*

Prompt engineering focuses on crafting better input prompts to guide LLMs toward more accurate outputs.

- *Common Techniques Include:*

✓ Zero-shot prompting: Zero-shot prompting means asking a question to the model without giving any example beforehand.
✓ Few-shot prompting: Providing a few examples to set the context.
✓ Example: For users who do not have a technical background, prompt engineering can simply be seen as the way we ask or frame questions so that the AI gives a better and clearer response.
✓ Chain-of-thought prompting: Encouraging step-by-step reasoning to improve factual accuracy.

Studies such as Bang et al. (2023) have shown that well-structured prompts can significantly reduce hallucinations, especially in conversational tasks.

➢ *Fine-Tuning*

Fine-tuning involves training a pre-trained model on verified, domain-specific data to improve accuracy in that field.

- *For Example:*

✓ Medical domain: Models fine-tuned on PubMed data make fewer medical errors.
✓ Legal domain: Models trained on legal documents provide more reliable case references.

Fine-tuning helps models internalize correct domain knowledge, reducing false or fabricated responses.

“In simple terms, fine-tuning is similar to training a graduate for a specific profession. While basic education provides general knowledge, specialized training improves accuracy and performance in a specific domain.”

➢ *Retrieval-Augmented Generation (RAG)*

Retrieval-Augmented Generation (RAG) works by connecting a language model with external sources of information to produce more accurate answers.

The model first retrieves relevant documents and then uses that information to generate answers.

- *Benefits Include:*

✓ Better factual grounding for complex questions.
✓ Consistency across multi-step reasoning tasks.
✓ Lewis et al. (2020) demonstrated that RAG improves the factual accuracy of open-domain question answering systems.

➢ *Evaluation and Ethical Challenges*

Despite progress, challenges remain:

- Evaluation difficulties: Metrics like BLEU and ROUGE don’t fully measure factual accuracy.
- Ethical concerns: Many studies overlook the social implications of hallucinations.
- Cross-domain limitations: Methods that work in one domain may fail in others.

Overall, the literature emphasizes the need for comprehensive evaluation and multi-faceted mitigation strategies.

➢ *Research Gaps*

Even though we have a lot of great methods to try and fix hallucination in Large Language Models, there are still some huge hurdles we haven't cleared yet.

One of the biggest issues is that our evaluation tools are a mess. Current benchmarks are inconsistent and often fail to really tell the difference between one kind of hallucination and another. This makes it incredibly hard for researchers to accurately compare the performance of different models and know which technique truly works best.

Even our promising Retrieval-Augmented Systems (RAG) aren't perfect. They still run into problems when they retrieve bad information or have "noisy" context, meaning the system still ends up fabricating some of its output.

On top of that, most of our best mitigation tricks just don't travel well. They tend to work great on one model size or one specific language but completely fall apart when you switch to another language, especially in settings that have fewer resources. And let's not forget multimodal hallucinations—the weird, false stuff generated by AI models that combine vision and language. We've barely scratched the surface on understanding how or why those happen.

Ultimately, we have three big things to fix: We desperately need to agree on standardized ways to measure the problem, dig deeper to understand the mechanical reasons *why* models hallucinate in the first place, and develop mitigation strategies that are tough and reliable enough to work everywhere.

➢ *Objectives*

The main objectives of this study are to:

- Classify types of hallucinations—ranging from minor factual errors to major fabrications.
- Evaluate and compare mitigation techniques like prompt engineering, fine-tuning, and RAG.
- Assess the impact of hallucinations in domains such as education, healthcare, and software development.
- Analyze ethical and social implications to promote trustworthy AI.

III. METHODOLOGY

This study uses a survey-based research methodology to analyze how users experience hallucinations in Large Language Models (LLMs) and how these hallucinations affect trust, usage patterns, and expectations.

➢ *Research Approach*

A quantitative survey method was chosen as the primary research approach. This method helps gather real-world user experiences, perceptions, and opinions regarding incorrect or misleading outputs generated by AI models such as ChatGPT, Gemini, and Copilot.

➢ *Survey Design*

A structured questionnaire was created using Google Forms. The questions were designed to capture:

- User familiarity with AI tools
- Awareness of AI hallucinations
- Frequency and domains of hallucination encounters
- Impact on user trust
- Preferences for reducing hallucinations and improving AI reliability

➢ *Participants*

The survey targeted individuals who actively use AI tools, including:

- Students (UG/PG)
- Software developers
- Educators
- General AI users

➢ *Data Collection*

Data collection was carried out through a Google Form shared with participants online. The form included only multiple-choice questions, and all questions were marked as required. This ensured that every participant completed all parts of the survey. Using MCQs made the survey quick to answer and helped collect consistent, structured data, which is easier to compare, calculate, and analyze statistically.

➢ *Data Analysis*

Collected responses will be analyzed based on:

- Frequency counts (e.g., how often hallucinations occur)
- Percentage distributions (e.g., trust levels before and after hallucinations)
- Cross-domain insights (e.g., education vs. coding vs. healthcare)
- Patterns in user preferences for mitigation techniques

The analysis results will be used in the Findings and Discussion section to support the research conclusions.

IV. RESULT

➢ *Based on Prior Research:*

- Most users are aware of hallucinations but may not fully understand their causes.
- Hallucinations are expected to appear frequently in education, coding, and general knowledge queries.
- Users are likely to report reduced trust after encountering hallucinations.
- Users are expected to prefer improvements such as citations, verified information, and clarity features.

In this section, we break down the results from our quantitative survey, where we asked 122 people—a mix of students, developers, and general AI users—about their real-world experiences with AI hallucination. The data confirmed our suspicions about the problem's scope and gave us a clear, user-defined direction for building better AI systems.

➢ *First Things First: How Aware are People?*

We wanted to establish user familiarity with the core topic by asking, "Have you heard the term 'AI hallucination' before?" The responses were split, confirming a visible knowledge gap in the general user base:

- Almost half of the respondents (45.9\%) said "Yes," they were familiar with the term.
- However, a significant portion (38.5\%) said "No."
- The remaining 15.6\% were "Maybe" or uncertain.
- This indicates that while the problem of receiving incorrect AI output is widespread, the formal technical term is not yet universally known. This finding highlights the need for transparent communication from AI developers about the models' limitations.

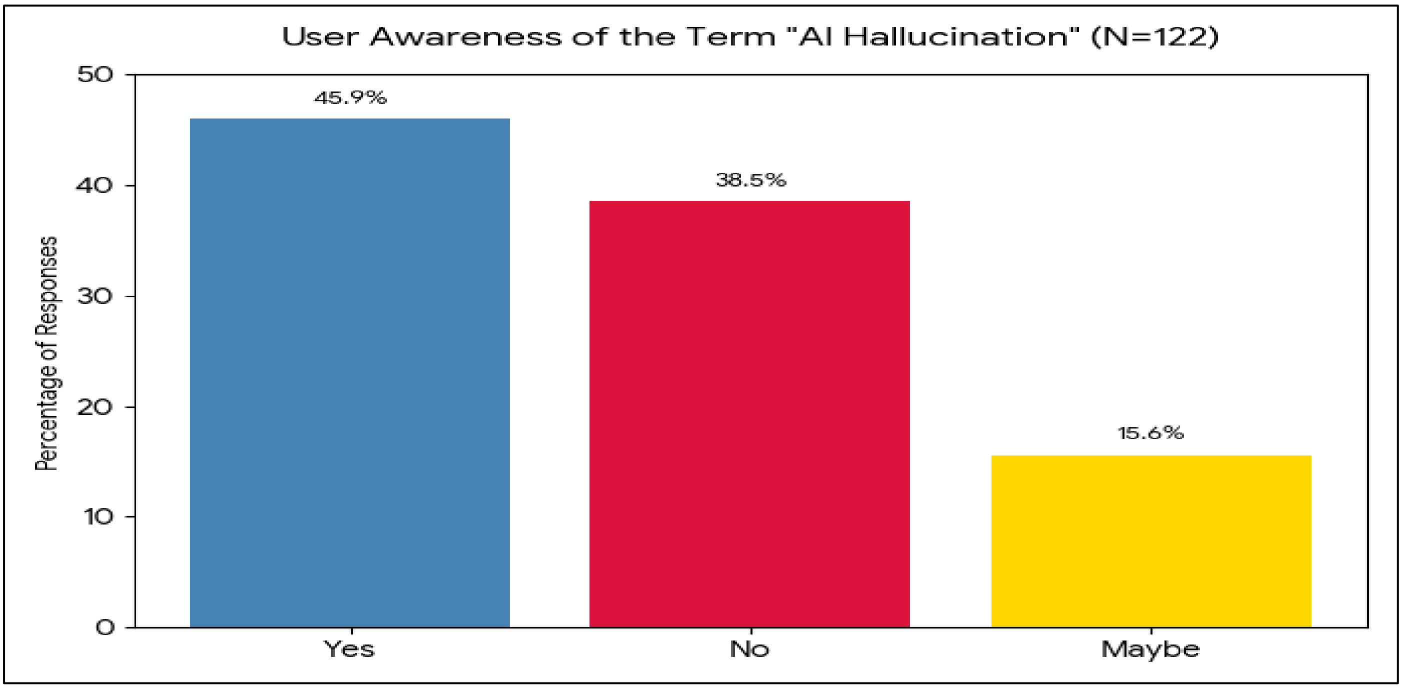

Fig 1 User Awareness of the Term “AI Hallucination” (N=122)

➢ *Where Hallucination Hits Hardest*

We asked users, "In which areas have you seen incorrect AI responses?" to quantify the problem across different domains. The answers were highly specific and validated our focus on applications where factual accuracy is critical. The results show that the problem is concentrated in professional and academic fields:

- Education was the top area, with 54.1\% of respondents reporting errors.
- Healthcare followed closely at 44.3\%.
- Coding was also high, with 41.8\% reporting faulty suggestions.

These numbers clearly show that the problem isn't just about general conversation; it’s happening most frequently in domains where accuracy is non-negotiable. This confirms that mitigation efforts must be prioritized for these high-stakes applications.

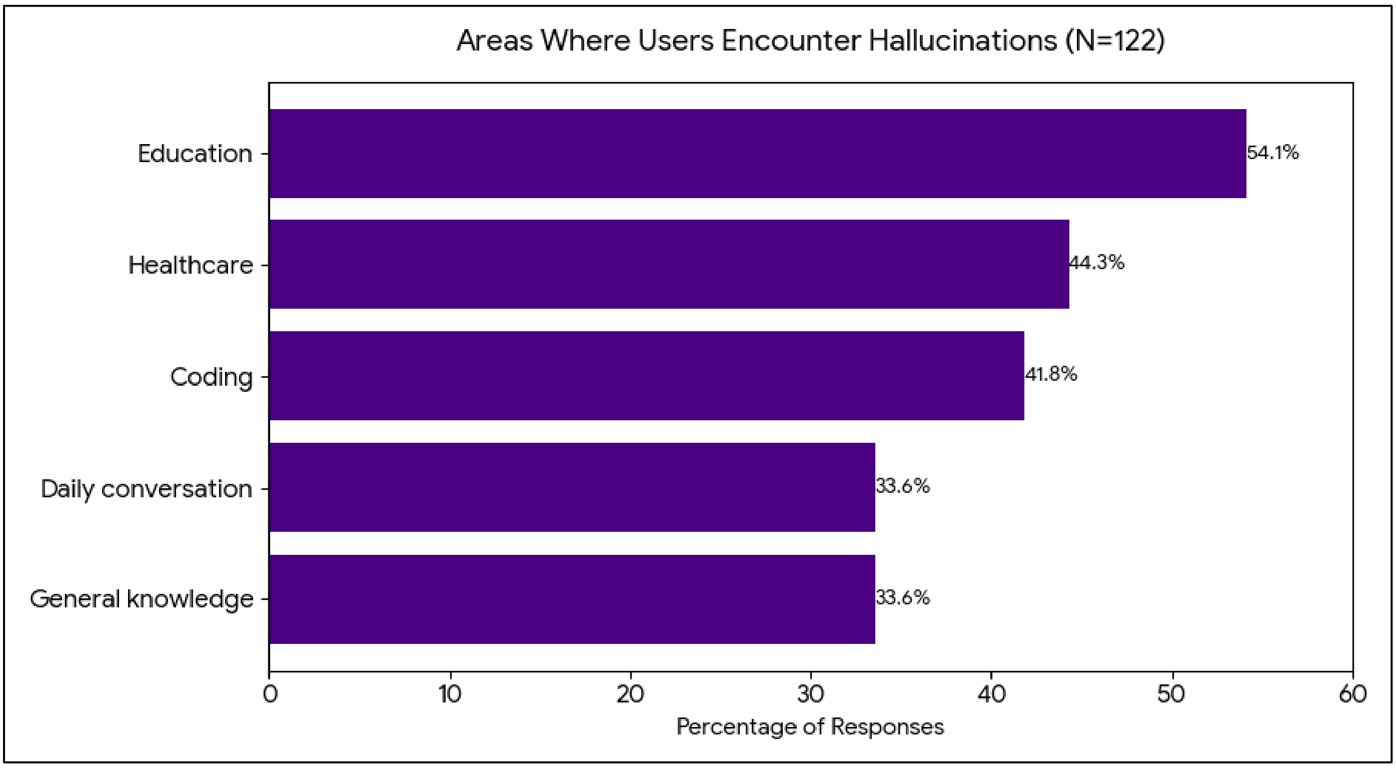

Fig 2 Areas Where Users Encounter Hallucination (N=122)

➢ *The Trust Factor*

One of the most critical objectives was to measure the impact of hallucination on user trust by asking, "Did incorrect AI responses affect your trust?" The results confirmed the expected impact, yet also revealed a high degree of user resilience:

- A notable 30.3\% of users said "Yes," their trust was directly reduced.
- Another 25.4\% were "Maybe" (uncertain about the full impact).
- The largest single group, 44.3\%, said "No," their trust was not affected.

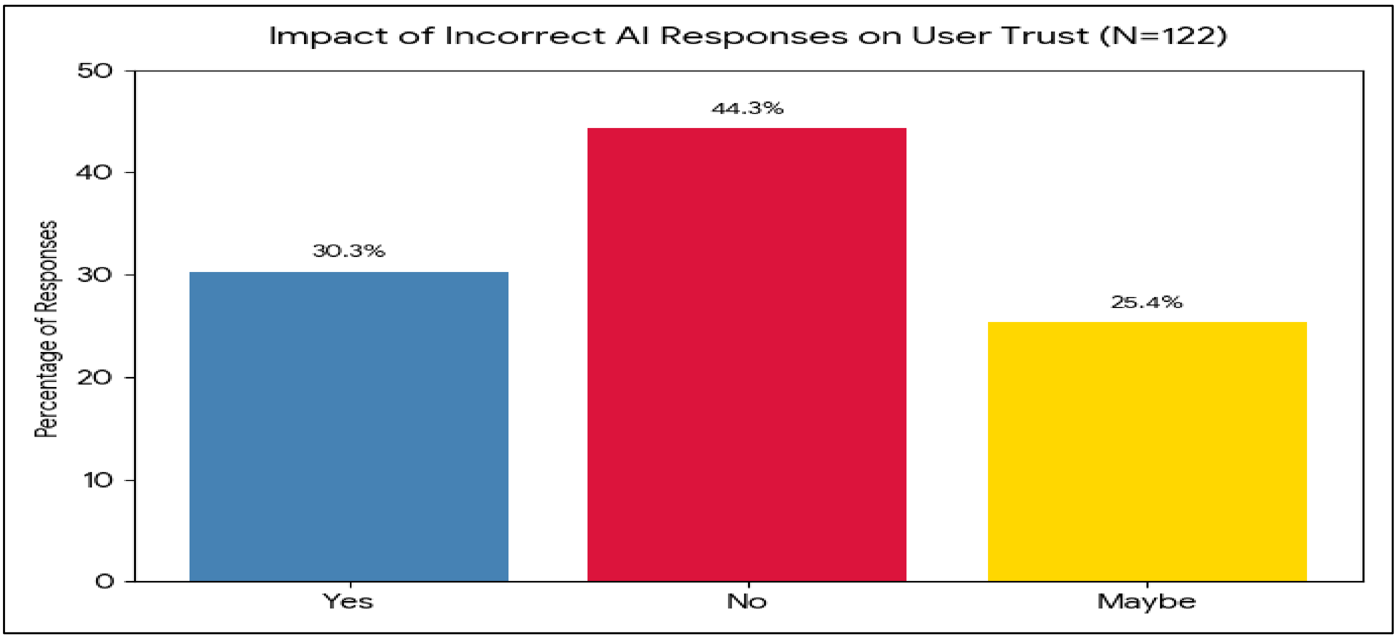

Fig 3 Impact of Incorrect AI Responses on User Trust (N=122)

Despite the high frequency of errors reported, a significant portion of users have maintained their trust. However, the fact that over half of the respondents (55.7\% total of "Yes" and "Maybe") felt negatively affected underscores that the erosion of user trust is a real and present danger that must be addressed to ensure long-term reliance on AI systems.

➢ *The User-Defined Solution Blueprint*

To gain practical direction for mitigation, we asked users: "Which features would increase your trust in AI?" Their answers provide a clear blueprint for developers:

- The number one request was simple: "More accurate responses" (42.6\%). This is the obvious, ultimate goal of mitigation.
- This was closely followed by a demand for greater transparency, with "Asking clarifying questions" (41.8\%) being highly valued.
- Users also emphasized structural proof, requesting "Showing sources/citations" (31.1\%) and using "Verified data" (25.4\%).

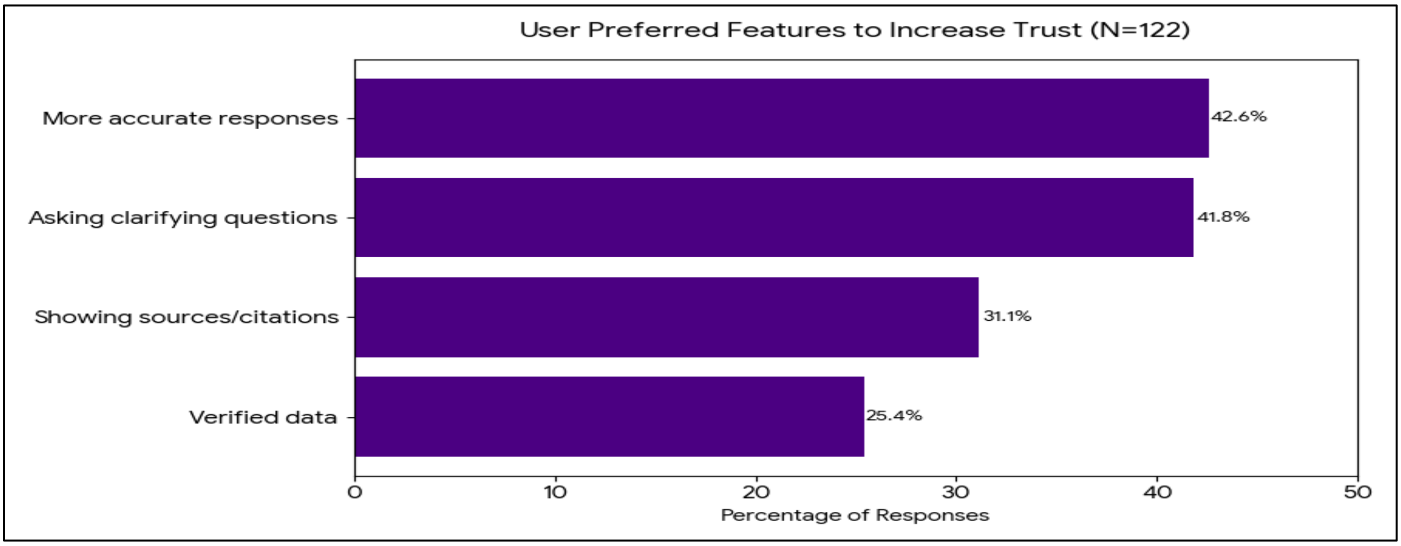

Fig 4 User Preferred Features to Increase Trust (N=122)

These findings confirm that users are looking for traceability and transparency. Their preference for citations and verified data directly supports the adoption of fact-grounding techniques like Retrieval-Augmented Generation (RAG) as the most effective path forward for building trustworthy AI.

V. APPLICATIONS

➢ *Education*

- AI tutors can provide accurate, verified information to students.
- Learning platforms can adapt lessons based on fact-checked data.
- Example: An AI explaining historical events with verified timelines.

➢ *Healthcare*

- Clinical decision systems can offer accurate medical guidance.
- Research summarizers can condense studies without introducing errors.
- Example: A medical assistant tool generating evidence-based treatment suggestions.

➢ *Software Development*

- AI tools can provide reliable coding assistance.
- Documentation generators can ensure technical accuracy.
- Example: AI recommending secure coding practices from verified sources.

VI. ETHICAL AND SOCIAL IMPLICATIONS

➢ *Reducing Hallucinations is Essential for Responsible AI use. Key Benefits Include:*

- Misinformation control: Prevents spread of false data.
- Bias reduction: Limit's reinforcement of harmful stereotypes.
- Trust building: Users gain confidence in AI systems.
- Data integrity: Ensures sensitive information is not fabricated.

Ethical deployment requires continuous monitoring, transparency, and user awareness.

VII. CONCLUSION AND FUTURE WORK

This study investigated user awareness, experiences, and trust-related impacts of hallucinations in Large Language Models (LLMs) through a structured online survey. The survey results highlight those hallucinations are commonly observed across multiple domains such as education, coding, and general knowledge.

Many users reported a decrease in trust after encountering incorrect or misleading responses, showing that hallucinations directly affect the reliability of AI systems.

The findings suggest that users prefer mitigation features like citations, verified information sources, and clarification prompts to improve accuracy. These insights emphasize the need for safer and more transparent AI models. The study concludes that understanding user experiences is essential for designing better strategies to reduce hallucinations and improve the overall trustworthiness of LLMs.

➢ *Future Work Future Research can Involve:*

- Collecting larger and more diverse user samples
- Comparing hallucination rates across different LLM tools
- Developing evaluation frameworks specifically for hallucination detection
- Testing the effectiveness of mitigation techniques based on user feedback

By focusing on real-world user experiences, this research contributes to building more dependable and responsible AI systems.

REFERENCES

[1]. Open AI. (2023). GPT-4 Technical Report. Open AI.
[2]. Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
[3]. Xu, W., et al. (2024). Hallucination in Large Language Models: A Survey. Journal of Artificial Intelligence Research.
[4]. Lewis, M., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.
[5]. Bang, Y., et al. (2023). Multitask Prompted Training Enables Zero-Shot Task Generalization. ICLR.
[6]. Zellers, R., et al. (2019). Defending Against Neural Fake News. NeurIPS.
[7]. Bhagavatula, C., et al. (2020). Abstractive Summarization with Faithfulness Constraints. ACL.
[8]. Roller, S., et al. (2021). Recipes for Building an Open-Domain Chatbot. arXiv:2004.13637.