International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EXPLORING THE BOUNDARIES OF ARTIFICIAL INTELLIGENCE HALLUCINATION

Praveen Gujjar J[1], Prasanna Kumar H R[2]

[1]Research Scholar, Visvesvaraya Technological University, and Faculty of Management Studies,

JAIN (Deemed-to-be University), Bengaluru, India

e-Mail: dr.praveengujjar@cms.ac.in

[2]Dept of Computer Science and Engineering, Visvesvaraya Technological University,

PESITM, Shivamogga, India

e-Mail: hrpbhat@gmail.com

Abstract

AI systems are being increasingly used to perform a wide variety of tasks. Large organization typically build foundation models, which are models trained on very large sets of data. These models are then fine-tuned for specific tasks in specific domains. A major issue with using such systems are hallucinations which refers to issues like incorrect detection, incorrect fact generation, etc. Such issues have been widely studied for identifying potential mitigation. In this study, we come up with a two category classification for the causes of hallucination. We then discuss about the ways to handle such hallucination. We briefly look at ways to identify hallucination and then discuss potential future research areas to reduce the hallucination.

Math. Subj. Classification 2020: 68T37, 97P80

Keywords— AI, Large Language Model, Hallucination

Introduction

The term Artificial Intelligence(AI) hallucination was coined in the context of image/object recognition, wherein AI systems incorrectly identified objects that were not actually present in images. With time, specifically with Large Language Models coming into the picture, the term was expanded to denote any incorrect, misleading or nonsensical output generated by AI models. Scholars over time have identified several different causes behind AI hallucination. Most of the AI systems found today are based on foundation models. Foundation models are broad models trained on very large datasets, which are later fine-tuned for specific applications. For example, LLaMA, DALL-E, and GPT are some well-known foundation models that are being used worldwide for different purposes. Hallucination is a problem that has been experienced in different application areas of foundation models. The extent of the problem that

Received: July 25, 2025

occurs due to hallucination varies across domains. For example, while analyzing images in healthcare, hallucination can have disastrous consequences, while somebody submitting a school/college assignment will have less severe consequences. We discuss hallucinations in different domains in later sections. Since hallucinations can cause large-scale problems, scholars have devoted a lot of time to study these hallucinations across different systems. Studies have tried to categorize sources of hallucinations to have a better understanding of their cause Sun et al. (2024). However, in this study, we generate a classification based on stages of model creation. The purpose of this is to suggest remedial measures to prevent and detect hallucination during model training and or usage. We propose that the sources of hallucination can be categorized into two broad categories:

1. Issues with training (during foundation model creation)
2. Issues with usage (during model fine tuning and usage)

Issues with training include (but is not limited to) less availability of data, use of biased/incorrect data for training, etc., lack of contextual information and overfitting of models. Issues while usage mostly include giving incorrect prompts, incorrect labelling of training data, etc.

Furthermore, hallucination leads to reduction in trust of AI systems leading to less adoption. Hence we also summarize ways to mitigate and reduce hallucination.

The major contributions of this study are:

1. Identification of sources of hallucination for object detection and large language models
2. Listing of mitigation strategies and future rese

We first discuss the basic idea of hallucination and then discuss the basic steps in the creation of foundation and fine-tuned models. We then list the causes of hallucination in these two domains. Finally, we look at suggested mitigation strategies for the identified sources of hallucination.

Related Work

The exact definition of hallucination has been widely debated by academicians Maleki et al. (2024). AI hallucination refers to situations when AI systems generate inaccurate and incoherent output. It may include identifying objects that are not present in an image or generating textual content that is not factually true.

Hallucination in context of object detection:

In context of image processing, one of the most important areas in which hallucination is observed is that of face recognition. In this context, face hallucination is a positive term which refers to a set of techniques to handle unclear or low resolution images to properly identify faces. Hallucination is also frequently observed in image recognition tasks like image captioning that requires computers to identify objects in an image to give the image proper caption. Algorithms often detect non-existing objects in images which has been termed as

Received: July 25, 2025

hallucination Kayhan et al. (2021). Once generative AI systems, came up, the term hallucination started gaining much more research prominence. Generative AI is AI AI-based system that can generate new content based on a user's prompt.

Specifically, the last few years have seen the rise of what are termed as foundation models. Foundation models are trained on a broad dataset and can be retrained for performing specific tasks Bommasani et al. (2021). While foundation models allow retraining of models for various types of tasks, LLMs or large language are usually trained to generate new text. These systems are also prone to hallucination in the sense that they generate incorrect or incoherent output.

We discuss some major areas in which hallucinations have been detected and studied by scholars. In the legal domain, there have been recorded instances of AI systems referring to fictitious cases 18. In the healthcare domain, a study was done to check instances of hallucination in scientific writing. An experiment showed that scientific papers generated using AI models generated discrepancies in the output. AI foundation models are also used for image processing, classification, and segmentation tasks. Studies have found models hallucinating while performing such tasks.

AI models are also being used in areas such as cybersecurity, wherein they have to perform tasks like the separation of malicious security breaches from non-harmful intrusions Sood et al. (2025). Foundation models are also used for activities like customer service in marketing, wherein chatbots are trained to respond to customer queries and issues Yaprak (2024). AI hallucinations are widely studied in different domains as the presence of hallucinations makes it difficult to trust the output generated by AI systems completely. In areas of healthcare, wrong diagnostics will create problems for the doctors and patients. Customers when interacting with AI-based systems, often will not trust the recommendations provided by AI systems. In the next section, we briefly describe the architecture of AI foundation models so that we can understand errors at which stage causes the systems to hallucinate.

Foundation Model Training

Different types of architecture of foundation models are found in the literature. Training foundation models is a complex, resource-intensive process that combines advanced machine learning techniques with large datasets to create versatile AI systems capable of handling a wide range of tasks. Training foundation models is computationally expensive and requires significant resources, including high energy consumption. The major steps involved in training a foundation model are:

1. Data Collection
2. Data Preprocessing
3. Model Architecture Setup
4. Pre-training
5. Fine Tuning
6. Evaluation and Testing

Received: July 25, 2025

The architecture of foundation models is shown in Figure 1. The first step is to collect data to train the model. As mentioned, training Foundation models requires vast amounts of data, typically sourced from the internet, including text from websites, books, articles, forums, and other publicly available content. Using techniques like scraping etc., data is collected for model training. The collected raw data is then cleaned to remove noise, irrelevant content, and potentially harmful or biased information. This process is also termed as data curation. The text data is tokenized, breaking down sentences into smaller units (words). The architecture for foundation models is usually a combination of adversarial networks, transformer networks, etc. The model is built with multiple layers of transformers, each containing attention mechanisms and feed-forward neural networks. The depth and size of the model depend on the computational resources available and the desired model capability. Because of the first dependency we see foundation models are mostly developed by large organizations.

The model is then trained using self-supervised learning techniques. Pre-training models like GPT has billions of parameters, which requires powerful GPUs/TPUs and distributed computing setups. Training can take weeks or months, depending on the model's size and complexity.

The model is evaluated on various benchmarks to test its performance on different tasks, ensuring it generalizes well and meets the desired accuracy or quality standards.

Using Foundation Models

The pre-trained foundation models are then fine-tuned to perform specific tasks that can be generative (writing a paragraph) or non-generative (sentiment classification).

This second level training (or fine–tuning) to teach the model to perform a specific task are mostly done using supervised and/or semi-supervised methods. Finally, the fine-tuned models are evaluated and tested for their accuracy.

The complete process is shown in Figure 1.

Received: July 25, 2025

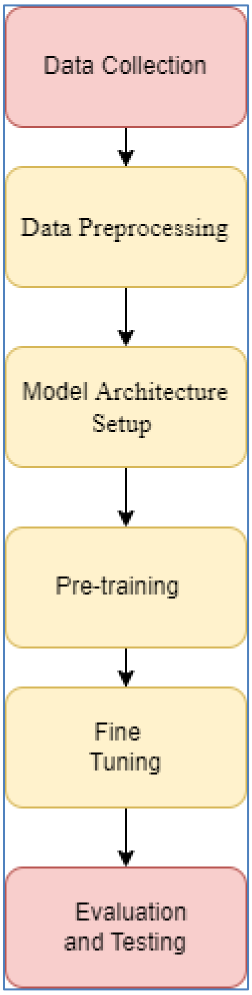

Figure 1. AI Foundation Model Training Process

Causes of AI Hallucination

Issues during Training:

Data has been identified as a major contributor to hallucination in the context of large language models Huang et al. (2024). Both quality and quantity of data have been found to influence hallucination.

1. Training from Incorrect data – To train foundation models often data from the internet is often collected. The problem with such data is that checking and verifying the facts of all the data collected is difficult. Challenges of fake news etc. increase the chances of getting incorrect data.
2. Training from Biased data – This is another challenge that is very difficult to detect. Specifically, the presence of undetected bias in the training dataset becomes a problem. For example, hate speeches that are not flagged, resumes of employees rejected due to bias etc. usually remain unidentifiable when these data are used for training it can generate biased content.

Received: July 25, 2025

3. Training with incomplete data – While training foundation models, many a times data till a particular date is used. Data till that particular date might be incomplete. While the next round of training will address the problem, using the system before the next iteration might give improper results.
4. Coverage of data – As mentioned, AI foundation models are trained with a broad dataset. However, limitations in collection make it unfeasible to collect data pertaining to all domains. If the model is later retrained with data in a relevant context, this limitation can be overcome to some extent.

The next major contributor to hallucination is the training process itself. Some of the major issues that result in hallucinations include:

1. Over-fitting – Over-fitting occurs when the model adapts to the training data so well that it is not able to perform well or test data. Overfitting often causes models to give incorrect output.
2. Under fitting – The opposite of overfitting, under fitting, or very less learning from data, also causes foundation models to hallucinate.

Both overfitting and under fitting makes the trained models unsuitable for generalization of the model to solve a large variety of problems.

Issues with usage

1. Giving improper prompts – Users who are not properly trained often can give incorrect prompts that will make the system give incoherent/incorrect output.

Scholars have identified several ways through which issues with prompts can be addressed through the training of users to provide adequate and sufficient prompts.

Addressing AI Hallucination

The following are some of the difficulties that foundation models face: Infrastructure needs - It is costly and time-consuming to create a foundation model from scratch, and training could take several months. Development of the front end. Developers must incorporate foundation models into a software stack that includes pipeline engineering, prompt engineering, and fine-tuning tools in order to create useful applications. Inability to understand. While foundation models can provide factually and grammatically correct answers, they struggle to understand the prompt's context. Furthermore, they lack psychological and social awareness.

untrustworthy responses. Responses to enquiries regarding specific topics may not always be trustworthy and occasionally be unsuitable, harmful, or inaccurate. Prejudice. Given that models can extract offensive language and inappropriate undertones from training datasets, bias is a real possibility. Developers should carefully filter training data and incorporate particular norms into their models to prevent this. Experimental studies demonstrated that AI hallucinations occur with varying frequency depending on the complexity of the task, the quality of input prompts, and the specific architecture of the AI model used. Hallucinations

Received: July 25, 2025

were more frequent in tasks involving ambiguous or open-ended prompts, where the AI model lacked sufficient context to generate accurate responses. Specific domains, such as creative writing or speculative responses, exhibited higher rates of hallucination compared to fact-based tasks like question answering or summarization. Larger models with more extensive training datasets showed a reduction in the frequency of hallucinations, but they were not entirely immune. The study found that while larger models were better at contextual understanding, they still occasionally produced confident but incorrect information, particularly in less common scenarios not well-represented in the training data. The results indicated that hallucinations often stem from the model's inability to maintain consistent context over long text sequences. Models that were prompted with detailed, specific information produced fewer hallucinations compared to those prompted with vague or broad inputs. Various strategies to mitigate hallucinations were tested, including prompt engineering, user interaction design, and post-generation validation techniques. Among these, prompt engineering—crafting more precise and contextually rich inputs—proved to be the most effective in reducing hallucinations. Additionally, incorporating human oversight and post-processing checks significantly improved the reliability of AI-generated content. Causes and steps for mitigating AI hallucinations are listed in Table 1.

Table 1 AI Hallucinations: Causes and Mitigation steps

SL. No	Hallucination	Handling
1	Data Issues	As discussed, the major issue here is incorrect or incomplete data. This can be handled in different ways. One way is to have systems like RAG, which can validate the information in real time. The second method is to have multiple systems with voting to determine what has to be done.
2	Model Overconfidence	Foundation models often generate content with high confidence which makes it difficult for users. (the confidence can be found out by asking the model to print its perceived confidence). Several methods have been proposed to mitigate, including that of using distractors, and post hoc calibration methods like temperature scalingChhikara (2025) .
3	Issues with prompts	Users often give prompts which are either incorrect or sometimes are misleading or incomplete. Studies have posited that prompts should be carefully typed with proper information about the problem to solve, the tone of the solution, context of the problem.
4	Overgeneralization	Overfitting and under fitting can lead to generalizability issues, wherein if a model is asked to solve a problem, it

Received: July 25, 2025

		might give an incorrect result. To prevent such issues, while training and subsequently fine-tuning the foundation models, model accuracy parameters should be monitored.
5	Long-Context Limitations	Large models can struggle to maintain context over long stretches of text. As the conversation or input gets longer, the model may lose track of key information, leading to inconsistent or hallucinatory outputs. Such issues can be tackled by providing the context at regular intervals.
6	Unsupervised Learning Nature	Foundation models are trained using self-supervised training methods which essentially means that no annotation takes place by the trainer. While the large dataset used makes human annotation impossible, this also leads to situations wherein the model learns incorrectly and generates hallucination. Scholars have also tried to use semi-supervised methods to reduce model bias 22.
7	Model Complexity	While complex transformer-based architectures enable powerful AI models, their reliance on statistical correlations rather than true comprehension can lead to errors when predictions are not strongly grounded in fact.

Hallucination and fairness

Fairness in intelligent systems has been researched for a long time. Fairness is of two broad categories: individual and group fairness (Wang, N., Tao, D., Gao, X., Li, X., & Li, 2014). The basic idea of fairness is that there should be no discrimination between individuals or groups based on any attribute. Hypothetically, unfairness can creep in two ways. During training, bias in training data is a major cause for hallucination in AI systems, and such hallucinated output will not be fair. The problem is specifically important since self-supervision of learning prevents the system from identifying potential sources of bias. The next source of bias will be prompts. If a user a biased, she will give biased prompts, which will lead to the model generating biased content. Fairness can be ensured only by making the AI foundation models more resilient to hallucination.

Detecting AI Hallucination

Scholars have come up with several different methods to detect hallucination. Most of the studies focus on detecting hallucinations by using other models. For example, GPT-4 has been extensively used to detect hallucination in outputs 26. Tools like WebGPT and Google's search augmented language models are used to verify facts of content generated by AI. Organizations like Fujitsu maintain continuous monitoring of AI generated content to check for anomalies.

Received: July 25, 2025

RAG or Retrieval Augmented Generation systems are now being extensively used to validate facts generated by AI systems.

Future research direction

AI hallucination is an important topic that has to be addressed so that AI adaptation becomes more trustworthy. Detection of hallucination has to take place after generation of the content. It is better if, during training, we can check for potential future hallucinations. Research can be done on ways to reduce potential hallucinations.

During Training – Subject to availability of data, we can create AI models to see whether we can predict from training data that the trained model is going to hallucinate or not. Further, as mentioned, the foundation models are trained using self-supervised learning. If in some way semi supervised learning can be introduced even with a small dataset, hallucination is expected to come down. Research has to be done to check the impact of such training on cost, time and other aspects of development of foundation models.

During Usage – Training users so that they know what prompts to give is an important way to mitigate hallucination. While scholars are looking at different prompt engineering measures, research also has to be done on using intelligent systems to fine-tune prompts. Such systems if integrated with foundation models, will enhance the model's robustness to incorrect prompts. RAG systems can be used to fetch context and provide augmented input base to models based on user's input.

A key aspect of foundation models is the cost of development of these models. It will be highly expensive to test these changes in the models directly. These enhancements can be introduced to smaller models to check their efficiency and once certain, they can be used for training foundation models.

Conclusion

The research underscores the importance of understanding and addressing AI hallucinations, particularly as AI systems become more integrated into critical decision-making processes. While larger and more sophisticated models tend to hallucinate less frequently, the issue persists and poses risks in high-stakes environments such as healthcare, legal advice, and finance. Mitigation strategies, especially those focusing on improved prompt design and human-in-the-loop systems, are essential for minimizing the impact of hallucinations. The study categorizes the sources of hallucination into two major categories. Such a categorization can be helpful in designing hallucination detection and mitigation strategies. Future research should continue to explore advanced techniques for hallucination detection and prevention, aiming to develop AI systems that are not only powerful but also reliable and safe for widespread use. If there is no reinforcement learning mechanism or feedback loop to correct wrong outputs during training, the model may continue to generate hallucinatory information without internal checks for accuracy. If the model is presented with inputs or queries that fall outside the distribution of

Received: July 25, 2025

data it was trained on, it may hallucinate because it cannot effectively generalize to those unfamiliar cases.

References

[1] Wang, N., Tao, D., Gao, X., Li, X., & Li, J. (2014). A comprehensive survey to face hallucination. *International journal of computer vision*, *106*, 9-30.

[2] Kayhan, O. S., Vredebregt, B., & Van Gemert, J. C. (2021). Hallucination in object Detection — A study in visual part VERIFICATION. 2022 IEEE International Conference on Image Processing (ICIP), 2234–2238. https://doi.org/10.1109/icip42928.2021.9506670

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2108.07258

[4] Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation models and fair use. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4404340

[5] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Transactions on Office Information Systems. https://doi.org/10.1145/3703155

[6] Kaate, I., Salminen, J., Jung, S., Xuan, T. T. T., Häyhänen, E., Azem, J. Y., & Jansen, B. J. (2025). "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. You Always Get an Answer": Analyzing Users' Interaction With AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination, 1624–1638. https://doi.org/10.1145/3708359.3712160 In *Proceedings of the 30th International Conference on Intelligent User Interfaces* (pp. 1624-1638).

[7] Chhikara, P. (2025). Mind the confidence gap: overconfidence, calibration, and distractor effects in large language models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2502.1102818.

[8] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*.

[9] Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating

Received: July 25, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. https://doi.org/10.7759/cureus.37432

[10] Sood, A. K., Zeadally, S., & Hong, E. (2025). The Paradigm of Hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations. Computers & Electrical Engineering, 124, 110307. https://doi.org/10.1016/j. compeleceng.2025.110307

[11] Yaprak, B. (2024). Generative Artificial intelligence in Marketing: The Invisible Danger of AI hallucinations. Journal of Economy Business and Management, 8(2), 133–158.https://doi.org/10.7596/jebm.158889722.

[12] Gan, K., & Wei, T. (2024). Erasing the bias: Fine-tuning foundation models for semi-supervised learning. *arXiv preprint arXiv:2405.11756*.

[13] Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: a misnomer worth clarifying. Ieee. https://doi.org/10.1109/cai59869.2024.00033

[14] Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications, 11(1). https://doi.org/10.1057/s41599-024-03811-x

[15] Pessach, D., & Shmueli, E. (2022). A review on Fairness in Machine Learning. ACM Computing Surveys, 55(3), 1–44. https://doi.org/10.1145/3494672

[16] Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., & Zhu, L. (2025). Detecting and mitigating hallucination in large vision language models via Fine-Grained AI feedback. Proceedings of the AAAI Conference on Artificial Intelligence, 39(24), 25543–25551. https://doi.org/10.1609/aaai.v39i24.34744

[17] Islam Tonmoy, S. M. T., Zaman, S. M. M., Vinija Jain, Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models.

Received: July 25, 2025