



AI-Based Detection of Hallucinated Outputs in Chatbots Using Retrieval-Augmented Generation: A Review and Conceptual Framework

¹Bhor Vaishnavi S., ²Prof.Bangar Abhishek P., ³Prof. Karad Vandana A., ⁴Prof. Bhosale Sachin B.

^{1,2,3,4,5}Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India

Abstract. Chatbots based on Large Language Models (LLMs) can provide quick, fluent, and informative answers, but they sometimes generate incorrect or unsupported information known as hallucinations. These hallucinations can reduce user trust and create problems in areas such as education, healthcare, research, and customer support. Therefore, detecting and preventing hallucinated outputs is important for developing reliable and trustworthy AI systems. Retrieval-Augmented Generation (RAG) is a useful approach that retrieves relevant information from trusted sources before generating an answer. This retrieved information can be used to check whether the chatbot's response is accurate and supported by available evidence. However, RAG alone cannot completely prevent hallucinations because the generated answer may still contain incorrect facts, irrelevant information, or unsupported statements. This paper proposes a conceptual framework for detecting hallucinations in chatbot responses by combining information retrieval, evidence-based verification, and rule-based validation. The system compares the generated answer with retrieved information and identifies statements that are incorrect or lack sufficient supporting evidence. Based on the verification results, the system classifies the response as reliable or flagged. If the response contains unsupported information, the system can suggest corrections or prevent the answer from reaching the user without proper verification. The proposed approach aims to improve answer accuracy, reduce misleading information, and increase transparency in AI-based conversational systems. Although this framework offers a promising direction, its effectiveness depends on the quality of retrieved information, verification methods, and decision thresholds. Future evaluation should measure detection accuracy, response time, and computational cost. This approach can support the development of safer, more reliable, and trustworthy chatbots across different application areas.

Keywords: Chatbots, hallucination detection, retrieval-augmented generation, large language models, answer verification, factual grounding

I. Introduction

Chatbots powered by Large Language Models (LLMs) are widely used in different fields, such as customer support, education, healthcare, industrial maintenance, and information search. These systems can understand user questions and generate human-like responses within a short time. They help users find information, solve problems, and complete tasks more efficiently. However, despite their ability to generate fluent and meaningful answers, LLM-based chatbots have an important limitation: they can produce incorrect, misleading, or unsupported information. This problem is known as hallucination. A chatbot may



provide an incorrect fact, generate a false reference, invent a numerical value, or give an answer that contradicts the information provided to it. Such responses may appear completely correct, making it difficult for users to identify mistakes. Therefore, detecting hallucinations is essential for improving the reliability, accuracy, and trustworthiness of AI-based conversational systems.

Retrieval-Augmented Generation (RAG) is an approach that combines information retrieval with text generation. Instead of depending only on the knowledge learned during training, a RAG-based system retrieves relevant information from external documents, databases, or other trusted sources before generating an answer. The retrieved information provides supporting evidence that can help the language model produce more accurate and relevant responses. It also provides a reference that can be used to check whether the generated answer is consistent with the available information. However, RAG does not completely eliminate hallucinations. A chatbot may misunderstand the retrieved information, combine facts incorrectly, or generate statements that are not supported by the retrieved passages. Therefore, an additional verification mechanism is needed to identify unsupported or incorrect outputs.

Hallucination detection focuses on checking the generated response against reliable evidence and identifying statements that may be incorrect or lack sufficient support. Different techniques can be used for this purpose, including semantic similarity checking, evidence-based verification, rule-based validation, and confidence thresholding. These techniques help determine whether an answer can be accepted, requires correction, or should be flagged for further verification. Combining these methods can provide a more systematic way to detect unreliable responses before they reach users.

This paper examines existing approaches to hallucination detection using RAG and related verification techniques. It compares the main ideas, methods, advantages, and limitations of the selected studies to understand how different detection mechanisms can improve chatbot reliability. Based on this analysis, a conceptual framework is proposed that combines relevant information retrieval, answer verification, and decision-making mechanisms. The framework aims to identify unsupported information and classify chatbot responses as reliable or flagged. It also highlights important challenges, including verification accuracy, response time, computational cost, and the quality of retrieved information. The proposed framework provides a foundation for future research and experimental evaluation toward developing more reliable and trustworthy conversational AI systems.

II. Background

1. Hallucination in Chatbot Outputs

Hallucination is one of the major challenges in chatbots based on Large Language Models (LLMs). It occurs when a chatbot generates information that is incorrect, misleading, unsupported by available evidence, or inconsistent with the given context. These responses may sound natural and convincing, even when they contain false information. For example, a chatbot may provide an incorrect date, generate a false reference, mention facts that are not present in the source document, or give an answer that contradicts trusted information. Such errors can affect the reliability of AI systems, especially in applications where users depend on accurate information for learning, decision-making, and problem-solving.

Hallucinations can occur for several reasons, including limitations in the model's training data, incomplete knowledge, unclear user questions, and incorrect interpretation of the available context. Even when relevant information is provided, the model may generate statements that are not directly supported by the source. Therefore, generating an answer is not enough; the system must also verify whether the answer is supported by reliable evidence.

The review in [5] groups hallucination mitigation methods into four main architectural families: unaugmented, post-hoc reactive, proactive, and integrated approaches. Unaugmented approaches depend mainly on the language model without adding a



separate verification mechanism. Post-hoc reactive approaches examine the generated answer after it has been produced and identify possible errors or unsupported statements. Proactive approaches attempt to reduce hallucinations before or during answer generation by guiding the model toward reliable information. Integrated approaches combine prevention and detection mechanisms to improve the overall reliability of the response.

This classification is important because different approaches detect and manage hallucinations at different stages. A detector must also avoid incorrectly identifying correct information as false. The review identifies seven possible output scenarios, ranging from correctly retaining an accurate answer to incorrectly changing a correct answer during the correction process. This highlights the importance of reducing both false positives and false negatives. A good detection system should identify unsupported information without unnecessarily rejecting or modifying correct responses.

2. RAG as a Reference for Detection

Retrieval-Augmented Generation (RAG) combines information retrieval with the text-generation capabilities of Large Language Models. In a standard RAG pipeline, the user submits a question, and the retrieval component searches an external knowledge source for relevant information. The retrieved passages are then provided to the language model as context, helping it generate an answer based on the available evidence rather than relying entirely on its internal knowledge.

In hallucination detection, the retrieved passages serve as reference material for verifying the generated response. After the answer is produced, the system compares its statements with the retrieved information to determine whether the answer is supported by the available evidence. Statements that contradict the retrieved passages or lack sufficient supporting information can be identified as potential hallucinations. Depending on the verification results, the system may accept the answer, flag questionable statements, request further retrieval, or generate a corrected response.

The effectiveness of this process depends on the quality of the retrieval component. Keyword-based retrieval searches for matching words, while dense retrieval uses semantic representations to identify passages with similar meanings. Hybrid retrieval combines keyword-based and semantic search to improve the chances of finding relevant information. Some systems also use reranking methods to reorder retrieved passages according to their relevance to the user's question. These techniques can improve the quality of evidence available for verification.

Another important factor is the amount and quality of context provided to the language model. If the retrieved passages are incomplete, outdated, irrelevant, or contradictory, the generated answer may still contain unsupported information. Similarly, passing too much text to the model may increase computational cost and make important evidence harder to identify. Therefore, effective RAG-based hallucination detection requires relevant retrieval, appropriate context selection, and reliable answer verification. Combining these components provides a useful foundation for developing chatbots that can identify potentially unreliable responses and improve the trustworthiness of generated information.

III. Review of the Selected Studies

1. Query-aware retrieval

Study [1] presents an adaptive Retrieval-Augmented Generation (RAG) approach that selects different retrieval strategies according to the type of question asked by the user. The study classifies queries into three categories: factual questions, definition-based questions, and reasoning-based questions. Each category requires a different type of information retrieval to provide suitable evidence for generating an answer. For example, a factual question may require a specific piece of information, whereas a reasoning question may require information collected from multiple sources. By selecting an appropriate retrieval strategy, the system attempts to improve the relevance of the retrieved information.



The study also uses context compression to reduce the amount of text passed to the language model. This helps remove unnecessary information while retaining important evidence. The query classifier achieved 92% accuracy. Experiments conducted on the HotpotQA, 2WikiMultiHopQA, and MuSiQue datasets showed a reduction of approximately 38% to 58% in the number of tokens, with only a one-to-two-point decrease in exact-match performance. These results suggest that adaptive retrieval can reduce computational requirements while maintaining reasonable answer quality.

For hallucination detection, this approach is useful because the quality of verification depends on the quality of the retrieved evidence. If the system retrieves relevant and sufficient information, it becomes easier to identify unsupported statements in the generated response. The study's ablation results suggest that adaptive retrieval contributes substantially to the observed quality improvements. However, the approach increased response latency to approximately 27.8 seconds compared with 9.6 seconds for the baseline system. This additional delay may affect real-time chatbot applications. Therefore, query-aware retrieval provides a useful foundation for evidence-based hallucination detection, but its retrieval quality and response time must be considered together.

2. Schema and Rule Validation

Study [2] investigates the use of structured output formats, predefined rules, and Retrieval-Augmented Generation to improve the reliability of AI-generated responses in industrial applications. The research discusses three industrial projects, including a helicopter maintenance simulator, where incorrect information can affect the usefulness and reliability of system outputs. In such applications, responses must not only contain correct information but also follow specific formats and operational requirements.

The study compares plain prompting, schema-guided prompting, and RAG combined with schema validation. In a 50-run evaluation using Qwen2.5-32B and GPT-4.1, plain prompting achieved reported factual accuracy values of 0.176 and 0.026, respectively. Schema-guided prompting improved these values to 0.28 and 0.50. According to the reported results, combining RAG with schema validation achieved complete factual accuracy and compliance in the evaluated setting.

Schema validation checks whether the generated response follows a predefined structure, while rule-based validation checks whether particular conditions or requirements are satisfied. For example, an industrial chatbot may be required to provide a maintenance instruction with specific fields, such as the component name, inspection procedure, and safety warning. If a required field is missing or the response violates a predefined rule, the system can flag the output for further checking.

This approach is important for hallucination detection because comparing an answer with retrieved evidence may not identify every formatting error, missing field, or rule violation. Structural and rule-based checks provide an additional layer of verification. However, the reported results should be interpreted within the specific industrial tasks and testing conditions used in the study. Complete accuracy in a limited evaluation does not guarantee the same performance in every application. Nevertheless, combining retrieval with deterministic validation offers a practical method for improving the reliability and consistency of chatbot outputs.

3. Model-Based Detection

Study [3] examines the ability of large language models to identify hallucinations in generated responses. The researchers developed a dataset containing 500 questions across 15 academic and professional disciplines, with controlled hallucinations introduced into the answers. This approach allows the detection system to be evaluated on responses containing known errors rather than relying only on general judgments about answer quality.



The study evaluates GPT-4 as a hallucination detector using several performance measures. Precision measures how many responses identified as hallucinated are actually hallucinated, while recall measures how many existing hallucinations the detector successfully identifies. The reported precision was 0.95, recall was 0.78, and the area under the receiver operating characteristic curve (AUC) was approximately 0.93. These results indicate that the model could identify many hallucinated responses while maintaining a relatively low rate of false alarms.

The study also investigates whether retrieval can improve model-based detection. When retrieved evidence was added during post-hoc verification, detection accuracy increased from 0.83 to 0.87, while the hallucination F1 score improved from 0.78 to 0.86. The F1 score combines precision and recall into a single measure, making it useful for evaluating detection performance. These improvements provide evidence that comparing generated answers with retrieved information can help a model make better detection decisions.

However, model-based detection also has limitations. A language model acting as a detector may misunderstand the evidence or incorrectly judge a correct statement as false. Its performance may depend on the quality of the prompt, the selected decision threshold, and the type of hallucination being evaluated. The study also presents some conceptual elements, including a threshold curve described as hypothetical, which should not be interpreted as experimentally established performance. Overall, this research supports the use of evidence-assisted model verification while highlighting the need for further testing across different datasets and real-world chatbot applications.

4. Similarity-Based Answer Validation

Study [4] presents a similarity-based method for validating generated answers without using a separate retrieval stage. In this approach, a document is divided into smaller text sections called chunks. The language model generates an answer using each chunk, and the system calculates the similarity between the generated answer and its corresponding source chunk. An answer is retained only when its similarity score exceeds a predefined threshold of 0.6. Responses that do not meet this condition are treated as insufficiently supported and are not accepted automatically.

The main advantage of this method is its simplicity. Since each answer is associated with a specific source chunk, the system has a clear reference for checking whether the generated content is related to the original information. The method also provides an explicit decision rule, making it easier to understand why an answer has been accepted or rejected. Such a design may be useful when answers must remain closely connected to particular documents.

The study reports an F1 score of up to 0.909 on the long-context TriviaQA dataset. It also reports scores of approximately 0.94 to 0.99 on a small scientific dataset evaluated through manual judgment. These findings suggest that similarity-based validation can identify potentially unsupported answers under the tested conditions. However, similarity alone cannot guarantee factual correctness. An answer may use words similar to the source while still containing an incorrect claim. Similarly, a correct answer expressed using different words may receive a lower similarity score.

Computational cost is another major limitation. The reported processing time was approximately 262 seconds per question when semantic chunks were used and 2,289 seconds when fixed-size chunks were used, compared with around 85 seconds for reranked RAG. These results show that the method may require substantial processing time, depending on the chunking strategy. Therefore, similarity-based validation offers a straightforward approach to answer checking, but its effectiveness depends on the similarity measure, threshold selection, and computational requirements.

5. Literature Evidence

Study [5] presents a systematic review of 125 research papers related to hallucination mitigation in large language models. The review examines the different methods researchers have developed to prevent, identify, and correct hallucinations. It



organizes these methods into architectural categories, helping explain how hallucination management can be performed at different stages of the response-generation process.

According to the review, post-hoc reactive approaches account for 44% of the reviewed work, while proactive approaches account for 43.2%. Integrated approaches represent only 8.8%. Post-hoc reactive methods examine generated answers to identify unsupported or incorrect information. Proactive methods attempt to prevent hallucinations during generation, for example, by improving the instructions or providing relevant supporting information. Integrated methods combine preventive and corrective mechanisms within a broader system.

The distribution indicates that researchers have mainly focused on detecting hallucinations after generation or reducing their occurrence before an answer is produced. Comparatively fewer studies have investigated systems that combine both approaches. This creates an opportunity to develop frameworks that use reliable information retrieval, answer verification, and correction mechanisms together.

The review also identifies question answering and multimodal applications as major areas of research. However, some application domains, including education and software engineering, remain comparatively under-studied. This is important because different domains have different accuracy requirements and consequences when incorrect information is generated. For example, an unsupported statement in an educational chatbot may confuse students, while an incorrect software instruction may cause implementation problems.

The findings of this review provide a broader understanding of current hallucination mitigation research and its limitations. They suggest that future work should investigate integrated approaches, improve evaluation methods, and test detection systems across different application areas. For the present study, this literature provides the basis for proposing a conceptual framework that combines retrieval-based evidence, model-assisted verification, and rule-based validation to identify potentially unreliable chatbot responses.

IV. Conceptual Detection Framework

The proposed conceptual detection framework combines the important ideas identified in the reviewed studies to improve the reliability of chatbot responses. It is designed to identify incorrect, misleading, or unsupported information before the final answer reaches the user. The framework integrates query-aware retrieval, answer generation, evidence-based verification, rule-based validation, and threshold-based decision-making into a single pipeline. It is a conceptual framework developed from the literature review and has not yet been implemented or experimentally evaluated.

The process begins when a user submits a question to the chatbot. The system analyzes the query and identifies its type, such as a factual question, a definition-based question, or a reasoning-based question. Based on this classification, the system selects a suitable retrieval strategy to obtain relevant information from trusted documents, databases, or knowledge repositories. This step follows the query-aware retrieval approach discussed in Study [1]. The retrieved passages provide supporting evidence that the language model can use to generate an answer.

After retrieving the relevant information, the language model generates a draft response based on the user's question and the available evidence. The draft is then passed to the verification layer, where its content is examined for possible hallucinations. The first verification method compares individual claims in the generated answer with the retrieved passages using a similarity measure, following the approach discussed in Study [4]. This helps identify statements that may not be sufficiently supported by the available evidence. However, similarity alone cannot guarantee factual correctness, so additional verification is required.

The second verification method uses predefined rules and schema validation, based on the ideas presented in Study [2]. This component checks whether the response follows the required structure, contains necessary information, and satisfies specified conditions. The results from both verification methods are combined to estimate the reliability of the generated answer. A predefined threshold is then used to determine the next action.

If the answer satisfies the verification requirements, it is marked as Reliable and delivered to the user. If the answer contains unsupported statements or fails the

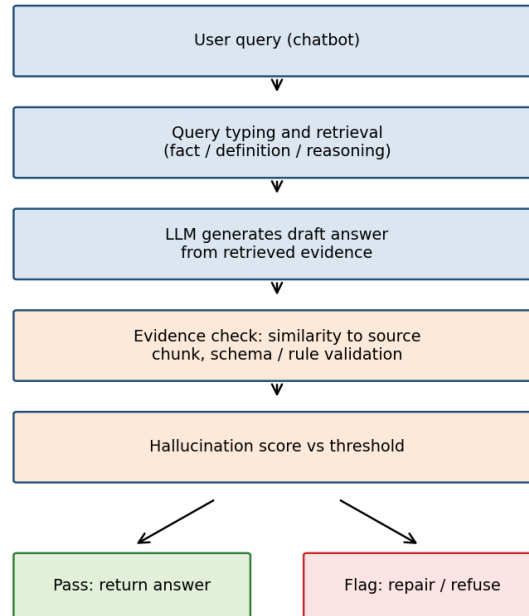


Figure 1: Conceptual hallucination detection pipeline for chatbots

Using an LLM as the verifier, as in [3], is an alternative to the similarity score and can be added as a second opinion when the first check is uncertain. The threshold should be tuned on labelled data, since a strict value raises recall but also raises the rate of false alarms and refusals.

V. Comparative Analysis

Table 1 summarises the five studies from the viewpoint of detection.

Table 1: Summary of the reviewed studies

Study	Core approach	Key reported result	Main caveat
[1] SAC-RAG	Query typing, type-specific retrieval, LLM context compression	38–58% fewer tokens for a 1–2 point EM drop	Latency about 27.8 s vs 9.6 s; single run
[2] Industrial	Guardrails, parsers, RAG with schema validation	RAG plus schema checks fully accurate in the safeguard test	Three bespoke projects; 50 runs

Study	Core approach	Key reported result	Main caveat
[3] Benchmark	Controlled-hallucination QA set, post-hoc retrieval	GPT-4 accuracy 0.87; hallucination F1 0.78 to 0.86	Partly conceptual framework
[4] CA-RAG	No retrieval; per-chunk answers filtered by similarity	F1 about 0.94–0.99 on small scientific set	Very slow; uncontrolled comparisons
[5] Review	PRISMA review of 125 papers, architecture taxonomy	Post-hoc 44%, proactive 43.2%, integrated 8.8%	Coverage ends January 2025

Chart 1 shows the distribution of architectures in the literature surveyed in [5]. Integrated designs, the closest match to the proposed framework, are the least represented.

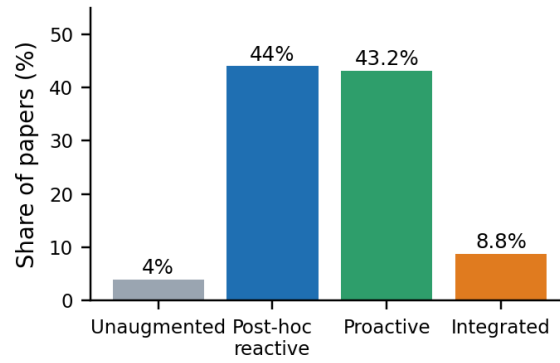


Chart 1: Architecture categories in the review of [5]

Figure 2 contrasts the latency reported in [1] and [4]. The two pairs of bars come from different studies and hardware, so only the ratio within each pair is meaningful.

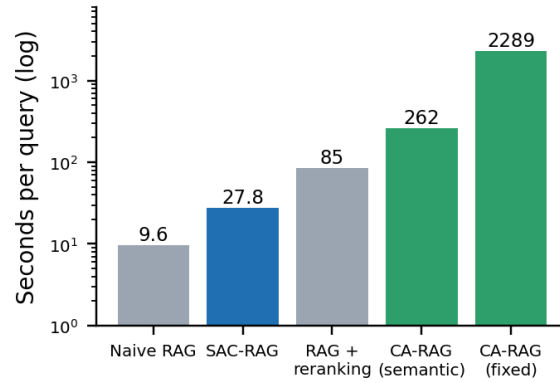


Figure 2: Reported latency per query (logarithmic scale)



VI. Discussion

1. Evidence plus structure

The reviewed studies show that combining evidence-based verification with structural and rule-based validation can provide a stronger approach to hallucination detection. Although Retrieval-Augmented Generation (RAG) helps language models generate answers using information from trusted sources, the retrieved information alone cannot guarantee that every statement in the final answer is correct. A language model may misunderstand the source, combine different facts incorrectly, or generate additional information that is not supported by the retrieved passages. Therefore, an additional verification mechanism is required to improve the reliability of generated responses.

Studies [2], [3], and [4] demonstrate different methods for checking chatbot outputs. Study [3] shows that adding retrieved evidence during post-hoc verification improves detection accuracy and the hallucination F1 score. This indicates that a detector can make better decisions when it has access to relevant supporting information. Study [4] uses similarity-based validation to determine whether generated answers are sufficiently related to their source passages. However, similarity cannot always confirm factual correctness because an answer may be similar to the source while containing an incorrect statement. Study [2] addresses another important issue by combining RAG with schema and rule validation. Its reported results show that structural validation can help ensure that generated responses follow predefined requirements.

These findings suggest that different verification methods serve different purposes. Evidence-based verification checks whether the answer is supported by available information, while structural validation checks whether the response follows the required format and rules. Combining these methods may help identify unsupported claims, missing information, and formatting errors more effectively than using a single method. Therefore, the proposed framework integrates retrieval, answer verification, and rule-based checking to improve the reliability of chatbot responses. However, the effectiveness of this combined approach must be confirmed through experimental evaluation.

2. Latency in Live Chat

Response time is an important factor when implementing hallucination detection in real-world chatbot applications. Users generally expect chatbots to provide accurate answers quickly. However, adding retrieval, verification, and correction stages may increase processing time because the system must perform additional operations before delivering the final response. As a result, improving detection accuracy while maintaining an acceptable response time is a major challenge.

The reviewed studies highlight this trade-off. In Study [1], the reported response time increased from approximately 9.6 seconds to 27.8 seconds when the adaptive retrieval approach was used. Although this method reduces the number of tokens processed, its additional retrieval and processing operations increase latency. Study [4] reports even higher processing times, approximately 262 seconds with semantic chunks and 2,289 seconds with fixed-size chunks. These results indicate that similarity-based validation can become computationally expensive under the tested conditions, making it difficult to use directly in interactive chatbot systems.

Several methods could help reduce this problem. First, caching can store previously retrieved information and verified answers so that repeated questions do not require the entire process to be performed again. Second, smaller and faster verification models could be used for routine checks, while more powerful models could handle complex or uncertain responses. Third, the system could apply additional verification only to high-risk answers, such as those involving medical information, financial decisions, or technical instructions. Context compression and efficient retrieval could also reduce processing requirements. The proposed framework should therefore consider both reliability and speed when making verification decisions. A practical chatbot should identify unsupported information without creating unnecessary delays. Future experiments should compare



different verification strategies to determine which combination provides an appropriate balance between detection performance and response time.

3. Limits of the Evidence

Although the reviewed studies provide useful information about hallucination detection, their results have several limitations. Some evaluations use relatively small datasets, limited experimental runs, or manually scored responses. These conditions may not fully represent the variety of questions, response styles, and hallucination types encountered in real-world chatbot applications. In addition, different studies use different datasets, models, evaluation metrics, and experimental settings, making direct numerical comparisons difficult.

Study [1] reports results that include a single experimental run, which limits the ability to determine whether the same performance would be achieved consistently across repeated tests. Study [2] evaluates its approach using three specific industrial projects and 50 runs. Although these results demonstrate the potential value of schema validation, they do not establish that the same level of accuracy can be achieved in every application. Study [3] includes a controlled-hallucination dataset and presents improvements from evidence-assisted detection, but some elements of its framework remain conceptual. Study [4] reports promising F1 scores on its evaluated datasets, but the small manually judged scientific dataset and uncontrolled comparisons limit the general conclusions that can be drawn. Study [5] provides a broad overview of the research field, but its findings are limited by the studies included in its review.

To address these limitations, the proposed framework requires systematic experimental evaluation using a labelled chatbot dataset containing both correct answers and hallucinated responses. The dataset should include different question types, levels of difficulty, and examples of unsupported, incorrect, and contradictory statements. The system should be tested under consistent conditions so that the contribution of each component can be measured separately.

Evaluation should include precision, recall, and F1 score to measure hallucination detection performance. The false-positive rate should also be measured to identify how frequently correct answers are incorrectly flagged. In addition, response latency and computational cost should be recorded to evaluate the system's practical suitability. Experiments should be repeated several times, and the results should be reported using averages and measures of variation. Comparing the proposed framework with a basic chatbot, a RAG-based system, and individual verification methods would help determine whether combining these components provides meaningful improvements. Such an evaluation would provide stronger evidence about the framework's effectiveness, limitations, and suitability for real-world applications.

VII. Conclusion

Detecting hallucinations in chatbots is essential for improving the accuracy and reliability of AI systems. Retrieval-Augmented Generation (RAG) provides supporting information to verify generated answers, while similarity-based verification and rule-based validation help identify unsupported or incorrect statements. Combining these techniques can improve the reliability of chatbot responses.

However, the reviewed studies highlight challenges such as increased response time, computational cost, and limited experimental evidence. The proposed framework combines retrieval, verification, and validation to classify chatbot responses as reliable or flagged. Since the framework is conceptual, future work will focus on implementing and testing it using a labelled dataset. Its performance will be evaluated based on detection accuracy, response time, and reliability. This approach aims to develop more accurate, transparent, and trustworthy chatbots.



Acknowledgements

We sincerely thank our respected guide and faculty members for their valuable guidance and support in completing this review paper. We are also thankful to Jaihind College of Engineering for providing the necessary resources. Finally, we thank all the researchers whose work helped us prepare this paper.

References

1. Zhang, D., Yu, H., Huo, J., Su, M. "SAC-RAG: Semantic Adaptive Context Compression for Retrieval-Augmented Generation". IEEE Access, 2026. DOI:- <https://doi.org/10.1109/ACCESS.2026.3718025>
2. Tikka, P., Karjalainen, J., Alesani, A., Goriachev, V. "Large Language Model Hallucination Mitigation in Three Industrial Use Cases". IEEE Access, 2026. DOI:- <https://doi.org/10.1109/ACCESS.2026.3659997>
3. Ajmal, R.H., Sarwar, M.U., Hanif, M.K., Khan, M.I. "Evaluating the Effectiveness of Advanced Language Models in Detecting and Mitigating Hallucinations Using Structured Question-Answering, Novel Metrics, and Post-Hoc Retrieval". IEEE Access, 2025. DOI:- <https://doi.org/10.1109/ACCESS.2025.3613851>
4. Collini, E., Kurniadi, F.I., Nesi, P., Pantaleo, G. "Context-Aware Retrieval Augmented Generation Using Similarity Validation to Handle Context Inconsistencies in Large Language Models". IEEE Access, 2025. DOI:- <https://doi.org/10.1109/ACCESS.2025.3614553>
5. C. Woesle, L. Fischer-Brandies, and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," *IEEE Access*, vol. 13, pp. 148231–148253, 2025, DOI:- <https://doi.org/10.1109/ACCESS.2025.3601206>