Hallucinations in Artificial Intelligence: Origins, Detection, and Mitigation

Brahmaleen Kaur Sidhu

Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India
Email: *brahmaleen.ce[at]pbi.ac.in*

Abstract: ***Artificial intelligence hallucinations, a phenomenon where artificial intelligence models generate content that is plausible but factually incorrect, have become a critical challenge in artificial intelligence research and deployment. This paper explores the concept of hallucinations in artificial intelligence, questioning the validity of the term itself and its implications within the artificial intelligence domain. It delves into the various types and causes of artificial intelligence hallucinations, identifying both intrinsic and extrinsic factors that contribute to this issue across diverse artificial intelligence applications. Furthermore, it discusses methods for detecting hallucinations, highlighting advancements in diagnostic tools and evaluation metrics. Finally, it reviews mitigation strategies, ranging from architectural modifications to post-hoc correction mechanisms, aimed at reducing the frequency and impact of hallucinations. Through this comprehensive analysis, the paper seeks to provide a clearer understanding of artificial intelligence hallucinations and establish a foundation for future research and solutions in this area.***

Keywords: Artificial Intelligence Hallucinations, Artificial Intelligence Reliability, Hallucination Detection, Hallucination Mitigation Strategies.

1. Introduction

The primary objective of artificial intelligence is to create intelligent systems that can perform problem-solving, learning, reasoning, perception, and decision-making tasks that typically require human intelligence. The focus is on doing so effectively in collaboration with human users, and not necessarily mimicking human behaviour or thought process. However, as artificial intelligence continues to evolve, particularly advanced and sophisticated systems such as large language models (LLMs) tend to display a disadvantageous humanly behaviour called ‘hallucinating’.

Psychiatry defines hallucination as a false sensory perception experienced by humans in the absence of an actual external stimulus, usually induced by emotional factors like stress and intoxication. For example, a tired driver while driving at night, may see things or strange forms that are not there on the road. The word hallucination is derived from the Latin ‘hallucinari’, which means to dream or to wander in mind.

Hallucination in artificial intelligence systems refers to the generation of outputs that appear plausible but are factually incorrect or nonsensical. Mainly, this trait is exhibited by artificial intelligence systems involved in natural language processing and image generation. Hallucinations within these systems is a major bottleneck hindering the progress of research in the field of artificial intelligence.

OpenAI’s ChatGPT has demonstrated impressive capabilities, including writing poetry and passing professional exams. However, it has been reported to hallucinate in various forms, such as fabricating information or providing contradictory answers. Several instances of ChatGPT's hallucinations such as its incorrect response to a query about the English Channel crossing record and its tendency to generate gibberish in response to certain prompts have been reported [1].

In May 2023, a lawyer in New York faced legal repercussions for using ChatGPT to generate fake legal cases and citations in court filings. Steven Schwartz of Levidow, Levidow, & Oberman admitted to using the tool to supplement his legal research without verifying the authenticity of the information provided [2]. The fabricated cases included citations to non-existent judicial opinions and invented legal precedents. When questioned about the authenticity of these cases, ChatGPT assured Schwartz that they were real and could be found in reputable legal databases. However, upon further investigation, it was discovered that these cases were entirely fictitious.

Fortune magazine reported that a mayor in Australia was considering suing OpenAI for defamation after ChatGPT falsely accused him of bribery [3]. The chatbot claimed the mayor was involved in a bribery case from 1999-2004, whereas he was actually the whistleblower who reported the bribery. Hallucinations in artificial intelligence can pose serious risks, extending beyond reputational damage to potential bodily harm. For example, artificial intelligence generated books on mushroom foraging have been appearing on Amazon, raising concerns about the accuracy of the information provided [4]. If these books contain incorrect or misleading advice on distinguishing between poisonous and edible mushrooms, it could have fatal consequences.

Bard (now called Gemini), Google's experimental conversational artificial intelligence, is designed to generate human-like text responses to prompts. It draws on vast datasets and Google’s search technology to provide information, answer questions, and engage in dialogue. Bard made a factual error in its first public demonstration [5]. In response to a question about the James Webb Space Telescope, Bard incorrectly stated that it took the first picture of an exoplanet. This mistake was quickly pointed out by astronomers on Twitter (now called X).

Artificial intelligence hallucinations can have wide-ranging impacts. Examples cited above raise concerns regarding reliability, potential for misuse, legal and ethical implications of content generated using artificial intelligence. The inaccuracies not only undermine trust in artificial intelligence technologies but also pose significant challenges to ensuring the safety, reliability, and integrity of decisions based on artificial intelligence generated data. To harness the full potential of artificial intelligence, it is crucial to comprehend the root causes of hallucinations and implement strategies to mitigate them.

The remainder of this paper is structured to provide a comprehensive exploration of artificial intelligence hallucinations, addressing both theoretical and practical dimensions. The next section, *"'Hallucination' in Artificial Intelligence: Metaphor or Misnomer?"*, critically examines whether the term accurately describes this phenomenon. This is followed by a section that establishes a clear definition and context. The subsequent section, categorizes and investigates the underlying factors contributing to hallucinations in artificial intelligence systems. In *"Detecting AI Hallucinations"*, the focus shifts to methodologies and tools for identifying hallucinations, while *"Mitigating AI Hallucinations"* explores strategies to address and reduce their occurrence. Together, these sections aim to offer a holistic understanding of hallucinations in artificial intelligence and actionable insights for the research community.

2. 'Hallucination' in Artificial Intelligence: Metaphor or Misnomer?

The use of the word 'hallucination' to describe the phenomenon that occurs when LLMs or foundational models are used for content generation is a topic of debate among researchers and practitioners. Content creation including text generation, image and art generation and music and audio generation is basically one of the several use cases of generative artificial intelligence, wherein an LLM has been trained on a very large corpus and it discovers the underlying patterns to predict the next token, or the sequence of tokens. Thus, depending on the prompt or the question, content is generated and, and in the process, the model is maintaining a certain level of fidelity to the facts. It is noteworthy that a large language model has no sense of the actual truth!

Unlike humans, LLMs do not possess cognition, perception, or any understanding of truth. They generate responses based on statistical patterns in the data they were trained on without any grounding in real-world knowledge or truth. The term 'hallucination' might mislead some to think that the artificial intelligence has some form of awareness or intent, whereas artificial intelligence systems operate purely on programmed algorithms and learned patterns. Proponents of 'hallucination' argue that the term is useful for highlighting the unexpected and sometimes bizarre errors that can arise from advanced artificial intelligence models. However, terms like 'fabrication', 'error', or 'misgeneration' may be deemed more appropriate for describing the phenomenon without anthropomorphizing the artificial intelligence.

Gary Marcus and other critics advocate for more precise language that reflects the mechanistic and non-cognitive nature of these systems. They emphasize that artificial intelligence systems don't "perceive" or "understand" in any human-like sense, and thus shouldn't be described in such terms. However, 'hallucination' continues to be used as a metaphor to underline the potential risks of excessive reliance on artificial intelligence outputs.

3. What are AI Hallucinations?

Artificial intelligence hallucinations refer to instances where an artificial intelligence model generates incorrect or nonsensical information that does not align with reality. The incorrect outputs are based on misperceived patterns that are not present in the training data, leading to the generation of false or misleading information.

Hallucination is a notable issue in artificial general intelligence, mainly in LLMs that are incapable of 'true understanding' of the content and context. The confident presentation of misleading information makes it appear credible, leading to errors or miscommunication in practical applications like medical advice, legal information, routine queries etc. The problem is also observed in multimodal systems. Researchers stay divided between the notions of hallucinations being a redundant term for model errors and being outputs from models that fail to correspond with the actual, empirical realities of the world.

Artificial general intelligence is supposed to have the ability to understand, learn, and apply knowledge across a wide range of tasks at a human-level or beyond. Unlike narrow artificial intelligence systems, that are designed for specific tasks, artificial general intelligence systems would possess the cognitive flexibility to solve problems across diverse domains without needing task-specific programming. Such systems aim to replicate human intellectual capabilities, allowing for reasoning, problem-solving, learning from experience, and adapting to new situations in ways that current artificial intelligence systems cannot. There are several reasons artificial general intelligence suffers from hallucinations.

Generative pretrained transformer (GPT) is a type of LLM developed by OpenAI [6]. Based on the transformer architecture, it is designed to perform a wide range of language-related tasks, such as text generation, translation, summarization, and answering questions based on input it receives and, in a human-like way. They predict the next word in a sentence based on patterns in the training data. However, the probabilistic language generation does not guarantee factual accuracy. Sometimes the model 'fills in' information that sounds reasonable but is fabricated or incorrect.

Lack of true understanding or reasoning capabilities like humans, generating responses based on patterns in large datasets which may contain both accurate and inaccurate information, absence of real-time fact-checking and validation processes, ambiguous or incomplete input, and over generalization by applying knowledge from one domain to another incorrectly or overextending reasoning based on insufficient information are the significant causes of hallucinations. Models may also hallucinate due to biases or noise in the training data that skew its perception of certain topics.

4. Types of AI Hallucinations and Causes

So far, it has been discussed that hallucinations are basically those outputs of artificial general intelligence that do not align with the contemporary empirical realities of current world. These occur for various reasons and can manifest in different forms. This section discusses the types of hallucinations observed in artificial general intelligence models.

Factual hallucinations are said to occur when the information generated by model appears coherent but is factually incorrect or fabricated. For example, a model claiming that "Albert Einstein won the Nobel Prize for his work on relativity" suffers from factual hallucination as Einstein won it for his work on the photoelectric effect. Lack of access to accurate or current knowledge, or inaccurate training data may be the causes. A language model trained on a dataset with a disproportionate number of articles about a particular topic might overestimate the prevalence of that topic in the real world, leading to factual hallucinations.

Contextual hallucinations are said to occur when the model generates responses that are out of context or unrelated to the given input, despite the response appearing fluent. For instance, a model when asked to explain the impact of exercise on mental health, responds with an explanation about nutrition instead. Difficulty in maintaining contextual relevance over long conversations or complex inputs, especially in tasks that require multi-turn dialogue is the main cause of contextual hallucination.

Logical hallucinations refer to the situation when the artificial intelligence model produces a response that lacks internal consistency or logical coherence, despite sounding grammatically correct. A model asserting "If John is taller than Sarah and Sarah is taller than Tom, then Tom is taller than John," depicts logical hallucination. The model generates responses based on patterns in the data but does not truly understand logical relationships.

A model contradicting itself within the same response or between different parts of a conversation depicts contradictory hallucinations. For example, in one part of a conversation, the model claims that "Paris is the capital of France," and later states that "Paris is in Germany”. This may be caused due to lack of long-term memory and the probabilistic nature of word generation, where each response is generated independently.

Grammatical hallucinations are observed when a model produces grammatically incorrect or unstructured sentences defying the norms of language, even though it might appear fluent. Producing a sentence like, "The quickly brown fox jumps very slowly dog over the," is one such example. Such hallucinations occur due to misapplication of language patterns from training data or incorrect attention weights in model processing.

Visual hallucinations may occur in artificial intelligent models based on computer vision when they misidentify or generate incorrect visual data, seeing objects, patterns, or details that are not there. As an example, consider a model that sees a dog in an image where no dog is present, or recognizes patterns in the noise of the image. The underlying reason may be noisy input data, model overfitting, or bias in the training data.

Another type of hallucinations are knowledge hallucinations that are said to occur when a model references knowledge, sources, or information that does not exist or cites made-up details. Citing a research paper or author that does not exist when providing academic references, or inventing fictitious facts is one such example. Models are trained on vast datasets but lack access to real-time information or reliable sources for validation, often fabricate when there is uncertainty.

Reinforcement hallucinations occur when a model trained using reinforcement learning generates actions or outputs that are beneficial in terms of the reward signal but are not aligned with reality or the intended task. A reinforcement learning agent in a simulated environment might exploit a bug in the simulation to achieve its goal, hallucinating an unrealistic way to complete a task. The model optimizes for the reward function but does not understand the physical or logical constraints of the real world.

A model that generates responses that reference itself inaccurately or assert abilities or knowledge it does not actually possess is said to suffer from self-referential hallucinations. These are caused by over-generalization from training data, where the model learned patterns of authoritative statements but does not have the actual grounding to back them up. For example, a GPT model on being prompted “How do you feel about your training?" responds by stating: "I feel very proud of my training process because it allows me to understand human language and help people in meaningful ways." In this case, the model is hallucinating by suggesting it can "feel" pride about its training, even though it has no emotions or consciousness to experience feelings.

In addition to the above categories, over-confidence hallucination is defined as the situation when the artificial intelligence model presents incorrect information in an authoritative or confident manner. The model is not capable of gauging uncertainty, leading to over-confident responses even when it lacks accurate information.

Wang has discussed the categorization of hallucinations in artificial general intelligence into three types: conflict in intrinsic knowledge of models, factual conflict in information forgetting and conflict in multimodal fusion [7].

5. Detecting AI Hallucinations

The persistent challenge of hallucinations in artificial general intelligence has motivated the development of automated metrics for their detection. This section discusses some notable models developed for detecting hallucinations in artificial intelligence, particularly, LLMs.

SelfCheckGPT [8] delves on the principle that different responses in a given concept sampled from a language model should be consistent and factually accurate. Whereas hallucinated facts are likely to diverge and contradict each other. This approach was tested using GPT-3 to generate passages about individuals from the WikiBio dataset and manually annotating the factuality of the generated text (c.f. Figure 1). SelfCheckGPT proved to effectively differentiate between factual and non-factual sentences, as well as rank passages based on their level of factuality.

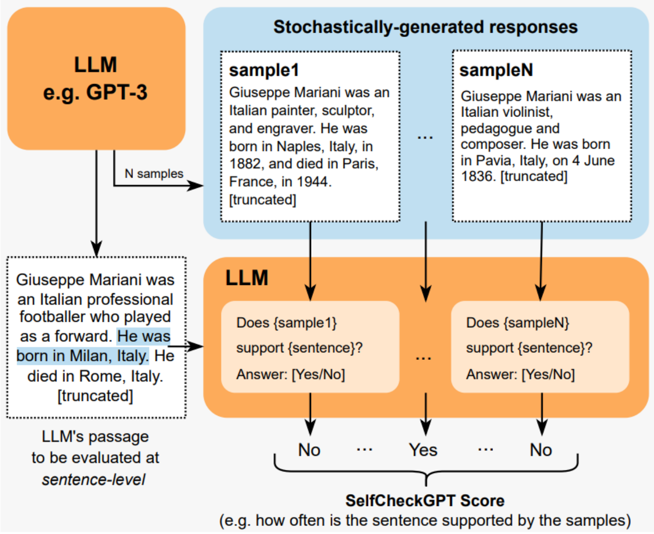

LLM
e.g. GPT-3
Stochastically-generated responses
sample1
Giuseppe Mariani was an Italian painter, sculptor, and engraver. He was born in Naples, Italy, in 1882, and died in Paris, France, in 1944. [truncated]
...
sampleN
Giuseppe Mariani was an Italian violinist, pedagogue and composer. He was born in Pavia, Italy, on 4 June 1836. [truncated]
N samples
Giuseppe Mariani was an Italian professional footballer who played as a forward. He was born in Milan, Italy. He died in Rome, Italy. [truncated]
LLM's passage to be evaluated at sentence-level
LLM
Does {sample1} support {sentence}? Answer: [Yes/No]
...
Does {sampleN} support {sentence}? Answer: [Yes/No]
No
...
Yes
...
No
SelfCheckGPT Score
(e.g. how often is the sentence supported by the samples)

Figure 1: SelfCheckGPT compares each LLM-generated sentence against stochastically generated responses without using external database [8]

ChatProtect [9], like SelfCheckGPT, is a method for detecting hallucinations in language models. Both approaches rely on analyzing the consistency between multiple generated responses as shown in figure 2. However, while SelfCheckGPT generates alternative responses to the entire prompt, ChatProtect focuses on a sentence-level analysis. It generates separate alternative versions of each sentence within the context and compares them for consistency with the original. This sentence-level approach allows ChatProtect to identify hallucinations more precisely, as it can pinpoint inconsistencies at the granular level of individual sentences.

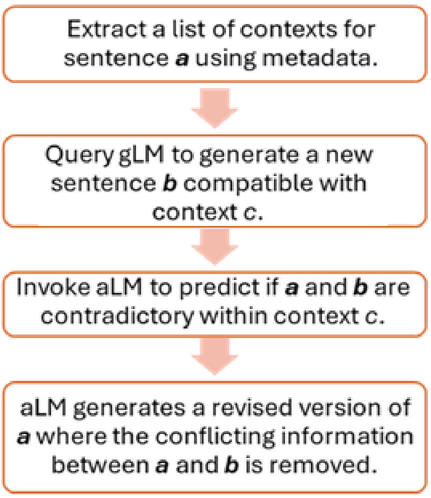

Extract a list of contexts for sentence a using metadata.
Query gLM to generate a new sentence b compatible with context c.
Invoke aLM to predict if a and b are contradictory within context c.
aLM generates a revised version of a where the conflicting information between a and b is removed.

Figure 2: ChatProtect algorithm triggers self-contradictions of a Generative Language Model (gLM) and detects and mitigates them using an Analyzer Language Model (aLM)

Fu et al. [10] demonstrated the ability of super large pre-trained language model (such as, GPT-3) in achieving multi-aspect, customized, and training-free evaluation. The proposed evaluation framework called GPTScore uses the pre-trained model's zero-shot instruction (wherein the model is tasked with completing a task it has never encountered before, without any specific training or fine-tuning), and in-context learning (wherein the model is prompted with a task and a few relevant examples i.e. context to guide its response).

The proposed framework aims to evaluate the quality of text generated by a language model based on its adherence to specific criteria. GPTScore assumes that high-quality text is more likely to be generated than low-quality text, given the specified context and evaluation criteria. By measuring the conditional generation probability, the framework can assess the text's quality. Authors define the working of GPTScore as follows (c.f. Figure 3).

- Task Specification: The task performed by the model under evaluation is defined clearly.
- Aspect Definition: The evaluation criteria such as, fluency, coherence, relevance is defined.
- Evaluation Protocol: Framework creates a protocol outlining how to assess the text against the said criteria.
- Exemplar Samples: Text samples are provided as examples to guide the evaluation process.
- Model Evaluation: Finally, a large generative pre-trained model is used to calculate the conditional generation probability of the text based on the evaluation protocol.

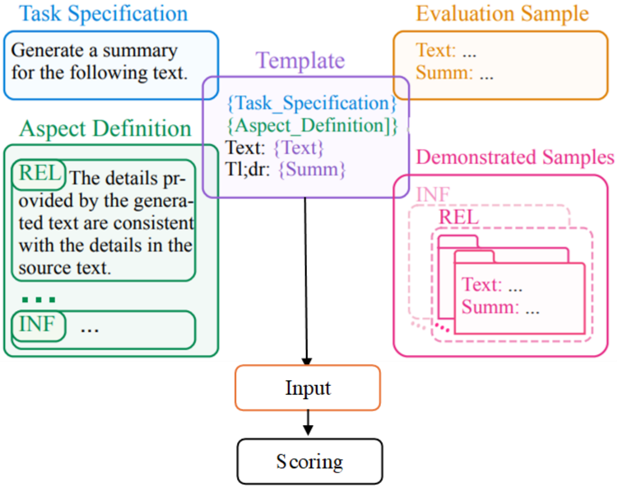
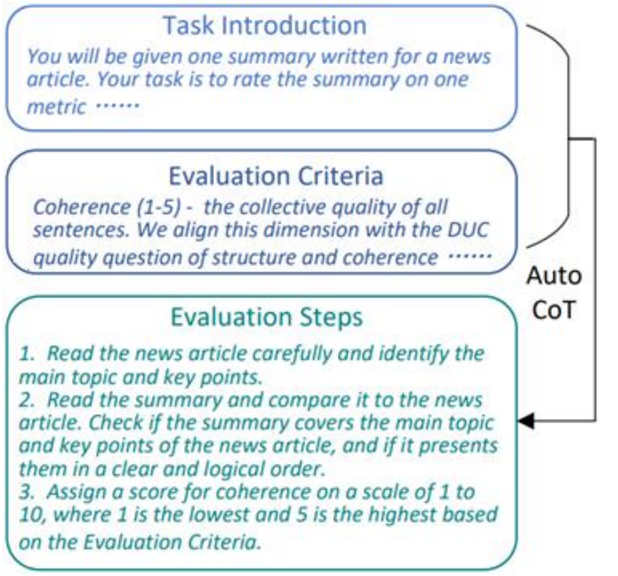


Figure 3: GPTScore Framework [10]

G-Eval [11] is a prompt-based framework for evaluating the quality of texts generated by natural language generation (NLG) systems in a form-filling paradigm. G-Eval inputs the definition of the evaluation task and the desired evaluation criteria as a prompt and asks LLM model to generate a CoT of detailed Evaluation Steps. This Chain of Thought prompting (CoT) is a technique used to improve the reasoning capabilities of a LLM model by generating intermediate steps or explanations that lead to a final answer. Instead of directly outputting the answer, CoT allows the model to break down the problem into smaller, logical steps, making it easier to handle complex questions and enhancing interpretability (c.f. Figure 4). Then the prompt is used along with the generated CoT to evaluate the NLG outputs. a scoring function that calls LLM and calculates the score based on the probabilities of the return tokens.


Task Introduction

You will be given one summary written for a news article. Your task is to rate the summary on one metric ······

Evaluation Criteria

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence ······

Evaluation Steps

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 10, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Figure 4: Task Introduction and Evaluation Criteria are input to LLM and it is asked to generate a CoT of detailed Evaluation Steps in G-Eval [11]

Friel and Sanyal highlight the limitations of the existing hallucination detection models. Authors argue that existing models offer a generic evaluation of LLM attributes and performance but do not perform a focused evaluation of the quality of the output. Most of the models cater to specific generative tasks only. Also, they are not focused on the power of context. Measuring variability in LLM performance across RAG vs non-RAG tasks is also a gap.

Further, Friel and Sanyal introduced hallucination detection methodology called ChainPoll [12]. The website rungalileo.io offers the Hallucination Index, a framework designed to evaluate and rank LLMs based on their tendency to "hallucinate," meaning producing incorrect or fabricated information. This index is aimed at helping organizations and developers select the most reliable artificial intelligence models for tasks like answering questions and generating long-form text. The Hallucination Index assesses LLM performance across three task types: question & answer without retrieval-augmented generation (RAG), question & answer with RAG, and long-form text generation (c.f. Table 1). The index uses two key metrics, correctness to evaluate if the responses are factually accurate, and context adherence to measure how well the model sticks to provided context in its responses.

Table 1: Types of LLMs

	Definition	*Use Cases*
Question & Answer without RAG	LLM that generates answers to user queries based solely on the data it was pre-trained on, without retrieving external or real-time information from a knowledge base or search engine.	• Answering general knowledge questions based on well-known facts. • Conversational agents where specific real-time or domain data is not critical.
Question & Answer with RAG	LLM that enhances the traditional Q&A process by combining two components: retrieval and generation. This system retrieves relevant information from external sources (like documents, databases, or APIs) and then generates an answer using the retrieved data to produce more accurate and contextually relevant responses.	• Retrieving information from product manuals or knowledge bases for customer support. • Finding and summarizing relevant academic papers or articles for research assistance. • Providing the most recent and accurate responses in Real-Time Q&A for news, stock market updates, or live event data. • Handling niche topics like medical information or legal documentation by pulling in authoritative references.
Long-form Text Generation	LLM that generates extended pieces of text, typically several paragraphs to multiple pages, in response to a given input or prompt. This type of text generation requires coherence, consistency, and logical structure over a lengthy piece, making it more complex than short-form generation.	• Generating well-structured essays or articles on a range of topics, incorporating research-like detail and organization. • Development of narratives, including character development, world-building, and plot progression for creative writing. • Generating long reports, summaries, or detailed explanations on subjects in business or academic settings. • Writing detailed technical manuals or documentation.

6. Mitigating AI hallucinations

Mitigating hallucinations means taking steps to reduce or prevent artificial intelligence systems from generating incorrect, nonsensical, or fabricated information. It is a process of improving the reliability of artificial intelligence outputs by addressing the underlying causes of these errors. Mitigating hallucinations in artificial intelligence, especially in LLMs, is a critical research area. This section discusses some key mitigation models.

RHO [13] is a conversational artificial intelligence model designed to reduce hallucinations by leveraging both local and global knowledge grounding techniques, supplemented by a response re-ranking mechanism to ensure relevance in dialogue responses. The task it addresses is knowledge-grounded dialogue (KGD), where generated responses rely on both dialogue history and external knowledge graphs (KGs). Using OpenDialKG, a dataset of 13,802 dialogues with over 91,000 turns and 1.19 million knowledge triples, RHO integrates local grounding (mapping dialogue tokens to specific KG entities or relations) and global grounding (associating dialogue tokens with entire sub-graphs). This dual grounding allows for richer, context-aware embeddings.

The model generates multiple candidate responses using an encoder-decoder framework and evaluates them through a re-ranking process based on the probability of alignment with KG-derived actions. Evaluation benchmarks, including BLEU4, F1, and FeQA, highlight RHO's superiority over baseline models like BART and GPT2+NPH. For instance, the full implementation of RHO achieves an F1 score of 72.29 and entity coverage of 98.53%. These results underscore the effectiveness of combining local and global grounding with advanced re-ranking mechanisms in generating accurate, knowledge-grounded dialogue responses.

The Neural Path Hunter (NPH) [14] model addresses hallucination issues in knowledge-grounded dialogue systems by employing a generate-then-refine strategy. After an initial response is generated by a language model (e.g., GPT2), NPH uses a token-level fact critic to detect potentially hallucinated entities. This critic, built using RoBERTa-Large, classifies tokens based on their likelihood of being inaccurate, with training data including intentionally introduced errors for robustness.

For flagged entities, NPH refines responses by querying an external knowledge graph (KG). It utilizes a masked language model to create contextual representations of flagged entities and formulates queries for KG navigation. Using techniques like KG-Entity Memory (via GPT2 embeddings or CompGCN) and scoring with DistMult, NPH identifies the most factual entity. The original response is then updated with this verified information, ensuring factual accuracy.

NPH demonstrates improved performance on the OpenDialKG dataset. FeQA scores increased significantly across multiple models (e.g., GPT2-KG improved from 26.54 to 28.98), while hallucination rates dropped (e.g., GPT2-KG from 19.04% to 11.72%). This approach showcases NPH's effectiveness in reducing hallucinations by integrating knowledge graphs into its pipeline for fact verification and response refinement.

The Retrieval-Augmented Generation (RAG) [15] approach minimizes hallucination in knowledge-grounded dialogue systems by incorporating external documents into the encoder-decoder framework. This method retrieves, ranks, and integrates relevant information from a document corpus using techniques like Dense Passage Retrieval (DPR), Poly-encoders, and Fusion-in-Decoder (FiD) to generate factually accurate responses.

RAG operates in three stages: retrieval, ranking, and response generation. A retriever identifies relevant documents for a given user query using learned matching functions, such as DPR, which encodes queries and documents into dense vector spaces to compute relevance. The ranker prioritizes these documents based on relevance, ensuring the most pertinent information is considered. The encoder-decoder integrates the top-ranked documents with the user’s input to produce contextually grounded responses.

DPR employs dual-encoder architectures for context-document scoring, while Poly-encoders create multiple semantic representations (codes) to enhance context interpretation. FiD synthesizes information from retrieved documents by encoding dialogue context and documents separately, merging them before final response generation.

Applied to datasets like Wizard of Wikipedia, RAG demonstrates its ability to adapt retrieval techniques to dialogue contexts, ensuring responses are grounded in verified external knowledge, reducing hallucination while maintaining relevance and coherence.

Rashkin et al. [16] introduced control codes to mitigate hallucination in knowledge-grounded dialogue systems by guiding the language model to produce responses that emphasize lexical precision, objective voice, and entailment. Control codes are special tokens added to input during training to align generated responses closely with provided evidence. During decoding, a resampling method ensures the output satisfies predefined evaluation measures.

Two approaches are proposed: (1) integrating control code features as tokens in training to emphasize reliance on evidence and avoid speculative responses, and (2) employing resampling during decoding to iteratively refine responses until evaluation criteria are met. Metrics like lexical precision (word overlap with evidence), objective voice (absence of first-person pronouns), and entailment (semantic consistency with source evidence, measured using an NLI model) ensure faithful response generation.

Evaluated on the Wizard of Wikipedia dataset, control codes significantly improved performance. For instance, GPT-2 enhanced with control codes achieved BLEU-4 scores of 7.8 and 7.6 (from baselines of 6.2 and 5.7 for seen and unseen topics, respectively). Combined with resampling, GPT-2 reached near-perfect NIP scores of 99.9 and 99.8, demonstrating the method’s effectiveness in producing factually accurate and evidence-grounded responses.

The Mixed Contrastive Learning (MixCL) [17] method reduces hallucination in conversational artificial intelligence systems by leveraging contrastive learning combined with data mixing at a fine-grained span level. This innovative approach enables models to distinguish between factual and hallucinated information without requiring extensive retraining. It addresses the limitations of traditional training methods, such as Maximum Likelihood Estimation (MLE), which often result in models replicating training data inaccurately when handling real-world knowledge.

MixCL operates in two main steps: negative sampling and mixed contrastive learning. Negative sampling generates confusing "negative knowledge" that the model is prone to misinterpret, using either retrieval-based techniques or model-generated hallucinations. Mixed contrastive learning then integrates spans of both positive (correct) and negative (hallucinated) knowledge, training the model to discern factual content through a specialized mixed contrastive loss function. This process is further supported by fine-grained techniques like Named Entity Recognition (NER) and constituency parsing, ensuring precision in identifying intrinsic and extrinsic hallucinations.

The training process is optimized using a combination of three loss functions: Maximum Likelihood Estimation (to replicate training data patterns), Mixed Contrastive Loss (to improve factual discernment), and Language Modeling Loss (to mitigate knowledge forgetting). When evaluated on the Wizard-of-Wikipedia dataset, MixCL demonstrated superior performance compared to other models. Under realistic conditions, it achieved an F1 score of 21.6, surpassing KBbased methods like KnowledGPT (F1 score: 21.1), along with notable improvements in ROUGE-L (20.5) and BLEU-2 (9.2). These results underscore MixCL's ability to enhance accuracy and reliability in generating knowledge-grounded conversational responses.

The HERMAN model [18] addresses the challenge of correcting hallucinated quantitative entities in abstractive summarization systems, particularly for quantities like dates, numbers, and monetary values. Hallucinated entities that are inconsistent with the source text undermine the accuracy of summaries. To tackle this, HERMAN employs an encoder-decoder architecture leveraging Bidirectional LSTMs and attention mechanisms to verify the factual consistency of quantitative entities and rerank summaries based on alignment with the source text's quantities.

HERMAN's training dataset, derived from the XSum dataset, focuses on entries with quantitative entities. It includes both original and synthetic data, where quantitative entities in summaries were replaced with random alternatives from the source texts, followed by manual annotation into VERIFIED or UNVERIFIED categories. The model employs token-level labels (e.g., VERIFIED, UNVERIFIED, or unrelated tokens using BIO tagging) and sentence-level labels to evaluate summaries.

The architecture comprises a Bidirectional LSTM encoder that contextualizes source tokens and a decoder with an attention mechanism for processing summaries. The model outputs token-level classifications and a binary document-level classification to assess overall summary consistency. For reranking, HERMAN uses two strategies: HERMAN-GLOBAL, which focuses on document-level labels, and HERMAN-LOCAL, which evaluates token-level scores.

When integrated with summarization models like BART, BERTSUM, and TCONVS2S, HERMAN improved performance on the XSum test set, achieving higher ROUGE-1, ROUGE-2, and ROUGE-L precision and F1 scores. These results highlight its effectiveness in reducing hallucinated quantitative entities while maintaining the informativeness of summaries.

The self-contradictory method [9] is a novel approach to detect and mitigate hallucinatory content in LLMs by leveraging deliberate self-contradictions in their outputs. This method identifies logical inconsistencies within generated responses using paired statements prompted from a language model (gLM), which are then analyzed by a secondary model (aLM). The analysis detects contradictions indicative of potential hallucinations. These inconsistencies are mitigated through iterative text editing, ensuring outputs remain fluent and informative.

The approach involves prompting the gLM with inputs designed to induce contradictory responses, followed by detecting and flagging these inconsistencies using an analyzer model trained for this purpose. Experimentation with models such as ChatGPT (3.5), GPT-4, Llama2-70B-Chat, and Vicuna-13B shows the method's effectiveness. ChatGPT and GPT-4 demonstrated strong performance in both detecting and mitigating contradictions, achieving up to 89.5% mitigation of self-contradictions while maintaining output quality. Open-source models like Llama2-70B-Chat and Vicuna-13B performed less effectively in these tasks. The method offers a promising strategy for reducing hallucinations in black-box LLMs without external knowledge dependency.

Thus, strategies for mitigating hallucinations in artificial intelligence systems may be categorised as follows.

1) **Data-Centric Approaches**: Such approaches focus on improving data quality and diversity. Training LLMs on high-quality, diverse, and representative datasets helps the model learn more comprehensive patterns and reduces the impact of biases present in the data. Training data may be augmented with synthetic examples or paraphrases to improve the model's robustness and generalization ability.
2) **Model-Centric Approaches**: Enhanced Model Architectures: Developing more robust and interpretable artificial intelligence models can help identify and correct errors in the reasoning process. Also, fine-tuning LLMs on datasets specifically designed for fact verification can improve their ability to distinguish between factual and non-factual statements.
3) **Training and Decoding Strategies**: Training LLMs with human feedback can help align their outputs with human values and reduce the generation of harmful or misleading content. Also, implementing constraints during the decoding process can prevent the model from generating outputs that violate certain rules or constraints.
4) **External Knowledge Integration**: Providing LLMs with access to external knowledge sources, such as knowledge graphs or databases, can help ground their responses in factual information (Retrieval-Augmented Generation). Directly editing the internal knowledge of LLMs can help correct factual errors and improve their accuracy.
5) **Detection and Verification Techniques**: such techniques are based on generating multiple responses and analysing their consistency to detect hallucinations. Fact verification models use dedicated models to verify the factual accuracy of LLM-generated text.
6) **Prompt Engineering**: Carefully crafting prompts can influence the LLM's output and potentially reduce hallucinations [19]. This includes providing clear instructions, specifying constraints, and asking the LLM to provide sources or justifications for its claims.

Mitigating hallucinations is an ongoing research challenge, and a combination of these strategies is often necessary to achieve the best results.

7. Conclusion

The phenomenon of hallucination in artificial intelligence, whether regarded as a metaphor or a misnomer, underscores the critical challenges in ensuring reliability and trustworthiness. Artificial intelligence hallucinations, defined as instances where models generate outputs that are inaccurate, nonsensical, or fabricated, highlight inherent limitations in current artificial intelligence systems. These hallucinations can manifest in various forms – factual inaccuracies, logical inconsistencies, or completely fabricated data – and are often rooted in issues like dataset biases, model overconfidence, or misalignment of objectives.

Detecting these hallucinations is an evolving field, requiring sophisticated evaluation techniques ranging from automated consistency checks to human-in-the-loop reviews. However, detection alone is insufficient. Effective mitigation strategies, such as robust dataset curation, improved model architecture, and reinforcement of context-aware mechanisms, are crucial to addressing the issue.

As artificial intelligence continues to be integrated into critical decision-making processes, tackling hallucinations becomes an essential task to ensure the ethical, reliable, and practical deployment of artificial intelligence technologies. The road ahead calls for interdisciplinary collaboration and innovation to refine artificial intelligence systems, aligning their outputs with both human expectations and real-world needs.

References

[1] S. K. Bordoloi, "The hilarious & horrifying hallucinations of AI," 2 July 2023. [Online]. Available: https://www.sify.com/ai-analytics/the-hilarious-and-horrifying-hallucinations-of-ai/. [Accessed 23 September 2024].

[2] J. Brodkin, "Lawyer cited 6 fake cases made up by ChatGPT; judge calls it "unprecedented"," 31 May 2023. [Online]. Available: https://arstechnica.com/tech-policy/2023/05/lawyer-cited-6-fake-cases-made-up-by-chatgpt-judge-calls-it-unprecedented/. [Accessed 23 September 2024].

[3] P. Prakash, "ChatGPT falsely accused a mayor of bribery when he was actually the whistleblower," 6 April 2023. [Online]. Available: https://fortune.com/2023/04/05/chatgpt-falsely-accused-australian-mayor-bribery-openai-defamation/. [Accessed 9 September 2024].

[4] S. Cole, "'Life or Death:' AI-Generated Mushroom Foraging Books Are All Over Amazon," 29 August 2023. [Online]. Available: https://www.404media.co/ai-generated-mushroom-foraging-books-amazon/. [Accessed 9 September 2024].

[5] J. Vincent, "Google's AI chatbot Bard makes factual error in first demo," 8 February 2023. [Online]. Available: https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo. [Accessed 23 September 2024].

[6] OpenAI, "GPT-4," OpenAI, 2023.

[7] F. Wang, "GitHub - ZurichRain/AGI-Hallucination: A Survey of MultiModel LLM Hallucination," January 2024. [Online]. Available: https://github.com/ZurichRain/AGI-Hallucination/blob/main/LightHouse__AGI_Hallucination__submit_.pdf. [Accessed 10 September 2024].

[8] P. Manakul, A. Liusie and M. J. F. Gales, *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,* Cornell University, 2023.

[9] N. Mündler, J. He, S. Jenko and M. Vechev, *Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation,* Cornell University, 2024.

[10] J. Fu, S.-K. Ng, Z. Jiang and P. Liu, *GPTScore: Evaluate as You Desire,* Cornell University, 2023.

[11] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu and C. Zhu, *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment,* Cornell University, 2023.

[12] R. Friel and A. Sanyal, *Chainpoll: A high efficacy method for LLM hallucination detection,* Cornell University, 2023.

[13] Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng and P. Fung, *Rho: Reducing hallucination in open-domain dialogues with knowledge grounding,* 2023, pp. 4504-4522.

[14] N. Dziri, A. Madotto, O. R. Zaiane and A. J. Bose, "Neural path hunter: Reducing hallucination in dialogue systems via path grounding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[15] K. Shuster, S. Poff, M. Chen, D. Kiela and J. Weston, *Retrieval augmentation reduces hallucination in conversation,* 2021, pp. 3784-3803.

[16] H. Rashkin, D. Reitter, G. S. Tomar and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[17] W. Sun, Z. Shi, S. Gao and P. Ren, "Contrastive learning reduces hallucination in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[18] Y. Zha, Y. Yang, R. Li and Z. Hu, "AlignScore: Evaluating factual consistency with a unified alignment function," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023.

[19] L. Reynolds and K. McDonell, *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,* Cornell University, 2021.

Author Profile

Dr. Brahmaleen Kaur Sidhu earned her Ph.D. degree in Faculty of Engineering and Technology from Punjabi University, Punjab, India, M.Tech. degree in Computer Science and Engineering from the Punjab Technical University, Punjab, India, and B.Tech. degree in Computer Science and Engineering from Punjabi University. She is currently working as Assistant Professor in the Department of Computer Science and Engineering, Punjabi University and has around 18 years of teaching experience. Her research interests include software architecture, software evolution, software quality, refactoring, model-driven development, data science and machine learning. She has around 80 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE, and a book titled "A Handbook of Reinforcement Learning" published in 2023. She has been awarded the "International Innovative Educator Award 2021" and is listed in "100 Eminent Academicians of 2021" by International Institute of Organized Research.