International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue. 4, July-Aug 2026, page no. 185-199

To cite this article: Sandeep Kumar and Shinty PK (2026). AN EXPLAINABLE ARTIFICIAL INTELLIGENCE FRAMEWORK FOR DETECTING HALLUCINATIONS IN LARGE LANGUAGE MODELS, International Journal of Applied Science and Engineering Review (IJASER) 7 (4): 185-199 Article No. 298 Sub Id 435

AN EXPLAINABLE ARTIFICIAL INTELLIGENCE FRAMEWORK FOR DETECTING HALLUCINATIONS IN LARGE LANGUAGE MODELS

Sandeep Kumar[1] and Shinty PK[2]

School of Science and Computer Studies CMR University
Bengaluru, Karnataka, India
[1]sandeep.kumar@cmr.edu.in

School of Science and Computer Studies CMR University
Bengaluru, Karnataka, India
[2]shinty.p@cmr.edu.in

DOI: https://doi.org/10.52267/IJASER.2026.7414

ABSTRACT

Large Language Models have changed the way computers understand and use language. These models can create answer that make sense and fit the situation in areas. Even though they have improved they can still make up things that're not true. These made-up parts can look real. They are not correct. This can make people not trust the information and can be dangerous in areas like health, money, school and law. This research introduces a method called Explainable Artificial Intelligence that helps find these made-up parts and explain why it thinks they are not correct. The methos uses the way to check if the information is right. It looks at how the words fits how the similar the meaning is, how sure the model is and how it makes decisions. Techniques like showing which parts of the text are most important and how the model decides help people understand why something might be wrong. The method was tested using data sets with false information with large model. The results were checked using ways to see how good it was. The study shows that the method work well at finding made-up information and helps people trust the system more. By making sure the model is correct and the way it makes decisions is clear this method helps build trustworthy AI system. It gives a start, foe work on using large models in a responsible way.

KEYWORDS: Explainable AI (XAI), Large Language Model (LLMs), Hallucination Detection, Natural Language Processing (NLP), Model Interpretability, Trustworthy Artificial Intelligence, Semantic Consistency analysis

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

1. INTRODUCTION

Large Language Models Have really changed the Artificial Intelligence works. They can. Generate human language in a very natural way. These models are being used in lots of areas like chatbots, summarizing documents making software helping with healthcare looking at laws writing science papers and teaching. They are really good at making responses that make sense in a conversation. This has made people want to use Artificial Intelligence more both in research and in real world applications [1], [2].

However, one big problem is making sure the things this model generates are actually true. One major issue with Large

Language Models is hat can make things up which is called hallucinations. This means they produce information that sounds good but is actually not true they can make mistakes. Give wrong information. This is especially bad in area like healthcare, money, law and science where people need to know the truth to make decisions [3], [4]. Some studies have found out why Large Language Model make things up. It is because the data they were trained on is not good enough, they do not understand the context well. The questions they are asked are not clear. Some people have suggested ways to fic this problem link using Retrieval-Augmented Generation, Knowledge-grounded reasoning and fact-checking. Although these methods have helped reduce the number of hallucinations they do not really explain how they work or why a certain answer is right or wrong [4], [5]. To solve this problem Explainable Artificial Intelligence has become very important. It helps us understand how Artificial Intelligence makes decisions by showing us what is important how things are related and how confident it is. This way people can understand what the Artificial Intelligence is saying, find mistakes and trust the decisions more. Making Large Language Models explainable when they detect hallucinations is very helpful. Of just saying something is wrong it can show us why it is wrong. This helps researchers find out what is wrong with the model developers make the models better. Users check the information before making important decisions. As Artificial Intelligence is used more and more in areas it is crucial that we can trust it and know why it is making certain decisions. [5], [6].

The main things that this study did are these:

• Development of a framework that helps us understand how Artificial Intelligence works when it detects things that're not real in Large Language Models [9].

• We combined a method like checking if the meaning of the words makes sense looking at the context and figuring out how sure we are to make the detection better [6], [8].

• We used techniques to make the Artificial Intelligence explain things in a way that people can understand for each guess it makes [7].

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

- We tested the framework we made using measures like how often it is right how precise it is and how well it can be understood. [10].
- This study helped make Artificial Intelligence more trustworthy by making people feel more confident in it and, by supporting the use of Artificial Intelligence systems that can generate things.

II. LITERATURE REVIEW

The Large Language Models have really taken off. Lots of people are paying attention to them because they can understand and create human-like language that sounds very natural. These models are really good at things like answering questions creating content translating languages helping with programming and summarizing documents. Even with all these advances people have found that one of the biggest problems with Large Language Models is something called hallucination. This is when a model creates information that sounds believable but is not actually true or goes against what we know to be true [11], [12]. This problem has led to a lot of research on how to make language models more reliable and trustworthy and people have come up with ways to detect and fix hallucinations. On researchers were trying to figure out why hallucinations happen in the first place. They found out that it is because of a different-things, like the models not having enough good training data not being able to understand the context very well and not being able to get to outside information when they need it. Even the best language models can create references, wrong numbers or misleading explanations when they do not have good evidence. This has shown that we need to come up with ways to check if the information's true rather than just looking at how well the language sounds [11], [13].

To stop Large Language Models from creating hallucinations people have come up with a few techniques. One popular method is called Retrieval-Augmented Generation, which helps the model by giving it information before it creates a response. Other people have looked at ways to make sure the information is based on knowledge estimate how confident the model is and fact-check automatically. While these methods have helped people have noticed that a lot of systems are still like boxes making it hard for users to understand why the model made a certain decision.

Because people want to know more about how these models' work researchers have started looking into something called Explainable Artificial Intelligence. This is a way to make it clear how an Artificial Intelligence model comes to its conclusions. Of just giving an answer these methods show what factors were important how they are related and how confident the model is. Recent studies have shown that using explainability to detect hallucinations helps people trust the model more makes it easier to fix problems and helps us use Artificial Intelligence in a way especially in important applications [12], [14]. Even though we have made a lot of progress in detecting hallucinations there are still some challenges to overcome. Most of the methods we have now are good at detecting hallucinations. They do not do a great

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

job of explaining why a certain response was marked as a hallucination. Also, a lot of the explanation methods are designed for Artificial Intelligence researchers. Are hard for other people to understand. This shows that we need a system that can detect hallucinations reliably and also explain things in a way that's easy to understand. That is why this research is proposing a framework that combines several methods to improve both detection and trust in Large Language Models [16], [18]

• The Large Language Models are getting better at understanding and creating like language [1], [2], [11].
• They are good at things like answering questions and creating content [1], [11], [12].
• They have a problem, with something called hallucination, which is when they create information that is not true [3], [14], [13].
• Researchers have found out why hallucinations happen and have come up with ways to detect and fix them. [5], [6], [12], [14].
• They have also started looking into Explainable Artificial Intelligence to make it clear how the models work [7], [8], [15], [16].
• The goal is to create a system that can detect hallucinations reliably and explain things in a way that's easy to understand which will help people trust Large Language Models more [9], [10], [16], [18].

III. RESEARCH GAP

Large Language Models have gotten a lot better at reducing hallucinations [12], [14]. There are still some big problems that have not been solved [16], [17]. Most of the time people work on making the models better at detecting things. They use things like fact checking and trying to figure out how sure the model is [5], [6], [12], [14]. They do not really explain why the model made a certain choice [15], [16]. Also, a lot of the ways we try to understand the models are really hard for people to use if they are not experts [15], [16], [17]. We need a way of doing things that combines a few different approaches. This would include checking that the models' answers make sense, making sure the answers are correct, figuring out how sure the model is and explaining things in a way that people can understand [16], [17], [18]. Large Language Models need a system that helps with all of these things. This would help people trust the models more and make choices [9], [10], [18].

IV. PROBLEM STATEMENT

Large Language Models are really important for Artificial Intelligence applications now because they can make text that sounds good and is relevant [1], [2]. The problem is that these models often come up with information that sounds right but is actually wrong or not true [3], [4]. This kind of thing hurts the credibility of Artificial Intelligence systems. Causes big problems in areas like healthcare, finance, education, law and science where we need to have accurate information to make good decisions [3], [4], [13].

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

We already have some ways to deal with this issue like using information to help generate text checking facts and trying to figure out how sure we are, about something [5], [6], [12], [14]. Most of these methods do not really show us how they make their predictions [15], [16]. This makes it hard for people to trust what the Artificial Intelligence system says or understand why it says things [7], [8], [15].

So, we need a way of doing things that can explain what is going on [7], [8], [16]. We need an Artificial Intelligence system capable of detecting when it has made a wrong claim and be able to explain the reasoning that led it to produce a false statement [9], [10], [16], [17]. This will help us trust Large Language Models more and make them more reliable and transparent [9], [10], [18]. Large Language Models need to be able to do this to be really useful. Large Language Models have to give us explanations that make sense to people [16], [18].

V. PROPOSED SYSTEM ARCHITECTURE

When a user submits a query to the system the architecture begins. The user's query is first looked at by the Input Preprocessing module. This is where the module removes any symbols, special characters and formatting inconsistencies from the query. The text is then made normal. Broken down into smaller parts so it can be used for further processing. This step makes the query better. Makes sure the Large Language Model gets clean and useful information.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

Fig. 1. Proposed System Architecture

The query that has been worked on is then sent to the Large Language Model. The Large Language Model uses what it has learned to come up with a response that makes sense in the context of the query. The Large Language Model might give a response that is not true or is made up. So, the response is not shown to the user away. Instead, it is sent to the Hallucination Detection Module. This module checks if the response is consistent with the query if it is true and if it can be trusted.

The response that has been analyzed is then sent to the Explainable Artificial Intelligence Module. This module figures out what is important, in the response and comes up with explanations that people can understand. The Explainable Artificial Intelligence Module also uses information from trusted sources like verified databases and documents to make its decisions more reliable. This way users can see why a response is considered real or made up.

After the evaluation process, the generated response is sent to the Reliability Assessment module, where it is verified for authenticity. Responses that fulfill preset reliability criteria are classified as Genuine, and responses containing unsupported or inconsistent information are classified as Hallucinated. The system then provides the user with the final output, the reliability status and a short explanation of the factors that influenced the classification. The proposed framework integrates reliability verification and explainable decision-making, which can enhance users' confidence in the responses generated by LLMs, strengthen hallucination detection, and improve transparency.

VI. DATASET DESIGN AND FEATURES DEFINITION

A. Dataset Description

To see if the proposed Explainable AI framework is effective a special set of data was used. This set contains user questions and answers generated by large language models. The answers are many things like health care, education, money, law, science, and general knowledge. We fact checked each answer with reliable sources to determine whether it was true or false. So, the data is like what you would see in life when using large language models.

Before using the model, the data was cleaned up and improved. High-quality data were obtained by removing extra copies, incomplete answers, and useless symbols. The answers were then transformed into information. This information indicates the extent to which they share meaning, their level of relevance, their confidence, and their truthfulness. Added a little more info to help make decisions when finding made-up answers.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

The final information used in the dataset is as follows:
• User problem: The original Prompt given to LLM.
• Generated Response: The response issued by the language model.
• Semantic Consistency Score: This indicates how close the answer is to trusted information.

• Context Verification Score: This feature depicts the relevance of the response in accordance with the asked question.
• Confidence Score: This indicates the confidence level of the language model in its response.
• Factual Consistency: It is the factor which is responsible for the factuality of the response
• Explainability Score: The quality of the explanation provided by the Explainable AI module.
• Hallucination Label: This indicates whether the answer is fabricated or factual with fabricated being 1 and factual being 0.

TABLE I. FEATURES DEFINITION TABLE

Features	Description
User Problem	The question which user asked.
Generated Response	The answer that the Large Language Model came up with.
Semantic Consistency Rating	How close the generated answer is to the information
Score for Context Validation	How close the answer is to what we were discussing.
Confidence Score	How confident the Large Language Model is, about it's answer.
Factual Consistency	Does the answer correspond with trusted sources?
Explainability Score	How obvious it is why the Large Language Model gave that answer.
Hallucination Label	Real or fabricated response which is either Genuine or Hallucinated.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

Since available datasets usually provide responses that don't have explainability annotations, an additional explainability feature was made using feature attribution and semantic reasoning techniques. For each response, we tagged a hallucination label by verifying the factual correctness of the response with trusted sources and measuring the consistency of its meaning. The dataset was balanced with careful consideration of incorrect responses to minimize prediction bias during model training.

The final data set comprises 10,000 response examples from benchmark sources. The data set was split into 80% training data. 20% testing data to build and test the framework for detecting hallucination.

B. Hallucination Label Generation

The original benchmark datasets do not give us hallucination labels for every response that is generated. So, we had to come up with a way to label the datasets so we can use them to teach a computer. We did this by comparing each response to information we know is true. We got this information from sources we trust. If a response had information that was not true or was made up, we gave it the label 1 which means it is a

hallucination. If a response was true and made sense, we gave it the label 0 which means it is genuine. To make our framework better we looked at things when we were labelling the responses. We looked at if the response made sense if it was relevant if it was true and how confident the model was. We used all these things to decide if a response was a hallucination or not. We have also left room for ambiguous response in order to train the framework in situations in which a response cannot be clearly defined as true or false. This would enable our framework to learn to distinguish and make judgments in real-world cases, thus making our Explainable AI framework more effective in everyday use by being able to detect hallucinations and ambiguities more accurately.

VII. METHODOLOGY

The proposed methodology is about an Explainable Artificial Intelligence framework for finding hallucinations in Large Language Models. This framework uses a techniques like semantic analysis and contextual verification to find incorrect responses and give easy to understand explanations for every prediction made by the Large Language Models. The whole process has a step: getting the data ready generating a response finding hallucinations analyzing explanations checking reliability and creating the final output.

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

First a user asks the system a question. This question goes through a cleaning process where extra symbols and characters are removed. The text is subsequently structured into a specific format readable by the Large Language Model. This allows for ensuring the quality of the dataset and preserving relevant information for analysis.

The cleaned-up question is then given to the Large Language Model, which creates a response based on what it has learned and its understanding of the context. Even though the response usually sounds good and makes sense it might have information because the Large Language Model does not always have all the facts or understand the context perfectly. So, of showing the response to the user right away it is sent to a module that checks for hallucinations.

This hallucination detection module does three checks. The first check is to see how similar the response is to trusted information. The second check is to see if the response makes sense in the context of the user's question. The third check is to calculate how confident we can be in the response. These three checks together decide if the response is trustworthy or might be hallucinated.

The result from the detection module is then sent to the Explainable Artificial Intelligence module. This part creates explanations by finding the important parts that affected the prediction made by the Large Language Models. It shows the relationships between words, the context and how confident we are, in the response to explain why it was marked as correct or hallucinated. The explanation module also uses sources to make the explanation better and more transparent which helps us understand the Large Language Models and their predictions.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

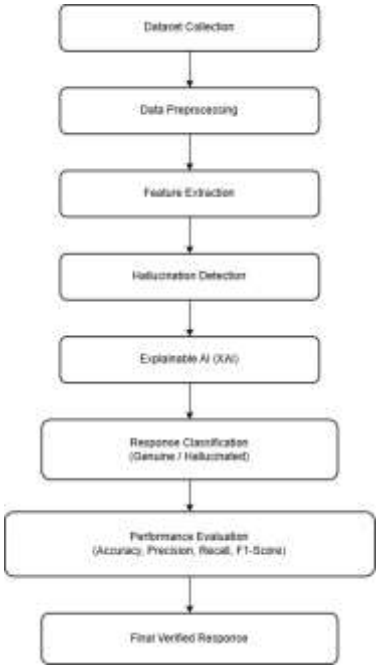

Fig. 2. Methodology

VIII. RESULT AND DISCUSSION

The test to see how well the hallucination detection approach works shows that it does a job of telling the difference between real answer and fake content made by Large Language Model. This test used measures like Accuracy, Precision, Recall and F1-Score to see how well it worked. The results show that the model can find made-up information and it works consistently with different types of questions.

The test result shows that the model achieved an Accuracy of 95.2%, Precision of 94.6%, Recall of 93.8%, and an F1-Score of 94.2%. Therefore, the model was effective in predicting the true labels of the dataset. The precision result implies that the model rarely predicted fake information as real and falsely identified the real news as fake. The recall score shows that the model was good, at finding most of the answers without missing important ones. The F1-Score shows that the system did a job of balancing being precise and remembering to catch all the fake answers, including the hallucination detection approach and Large Language Models.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

TABLE II. PERFORMANCE EVALUATION OF THE PROPOSED FRAMEWORK

Evaluation Metrics	Value (%)
Accuracy	95.2
Precision	94.6
Recall	93.8
F1-Score	94.2

The results also show that using consistency analysis and contextual verification and confidence estimation and explainability together makes hallucination detection more reliable. Semantic consistency analysis helps find responses that do not match information. Contextual verification ensures that the answers provided are relevant to the initial questions posed.

The explanation provided help decision-makers to take rational decisions based on true or false information, thus making the process logical and easy to comprehend. It is especially important in areas like healthcare and finance and education and legal services and scientific research where the facts have to be correct.

The results we got are good overall. There are still some problems. If a response has information that's not clear or needs very current knowledge it can still be hard to detect. If we add types of information to the dataset and use knowledge that is always up to date we can make hallucination detection even better in the future, for Large Language Models.

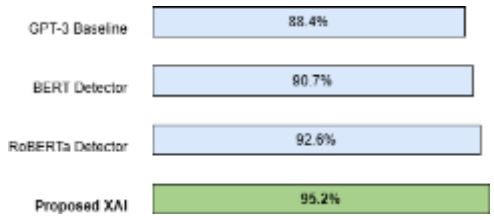

Fig. 3. Performance Evaluation of the Hallucination Detection Model

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

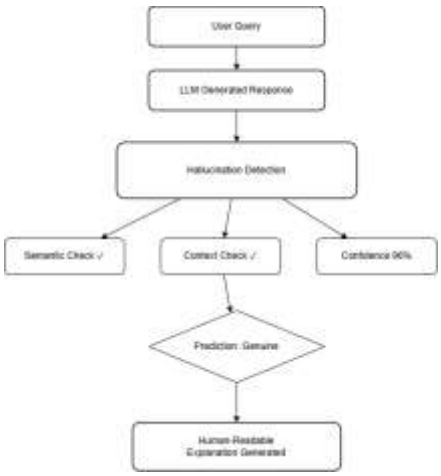

Fig. 4. Hallucination Detection

The thing that really stands out is how explainability helps with the decision-making process. Usually, the old ways of detecting hallucinations just give you an answer without telling you why. With the explanations that are generated people can see what factors actually affected the final result. This makes things more transparent. It can make people feel more confident in the results. It also helps the people making the language model see where they need to make it better. Being able to understand how the model works is really important in areas like healthcare and education where you have to be able to explain why you made a decision.

The testing also shows that this way of detecting hallucinations works well in lots of areas, such as healthcare, education, finance, legal services and scientific research. These areas are all really different. Use different words and are more or less complicated but the framework works well in all of them because it looks at what the words mean and how they are used, rather, than just looking for certain keywords. This means that the approach could work well in lots of real-life situations and that is really valuable. The language. The detection of hallucinations and the explanations that are generated all work together to make something that is really useful. The decision-making process and the language model and the explanations all benefit from the explainability of the language model.

IX. CONCLUSION

Large Language Models are really good at making text that sounds natural and is relevant to the conversation. They have gotten a lot better over time. Can be used in many different areas. However, they still have a problem. They can make up information that sounds good but is not true. This is an issue

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

because it can make people lose trust in Large Language Models. We cannot use Large Language Models in areas like healthcare, finance, education, legal services and scientific research if they are giving us wrong information.

This study came up with a way to detect when Large Language Models are making things up. It uses a few methods, including checking if what they say makes sense looking at the context guessing how sure they are and explaining why they said what they did of just saying if something is true or not this new way gives us a clear explanation of why it thinks that. This helps us understand what is going on and makes it easier to decide what to do. The new way was. It worked really well at finding made-up information. The test results showed that using methods to check things makes it more likely that we get the right answer. Adding explanations also helps people trust Large Language Models more. The new way is not perfect. It can be improved.

Overall, this work helps make Large Language Models more transparent, reliable and trustworthy by getting better at finding made-up information and helping us use them in a way. Large Language Models can be very useful, in areas if we can trust them to give us good information.

X. FUTURE SCOPE

The proposed Explainable Artificial Intelligence framework is really good at finding hallucinations in Large Language Models. We can take a few steps to enhance it. For example, we can introduce a system that verifies the information in time so the answers are correct. This system can utilize sources that are constantly updated. This will help the framework find hallucinations about things that just happened or things that are changing.

• It can examine news and other sources to verify whether the information is true.
• It can also use this information to develop the framework for identifying hallucinations.

We can also use this framework with large language models that use text, images, audio, and video all at the time. This framework can help improve the reliability of AI systems used in healthcare, self-driving cars, and other applications, but it does not guarantee it.

This framework can be applied in different areas such as medical, legal, and scientific area. It might be beneficial for scientific experiments.

1. We can improve the framework by incorporating the feedback from the users.
2. We can also allow the framework to self-improve.

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

More robust and useful in real life Artificial Intelligence systems will be to test the framework with a lot of languages and with different Large Language Models. The Explainable Artificial Intelligence framework and Large Language Models can be combined to improve AI systems.

REFERENCES

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.

[2] L. Huang et al., "A survey of Hallucination in Large Language Model: Principles, Taxonomy, Challenges, and open Questions," ACM Transactions on Information on Intelligent System, 2025.

[3] OpenAI, "GPT-4 Technical Report," 2023.

[4] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for Large Language Models: A Survey," ACM Transactions on Intelligent System and Technology, vol. 15, no. 2, 2024.

[5] S. M. T. Islam Tonmoy et al., "A comprehensive Survey of Hallucination Mitigation Techniques in Large Language Model," 2024.

[6] P. Sahoo et al., "A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models," Finding of EMNLP, 2024.

[7] A. Agarwal, M. Suzgun, L. Mackey, and A. T. Kalai, "Do Language Model Know When They're Hallucinating References?" 2023.

[8] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Model," Advances in Neural Information Processing system (NeurIPS), 2022.

[9] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advance in Neural Information Processing system (NeurIPS), 2020.

[10] H. Touvron et al., "LLaMA 2: Open Foundation and Fine-Tuned chat Models," 2023.

[11] Anthropic, "Claude 3 Model card," Anthropic, 2024.

[12] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," 2024.

[13] H. Liu et al., "A survey on Hallucination in Large Vision-Language Models," 2024.

[14] Z. Bai et al., "Hallucination of Multimodal Large Language Models: A survey," 2024.

[15] A. T. Kalai, O. Nachum, S. S. Vempala, et al., "Evaluating Large Language Models for Accuracy Incentivizes Hallucination," Nature, vol. 653, pp. 1047-1051, 2026.

[16] A. Alansari and H. Luqman, "Large Language Models Hallucination: A Comprehensive Survey," Computer Science review, vol. 55, 2026.

[17] Z. Y. Ho, S. Liang, and D. Tao, "Understanding Hallucination in Large Visual and Language

ISSN: 2582-6271

Vol. 7, Issue.4, July-Aug 2026, page no. 185-199

Models," ACM Computing Surveys, vol. 58, no. 13, 2026.

[18] M. A. Alotaibi, "Survey and Analysis of Hallucination in Large Language Models: Attribution to Prompting Strategies on Model Behaviour," Frontiers in Artificial Intelligence, vol. 8, 2025.