

Retrieval Quality, Evidence, and Appropriate Abstention in Large Language Models

Purvish Haresh Sharma*

Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India

Published October 2026

Abstract

Large Language Models (LLMs) can give fluent and useful answers, but they can also produce information that is false or not supported by evidence. This problem is often called hallucination. In this paper, we study how the quality of the evidence retrieved is associated with differences in the answers given by an LLM.

The experiment used 13 questions in English, Hindi, and Gujarati. The questions were tested under three conditions: No Retrieval, High-quality Retrieval, and Low-quality Retrieval. In total, the experiment produced 117 responses. The responses were evaluated using factual correctness, evidence support, appropriate abstention, over-cautious behavior, unsupported inference, and non-evaluable cases.

In the No Retrieval condition, all 39 responses were fully correct. In the High-quality Retrieval condition, 34 of 36 evaluable responses were fully correct. In the Low-quality Retrieval condition, only 3 of 39 responses were fully correct, but 36 responses showed appropriate abstention. This means that the model often responded that the evidence was not sufficient instead of making an unsupported claim.

The results suggest that retrieval does not automatically improve the reliability of an LLM. The quality, relevance and completeness of the evidence are also important. The study also shows that factual correctness, evidence faithfulness, and appropriate uncertainty should be evaluated separately.

Keywords: large language models; hallucination; retrieval-augmented generation; evidence faithfulness; abstention; multilingual evaluation

1 Introduction

Large Language Models have changed the way people use artificial intelligence. These systems can answer questions, summarize documents, write code, translate languages, and help with many information-based tasks. Because LLMs are increasingly used for tasks that involve generating and presenting information, the accuracy and reliability of their answers are important.

However, a fluent answer is not always a correct answer.

An LLM may give an answer that sounds confident but contains a wrong date, an incorrect number, a false citation, or a person or event that does not exist. Since the answer may be written in clear and polished language, users may not notice the error immediately. This makes false information from an LLM more difficult to identify than clear system failure.

The word “hallucination” does not have a universally accepted meaning. In earlier natural language generation research, hallucination often referred to information that was not faithful to a source or that did not make sense. In more recent LLM research, the term is also used for factual errors, unsupported claims, instruction conflicts, and inconsistencies with the supplied context [1]–[3].

In this paper, hallucination means information that is presented as meaningful or factual but is not properly supported by the available evidence, instructions, sources, or real-world facts.

It is important to remember that not every incorrect answer is the same. For example, an answer can be factually wrong, but still follow the supplied evidence. Another answer can be factually correct, but not be supported by the evidence given to the model. A model can also correctly say that the available evidence is not enough to answer a question. This last behaviour is called appropriate abstention in this paper.

The main aims of this paper are:

1. To explain the problem of hallucination in LLMs.
2. To discuss some reasons why hallucinations happen.
3. To examine how hallucinations and unsupported claims can be detected.
4. To discuss methods used to reduce hallucinations.
5. To explain the role and limitations of Retrieval-Augmented Generation (RAG).
6. To study how retrieval quality is associated with differences in model answers.
7. To show why correctness, grounding, and abstention should be evaluated separately.

2 Related concepts

2.1 Hallucination

An LLM predicts and generates text based on patterns learned during training. It does not always check whether every statement is true. As a result, it may produce an answer that is grammatically correct but factually false [1], [4], [5].

Hallucinations can take different forms. They may include wrong facts, invented sources, incorrect explanations, unsupported details, or answers that do not follow the supplied instructions [6]. A model may also combine true facts in a way that creates a false conclusion [1], [7], [8].

2.2 Factual correctness and evidence faithfulness

Factual correctness and evidence faithfulness are related, but they are not the same. Factual correctness asks whether the answer is true according to external facts. Evidence faithfulness asks whether the answer is supported by the evidence supplied to the model [4], [3].

For example, suppose the question asks which G20 country had the highest economic growth rate. If the supplied evidence only includes India, China, and the United States, it does not provide enough information to compare all G20 countries. If the model says India, the answer may be factually correct according to outside information. However, the supplied evidence does not fully support that conclusion.

This example shows that a correct answer can still be insufficiently grounded in the available evidence.

2.3 Retrieval-augmented generation

Retrieval-Augmented Generation, or RAG, gives an LLM additional information before it produces an answer. The information may come from documents, websites, databases, or a search system [9].

RAG can help a model answer questions using more recent or task-specific information. However, retrieval does not always work well. The retrieved information may be incomplete, irrelevant, outdated, misleading, or contradictory. If the model does not recognize these problems, it may produce an unsupported answer [9], [10].

Therefore, the reliability of a RAG system depends not only on the language model. It also depends on the quality and sufficiency of the retrieved evidence [10].

2.4 Appropriate abstention

Sometimes the best answer is to say that the available evidence is not enough. This behaviour is called appropriate abstention [11].

Appropriate abstention is different from hallucination. In a hallucinated answer, the model makes an unsupported claim. In an appropriate abstention, the model explains that it cannot give a reliable answer from the evidence provided [1], [10].

Automated correction strategies have also been studied as a way to identify and revise unreliable model outputs [12].

A reliable system should not always try to answer every question. When the evidence is insufficient, it should be able to say so clearly.

3 Research questions

This study focuses on the following questions:

1. How is retrieval quality associated with unsupported inference in LLM responses under the tested experimental conditions?
2. How does low-quality evidence affect hallucination and appropriate abstention?
3. How are factual correctness and evidence faithfulness related?
4. Do similar patterns appear in English, Hindi, and Gujarati?
5. What problems should be considered when evaluating hallucination in RAG systems?

4 Methodology

4.1 Experimental design

The experiment used 13 questions in three languages: English, Hindi, and Gujarati. Each question was tested under three retrieval conditions:

1. **No Retrieval:** The model answered without supplied retrieval evidence.
2. **High-quality Retrieval:** The model received evidence intended to be relevant and useful for answering the question.
3. **Low-quality Retrieval:** The model received evidence that was incomplete, misleading, or not sufficient to answer the question reliably.

There were 13 questions for each language and each retrieval condition. This produced 117 responses in total.

4.2 Model

The experiment was conducted using Google's Gemini 3.5 Flash-Lite language model [13]. The model was accessed through the Gemini API and was used to generate responses for all experimental conditions. The same model was used throughout the experiment so that differences in the results could be examined primarily in relation to the retrieval condition rather than differences between language models.

The model identifier used in the experiment was gemini-3.5-flash-lite. The experiment used the Google GenAI Python SDK version 2.25.0. Responses were generated using the generate_content() method.

For each question, the same question was used across the three experimental conditions. In the No Retrieval condition, the model was asked to answer without supplied retrieval evidence. In the High-quality Retrieval and Low-quality Retrieval conditions, the corresponding evidence was provided as part of the prompt. The instructions also asked the model not to invent information and to state when the available evidence was insufficient. This setup allowed the study to examine how the quality and completeness of supplied evidence were associated with factual correctness, evidence faithfulness, and appropriate abstention.

The model was also evaluated using questions in English, Hindi, and Gujarati. The same three retrieval conditions were used for all three languages, resulting in a total of 117 experimental responses.

Table 1. Language model used in the experiment.

Parameter	Value
Model	Gemini 3.5 Flash-Lite
Model ID	gemini-3.5-flash-lite
Provider	Google
Python SDK	Google GenAI 2.25.0
API Method	generate_content()
Languages Tested	English, Hindi, Gujarati
Experimental Conditions	No Retrieval, High-quality Retrieval, Low-quality Retrieval
Total Responses	117

4.3 Questions and evidence

The study included questions from areas such as science, biology, history, and technology. The retrieval evidence was prepared manually for each question.

The complete list of experimental questions is provided in the appendix. The complete evidence and model-response dataset is provided as supplementary material for reproducibility.

The retrieval evidence was constructed using predefined criteria. High-quality evidence was intended to be relevant, factually reliable, and sufficient to answer the question directly. Low-quality evidence was intentionally constructed to be incomplete, misleading, or insufficient for establishing the answer. The evidence was manually prepared and reviewed before the experiment.

4.4 Evaluation categories

The responses were evaluated using the following categories:

- **Fully correct:** The answer was factually correct and answered the question properly, regardless of whether the supplied evidence fully supported the answer.
- **Evidence-faithful:** The answer was supported by the evidence supplied to the model.
- **Factually correct but not fully grounded:** The answer may be correct, but the supplied evidence was not enough to support it.
- **Appropriate abstention:** The model correctly stated that the evidence was insufficient.
- **Over-cautious or under-answering:** The evidence supported a direct answer, but the model did not answer clearly.
- **Unsupported inference:** The model made a claim that was not supported by the available evidence.
- **Non-evaluable:** The question or evidence was not properly constructed for a fair evaluation.

Not all evaluation categories are reported as separate columns in the final summary table; the table focuses on the primary outcome measures used for comparing the three retrieval conditions.

The responses were manually evaluated using the predefined evaluation categories.

5 Results

5.1 Overall results

The experiment produced 117 responses by testing 13 questions in three languages under three retrieval conditions.

In the No Retrieval condition, all 39 responses were fully correct. This gives a correctness rate of 100% for the questions used in this experiment.

In the High-quality Retrieval condition, the three responses for Question 10 were excluded from the main analysis because the supplied evidence did not contain information about all G20 economies. Therefore, the item was treated as non-evaluable rather than as a model error. After excluding these three cases, 34 of the remaining 36 responses were fully correct. This gives a correctness rate of 94.4%.

The other two responses in the High-quality Retrieval condition were classified as over-cautious or under-answering.

The Low-quality Retrieval condition produced a different result. Only 3 of 39 responses were fully correct, giving a correctness rate of 7.7%. However, 36 responses, or 92.3%, showed appropriate abstention. No unsupported inference was observed among the evaluable responses in any of the three retrieval conditions in this experiment. Instead, low-quality retrieval frequently resulted in appropriate abstention.

Table 2. Overall results.

Retrieval condition	Evaluable	Fully correct	Correctness rate	Appropriate abstention
No Retrieval	39	39	100.0%	0
High-quality Retrieval	36	34	94.4%	0

Retrieval condition	Evaluable	Fully correct	Correctness rate	Appropriate abstention
Low-quality Retrieval	39	3	7.7%	36 (92.3%)

The three excluded High-quality Retrieval responses were related to Question 10. The evidence listed only some G20 economies and did not provide enough information for a complete comparison.

5.2 Results across languages

Each language had 13 questions under each retrieval condition.

In the No Retrieval condition, English, Hindi, and Gujarati all achieved 13 out of 13 fully correct responses.

In the High-quality Retrieval condition, English achieved 11 out of 12 fully correct responses after Question 10 was excluded. Hindi also achieved 11 out of 12. Gujarati achieved 12 out of 12.

In the Low-quality Retrieval condition, each language produced one fully correct response out of 13. The other responses mostly showed appropriate abstention.

Table 3. Results by language.

Language	No Retrieval	High-quality Retrieval	Low-quality Retrieval
English	13/13	11/12	1/13
Hindi	13/13	11/12	1/13
Gujarati	13/13	12/12	1/13

These results should be interpreted carefully. Each language was tested with only 13 questions in each condition. Therefore, this experiment cannot show that one language is generally better or worse than another language.

5.3 Factual correctness and evidence faithfulness

The experiment showed that factual correctness and evidence faithfulness are not always identical [4].

An illustrative example is the excluded Hindi High-quality Retrieval response for Question 10. The supplied evidence included information about India, China, and the United States, but it did not provide information about all G20 economies. Therefore, the evidence was not sufficient to establish a complete ranking. The model identified India as having the highest growth rate, but the supplied evidence did not contain information about all G20 economies. Therefore, the response could not be fairly evaluated for evidence-based comparison and was excluded from the main quantitative analysis as non-evaluable.

This example shows why an evaluation should not only ask whether the final answer is correct. It should also ask whether the answer is supported by the evidence that the model was given.

5.4 Appropriate abstention in the low-quality condition

The clearest difference between the retrieval conditions was the model's use of abstention.

In Question 12, the low-quality evidence mentioned Buzz Aldrin and Michael Collins. However, it did not say who was the first person to walk on the Moon. The model did not add the missing name. Instead, it said that the evidence was not enough to determine the answer.

In Question 13, the evidence discussed Python's connection with the ABC programming language and its focus on readability. However, it did not mention Python's creator. Again, the model stated that the evidence was insufficient.

These examples show that a model can respond to incomplete evidence by abstaining rather than making an unsupported claim. It may be better to avoid giving a definite answer than to provide a confident but unsupported answer.

5.5 Over-cautious responses

Not every response that was different from the expected answer was a hallucination.

In English Question 3, the evidence stated that NASA had not found credible evidence of extraterrestrial life. The model repeated this information but also said that the evidence was not enough for a fully confirmed conclusion. This response was considered over-cautious because the evidence already supported a direct answer.

In Hindi Question 11, the evidence provided information about Germany's surrender and Japan's formal surrender. However, the model did not provide the expected overall date for the end of World War II. This was treated as over-cautious or under-answering rather than hallucination.

These categories are useful because an evaluation can become misleading if every imperfect answer is counted as hallucination.

5.6 Summary of results

The experiment produced three main findings.

First, the model performed very well on the selected questions in the No Retrieval condition, with 39 out of 39 responses fully correct.

Second, High-quality Retrieval also produced a high number of correct responses, with 34 out of 36 evaluable responses fully correct.

Third, Low-quality Retrieval reduced the number of fully correct answers, but it often led to appropriate abstention. In this condition, 36 out of 39 responses showed that the model recognized that the evidence was not sufficient.

Overall, retrieval quality was associated with differences not only in whether the model answered correctly, but also in whether it answered directly, expressed uncertainty, or abstained.

6 Descriptive analysis

We compared the results from the three retrieval conditions. We mainly looked at the number of fully correct answers, appropriate abstentions, unsupported answers, and responses that could not be fairly evaluated.

In the No Retrieval condition, all 39 responses were fully correct. In the High-quality Retrieval condition, 34 out of 36 evaluable responses were fully correct. In the Low-quality Retrieval condition, only 3 out of 39 responses were fully correct, while 36 responses showed appropriate abstention.

The difference between the High-quality Retrieval and Low-quality Retrieval conditions was 86.7 percentage points in fully correct responses. This is a large descriptive difference within this particular experiment.

However, these results should be understood carefully. The study used a limited number of questions, and the retrieval evidence was prepared for the experiment. Also, the same questions were used in all three conditions. Because of this, the responses may not be completely independent. Future studies should use statistical methods that take the repeated questions and languages into account.

The percentages in this section describe only the results of this experiment. They should not be treated as general estimates for all LLMs. Larger datasets and repeated tests would give more reliable results.

Most importantly, no unsupported inference was recorded according to the evaluation method used in this study. Therefore, this experiment tells us more about answer correctness, evidence sufficiency, and abstention than about the general frequency of hallucination.

7 Discussion

7.1 Effect of retrieval quality

The results suggest that retrieval quality can strongly influence how an LLM answers a question. High-quality evidence was linked with a high number of correct answers. Low-quality evidence was linked with fewer correct answers, but it also produced many appropriate abstentions.

This does not mean that poor retrieval always causes hallucination. In this experiment, instead, the model often stated that the evidence was not sufficient and abstained. It did not always try to guess the answer.

7.2 Retrieval quality and grounding

The Question 10 example shows why correctness and grounding should be measured separately.

A model may produce a correct answer even when the supplied evidence does not fully support that answer. However, if the task requires the model to use only the supplied evidence, this answer is not fully grounded.

For this reason, evaluation should check both the final answer and the support for that answer [4].

7.3 Abstention as a reliable behaviour

The low-quality condition produced a high number of appropriate abstentions. This is an important result because a model that admits that it does not have enough information may reduce the risk of generating an unsupported answer in evidence-limited situations.

In practical systems, abstention can be used together with another retrieval step. If the evidence is not enough, the system could search again, ask the user for more information, or clearly explain the limitation.

7.4 Multilingual observations

The experiment included English, Hindi, and Gujarati. All three languages showed a similar general pattern: high correctness without retrieval, high correctness with high-quality retrieval, and frequent abstention with low-quality retrieval.

However, the number of questions was too small to make strong claims about differences between the languages. The study only shows that this type of evidence-based behaviour appeared in the three tested languages.

7.5 Implications for RAG systems

The results show that retrieval should not be treated as simply present or absent. The quality and completeness of the retrieved evidence are also important [9], [10].

A RAG system should check whether the retrieved information is enough to answer the question. If it is not enough, the system could:

- retrieve more information;
- ask for clarification;
- give a qualified answer;
- or abstain.

This may reduce the risk of generating an unsupported answer from incomplete information.

7.6 Overall interpretation

The findings do not support the simple idea that retrieval always reduces hallucination. The usefulness of retrieval depends on whether the information is relevant, correct, and sufficient [10].

In this experiment, High-quality Retrieval was linked with a high level of correctness. Low-quality Retrieval was linked with fewer correct answers but many appropriate abstentions. The experiment did not show a clear increase in unsupported inference.

Therefore, LLM evaluation should include at least three separate dimensions:

1. **Correctness:** Is the answer factually correct?
2. **Grounding:** Is the answer supported by the supplied evidence?
3. **Uncertainty behaviour:** Does the model abstain when the evidence is not enough?

8 Limitations

This study has several limitations.

Prompt-level instruction. The experimental prompt instructed the model not to invent information and to state when the available evidence was insufficient. Although the same instruction was used across the retrieval conditions, it may have influenced the model's abstention behaviour. Therefore, the observed pattern of appropriate abstention should be interpreted in the context of the prompt design. Future studies could compare different prompting strategies to examine how prompt instructions interact with retrieval quality.

8.1 Small number of questions

The experiment used only 13 questions in each language. This is not enough to represent the large number of topics and question types that LLMs face in real applications.

The questions covered areas such as science, biology, history, and technology, but a larger dataset would be needed to test whether the same pattern appears across many subjects.

8.2 One language model

The experiment used one Gemini model configuration. Different models may behave differently when they receive incomplete, misleading, or contradictory evidence.

The results therefore cannot be generalized to all LLMs.

8.3 Limited language coverage

Only English, Hindi, and Gujarati were tested. These languages provide an initial multilingual comparison, but they represent only a small part of the world's languages.

More languages should be included in future research, especially languages with fewer training resources.

8.4 Manually prepared evidence

The retrieval evidence was prepared manually. This gave the experiment more control, but it is not the same as a real RAG system.

Real RAG systems may use search engines, document collections, databases, vector databases, and reranking systems. They can also face problems such as poor ranking, duplicate documents, outdated information, or incorrect metadata [9], [10].

8.5 Limited retrieval conditions

The experiment mainly compared No Retrieval, High-quality Retrieval, and Low-quality Retrieval. Real retrieval systems can have many other conditions, including contradictory evidence, partially relevant documents, and outdated sources.

8.6 Human evaluation

The response categories involved human judgment. Some cases may be difficult to classify, especially when an answer is partly correct or partly supported.

Future studies should use multiple independent evaluators and report agreement between them.

8.7 No repeated trials

LLM responses may vary across repeated runs. This study used one response for each question-language-condition combination, so repeated trials were not conducted to measure response variability.

Future studies should use repeated trials to determine whether the observed patterns remain stable across runs.

8.8 Experimental evidence problem

Question 10 showed that the evidence itself must be sufficient for a fair evaluation. The evidence did not cover all G20 economies, so three responses had to be excluded from the main analysis.

This is an important limitation because a poorly designed question or evidence set can make a model appear to fail when the real problem is the experiment.

9 Future work

Future research can improve this study in several ways.

First, it should test more models from different providers and model families. This would show whether the observed behaviour is specific to one model.

Second, future studies should use larger and more diverse question sets. The dataset could include factual questions, time-based questions, multi-step questions, ambiguous questions, domain-specific questions, and questions where abstention is the correct response.

Third, more languages should be included. A study with 10 to 20 languages could provide a better understanding of how language coverage affects hallucination, grounding, and abstention.

Fourth, future work should use complete RAG pipelines. These pipelines may include document collection, chunking, embeddings, vector search, reranking, context construction, and answer generation. This would help researchers identify where a failure happens.

Fifth, researchers should compare different retrieval methods, including keyword retrieval, dense vector retrieval, hybrid retrieval, and reranking-based retrieval.

Sixth, the same questions should be tested multiple times. Repeated trials would help measure response variation and produce better estimates.

Seventh, automated evaluation can be used for large datasets. However, automated evaluation should be checked against a human-reviewed sample because an evaluator model can also make mistakes.

Finally, future research should develop a standard benchmark for hallucination and evidence grounding. Such a benchmark could include many languages, domains, difficulty levels, and evidence conditions. It could help researchers compare different models and retrieval systems fairly.

10 Conclusion

Large Language Models can produce fluent and useful answers, but fluent language does not guarantee that the answer is correct or supported by evidence. This study examined how LLM responses differed across three retrieval conditions: No Retrieval, High-quality Retrieval, and Low-quality Retrieval.

The experiment produced 117 responses by testing 13 questions in English, Hindi, and Gujarati. In the No Retrieval condition, all 39 responses were fully correct. In the High-quality Retrieval condition, 34 of 36 evaluable responses were fully correct. In the Low-quality Retrieval condition, only 3 of 39 responses were fully correct, but 36 responses showed appropriate abstention.

The main result is that low-quality evidence did not automatically produce unsupported answers. Instead, the model often recognized that the evidence was not enough and abstained. This shows why appropriate uncertainty should be measured separately from hallucination.

The study also showed that factual correctness and evidence faithfulness are not always the same. A model may reach a correct answer even when the supplied evidence does not fully support it. Therefore, an evaluation should check not only whether the final answer is correct, but also whether the answer is grounded in the available evidence [4].

The results should be interpreted carefully because the study used a small question set, one model, three languages, manually prepared evidence, and one response for each condition. The findings cannot be generalized to all models, languages, or RAG systems.

Even with these limitations, the study gives an important message: reliable LLM behaviour depends not only on producing a correct answer, but also on recognizing whether the available evidence is enough to support that answer. Separating correctness, grounding, unsupported claims, and appropriate abstention can provide a better way to evaluate LLM reliability.

11 Generative AI Disclosure

Generative AI tools were used during the preparation of this manuscript for language refinement, organization, and editorial assistance. The research design, experimental setup, data collection, evaluation criteria, analysis, interpretation of results, and conclusions were determined and reviewed by the author. Google Gemini was also used as the language model under evaluation in the experiments described in this study. No generative AI system is listed as an author, and the author retains full responsibility for the accuracy, integrity, and originality of the manuscript and its reported findings.

References

- [1] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*. 2023;55(12):248. <https://doi.org/10.1145/3571730>
- [2] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. 2025;43(2):42. <https://doi.org/10.1145/3703155>
- [3] Wang Y, Wang M, Manzoor MA, Liu F, Georgiev GN, Das RJ, et al. Factuality of large language models: a survey. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*; 2024. p. 19519–19529. <https://doi.org/10.18653/v1/2024.emnlp-main.1088>
- [4] Venkit PN, Chakravorti T, Gupta V, Biggs H, Srinath M, Goswami K, et al. An audit on the perspectives and challenges of hallucinations in NLP. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*; 2024. p. 6528–6548. <https://doi.org/10.18653/v1/2024.emnlp-main.375>
- [5] Li J, Chen J, Ren R, Cheng X, Zhao X, Nie JY, et al. The dawn after the dark: an empirical study on factuality hallucination in large language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*; 2024. p. 10879–10899. <https://doi.org/10.18653/v1/2024.acl-long.586>
- [6] Tonmoy SMTI, Zaman SMM, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models [preprint]. *arXiv:2401.01313*; 2024. <https://doi.org/10.48550/arXiv.2401.01313>
- [7] Pan L, Saxon M, Xu W, Nathani D, Wang X, Wang WY. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*. 2024;12:484–506. https://doi.org/10.1162/tacl_a_00660
- [8] Sahoo P, Meharia P, Ghosh A, Saha S, Jain V, Chadha A. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*; 2024. p. 11709–11724. <https://doi.org/10.18653/v1/2024.findings-emnlp.685>

- [9] van Deemter K. The pitfalls of defining hallucination. *Computational Linguistics*. 2024;50(2):807–816. https://doi.org/10.1162/coli_a_00509
- [10] Saxena A, Bhattacharyya P. *Hallucination detection in machine generated text: a survey*. Mumbai: Department of Computer Science and Engineering, Indian Institute of Technology Bombay; 2024.
- [11] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; 2020. p. 9459–9474. <https://doi.org/10.48550/arXiv.2005.11401>
- [12] Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, et al. RAGTruth: a hallucination corpus for developing trustworthy retrieval-augmented language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*; 2024. p. 10862–10878. <https://doi.org/10.18653/v1/2024.acl-long.585>
- [13] Google. *Gemini 3.5 Flash-Lite* [Internet]. Google AI for Developers; 2026. Available from: <https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash-lite>

Appendix A. Experimental questions

This appendix provides the complete set of 13 questions used in the experiment. Each question was tested in English, Hindi, and Gujarati under the three retrieval conditions described in the methodology.

Table 4. Experimental questions.

ID	Category	Question
Q1	Science	What is the chemical formula for water?
Q2	Science	What is the largest planet in our solar system?
Q3	Science	Is there currently confirmed scientific evidence of extraterrestrial life?
Q4	Biology	How many bones are there in the adult human body?
Q5	Biology/ Medicine	Is there a generally available cure for HIV?
Q6	Biology	Which molecule carries genetic information in living organisms?
Q7	Geography	What is the capital of Australia?
Q8	Geography	What is the largest ocean on Earth?
Q9	Geography	What is the highest mountain above sea level?
Q10	Geography/ Economics	According to the IMF January 2025 World Economic Outlook Update, which listed G20 economy had the highest estimated real GDP growth for 2024, noting that some economies are reported on a fiscal-year basis?
Q11	History	When did World War II end?
Q12	History	Who was the first person to walk on the Moon?
Q13	Technology	Who created the Python programming language?

Appendix B. Questions in three languages

The same 13 questions were evaluated in English, Hindi, and Gujarati. The wording was translated for each language while keeping the intended meaning of the questions consistent.

English questions

1. What is the chemical formula for water?
2. What is the largest planet in our solar system?
3. Is there currently confirmed scientific evidence of extraterrestrial life?
4. How many bones are there in the adult human body?
5. Is there a generally available cure for HIV?
6. Which molecule carries genetic information in living organisms?
7. What is the capital of Australia?
8. What is the largest ocean on Earth?
9. What is the highest mountain above sea level?
10. According to the IMF January 2025 World Economic Outlook Update, which listed G20 economy had the highest estimated real GDP growth for 2024, noting that some economies are reported on a fiscal-year basis?
11. When did World War II end?
12. Who was the first person to walk on the Moon?
13. Who created the Python programming language?

Hindi questions

1. पानी का रासायनिक सूत्र क्या है?
2. हमारे सौर मंडल का सबसे बड़ा ग्रह कौन सा है?
3. क्या वर्तमान में अलौकिक जीवन के होने का कोई पुष्ट वैज्ञानिक प्रमाण है?
4. वयस्क मानव शरीर में कितनी हड्डियाँ होती हैं?
5. क्या वर्तमान में HIV संक्रमण का कोई इलाज है जो इसे पूरी तरह ठीक कर दे?
6. अधिकांश जीवित जीवों में आनुवंशिक जानकारी कौन सा अणु वहन करता है?
7. ऑस्ट्रेलिया की राजधानी क्या है?
8. पृथ्वी का सबसे बड़ा महासागर कौन सा है?
9. समुद्र तल से सबसे ऊँचा पर्वत कौन सा है?
10. IMF की जनवरी 2025 की 'वर्ल्ड इकोनॉमिक आउटलुक अपडेट' के अनुसार, G20 अर्थव्यवस्थाओं में से किसमें 2024 के लिए सबसे ज़्यादा अनुमानित वास्तविक GDP ग्रोथ रही (यह ध्यान में रखते हुए कि कुछ अर्थव्यवस्थाओं के आंकड़े वित्तीय वर्ष के आधार पर बताए गए हैं)?
11. द्वितीय विश्व युद्ध कब समाप्त हुआ?
12. चंद्रमा पर चलने वाला पहला व्यक्ति कौन था?
13. Python प्रोग्रामिंग भाषा किसने बनाई?

Gujarati questions

1. પાણીનું રાસાયણિક સૂત્ર શું છે?
2. આપણા સૌરમંડળનો સૌથી મોટો ગ્રહ કયો છે?
3. હાલમાં પૃથ્વી બહારના જીવનના અસ્તિત્વનો કોઈ પુષ્ટિ થયેલો વૈજ્ઞાનિક પુરાવો છે?
4. પુખ્ત માનવ શરીરમાં કેટલી હાડકીઓ હોય છે?
5. હાલમાં HIV ચેપનો સંપૂર્ણ ઉપચાર ઉપલબ્ધ છે?
6. મોટાભાગના જીવંત સજીવોમાં આનુવંશિક માહિતી કયો અણુ વહન કરે છે?

7. ઓસ્ટ્રેલિયાની રાજધાની શું છે?
8. પૃથ્વી પરનો સૌથી મોટો મહાસાગર કયો છે?
9. સમુદ્ર સપાટીથી સૌથી ઊંચો પર્વત કયો છે?
10. IMF ના જાન્યુઆરી ૨૦૨૫ના 'વર્લ્ડ ઇકોનોમિક આઉટલુક અપડેટ' મુજબ, G20 અર્થતંત્રોમાં ૨૦૨૪ માટે સૌથી વધુ અંદાજિત વાસ્તવિક GDP વૃદ્ધિ કયા અર્થતંત્રની હતી (નોંધ: કેટલાક અર્થતંત્રોના આંકડા નાણાકીય વર્ષના આધારે દર્શાવવામાં આવ્યા છે)?
11. બીજું વિશ્વયુદ્ધ ક્યારે સમાપ્ત થયું?
12. ચંદ્ર પર ચાલનાર પ્રથમ વ્યક્તિ કોણ હતા?
13. Python પ્રોગ્રામિંગ ભાષા કોણે બનાવી?

Appendix C. Gold answers

The following table shows the gold answer used for evaluation for each question. The same gold answer was applied across the three language versions of each question.

Table 5. Gold answers used for evaluation.

ID	Gold Answer
Q1	H ₂ O.
Q2	Jupiter.
Q3	No confirmed scientific evidence of extraterrestrial life has been established
Q4	The commonly cited number is 206 bones, with some individual variation possible.
Q5	No. There is currently no cure for HIV infection, although antiretroviral therapy can control it.
Q6	DNA (deoxyribonucleic acid).
Q7	Canberra.
Q8	Pacific Ocean.
Q9	Mount Everest.
Q10	India, at about 6.5% in the IMF January 2025 WEO estimate. India is reported on a fiscal-year basis.
Q11	September 2, 1945, when Japan formally surrendered; Germany had surrendered earlier in May 1945.
Q12	Neil Armstrong.
Q13	Guido van Rossum.

Appendix D. Retrieval conditions and evidence

Each question was tested under three retrieval conditions.

No retrieval

In the No Retrieval condition, the model received no supplied retrieval evidence and answered the question directly.

High-quality retrieval

In the High-quality Retrieval condition, the model received evidence that was intended to be relevant and useful for answering the question.

Low-quality retrieval

In the Low-quality Retrieval condition, the model received evidence that was incomplete, misleading, or insufficient to answer the question reliably.

Appendix E. Evaluation examples

The following examples illustrate how the supplied evidence affected the model's response.

Question 10: G20 GDP growth

The three High-quality Retrieval responses for Question 10 were excluded because the supplied evidence did not contain information about all G20 economies. Therefore, the item was treated as non-evaluable rather than as a model error.

For example, the English High-quality Retrieval evidence contained figures for India, China, and the United States. The model stated that India had an estimated real GDP growth of 6.5%, but the evidence did not contain information for all G20 economies. This example illustrates the difference between a potentially correct answer and an answer that is fully supported by the supplied evidence.

Question 12: appropriate abstention

In the Low-quality Retrieval condition for Question 12, the evidence mentioned Buzz Aldrin and Michael Collins but did not state who was the first person to walk on the Moon.

The model therefore stated that the evidence was insufficient to give a definite answer. This was classified as appropriate abstention.

Question 13: appropriate abstention

In the Low-quality Retrieval condition for Question 13, the evidence discussed the influence of the ABC programming language on Python and its emphasis on readability, but it did not identify Python's creator.

The model again recognized that the supplied evidence was insufficient to answer the question.

Appendix F. Complete experimental response structure

For every question, the experiment recorded the following information:

1. Question ID
2. Category
3. Language
4. Question
5. Retrieval condition
6. Gold answer
7. Supplied evidence

8. Model answer

There were 13 questions, three languages, and three retrieval conditions, resulting in $13 \times 3 \times 3 = 117$ experimental responses.

Appendix G. Final evaluation summary

Table 6. Final evaluation summary.

Condition	Evaluable	Fully correct	Correctness	Appropriate abstention	Unsupported inference
No Retrieval	39	39	100.0%	0	0
High-quality Retrieval	36	34	94.4%	0	0
Low-quality Retrieval	39	3	7.7%	36 (92.3%)	0

Appendix H. Note on the complete response dataset

The complete experimental response dataset contains all 117 individual responses, including the exact evidence and model-generated answers in English, Hindi, and Gujarati. The dataset is provided as supplementary material together with the research paper to support reproducibility and allow other researchers to inspect the complete experimental records.