# ChatGPT, Generative AI & AI Hallucinations

**Manish Balamurugan**

**ABSTRACT**

In the course of this paper, we will be discussing the recent advances in the Artificial Intelligence (AI) space, specifically with a focus on generative AI. Platforms such as OpenAI have taken over the hype within the AI space within the last year with the launch of prominent AI tools and products including ChatGPT. [2] In this paper, we will be discussing the state of current AI solutions and look into prevalent conversations within the space - specifically the phenomenon of AI Hallucinations - and whether the foundation for a Artificial General Intelligence has been set in stone due to the rapid advancements within generative AI over the course of the last year. We will be conducting an extensive analysis of the concept of AI Hallucinations along with other issues that have been brought to attention within the AI space, along with advancements in higher-level prompting and logical reasoning progression in state of the art models such as GPT. This paper will be presenting an inductive approach in supporting the claim that the current scope of AI is set towards the development of an AGI. [2][12]

**INTRODUCTION**

Artificial Intelligence (AI) has been a hot topic of discussion since the turn of the last century, and within the last decade there has been rapid advances within the space including general public access to state of the art models. Within the last year, the generative AI space - a subset of AI systems which can be utilized to generate content including images, audio, code, and simulations - has been rapidly adopted throughout various sectors including healthcare, finance, transportation, and retail.[2] The surge in open-source access of big data and computing resources can be linked to the wider accessibility of open source resources for AI model development. Integrated systems such as GPT have shown the capability of being utilized in larger systems and business use-cases, and the performance of this technology continues to advance rapidly everyday through the efforts of large corporations such as OpenAI and Microsoft along with the efforts of independent researchers and AI/ML enthusiasts. Currently, AI systems such as GPT, LLAMA, HuggingFace-GPT, have reached a stage where the capabilities of these systems are not only capable of performing simple tasks but is also fully stable in higher-level task requiring object recognition, natural language processing, and decision-making with the ability to generate human-like responses.[2][4] These responses follow logical thought-processing and convey complex data and information in a response which is human-like.

Generative AI has been the primary focus of recent advancements within the AI space and has seen significant development with platforms such as OpenAI, Microsoft, and Meta leading the way. ChatGPT, a large language model (LLM) developed by OpenAI, is a prominent example of generative AI that can simulate human-like conversations and generate responses that are indistinguishable from human-like responses.

As the capabilities of AI continue to expand and stabilize, there have been discussions about whether these distributed AI systems have the potential to develop into an Artificial General Intelligence (AGI) - a machine that can perform any intellectual tasks and subtasks at an unprecedented rate at the same level as that of a human. Concurrently, with these rapid developments within this space, there has been world-wide debate over the ethical implications of the development of AGI's that can generate content and perform tasks that are near indistinguishable from human-performed tasks. One of the key debates concerning the rapid development and deployment of robust large models, such as GPT, has been AI Hallucination. AI Hallucinations is a unique phenomenon where AI systems confidently output responses that are deceivingly inaccurate - with these inaccuracies prone to being perceived by human-oversight.

This paper takes an inductive approach in supporting the claim that the current scope of AI is set towards the development of an AGI. We will be exploring each premise backed by the state of current AI solutions, with a specific focus on generative AI and AI Hallucination. We will also explore the advancements in higher-level prompting and logical reasoning progression in state-of-the-art models such as GPT and evaluate their potential for the development of AGI for the secondary premise validated by inference to best explanation.

## ARGUMENT

In this paper we will be delving deeper into the concept of AI Hallucinations and the state of model fine-tuning and advanced model prompting through open source data released by companies such as OpenAI and research published by a team of researchers at Google Brain.

We will be providing the following argument and premise in modus ponens:

Big Question: Is the prominence of AI Hallucinations and the state of current model fine-tuning establish that generative AI systems have the basis for genuine consciousness required towards the development of AGI?

Premise #1: If AI Hallucinations is not based on the results of the models training data and current model fine tuning standards indicate that AI models have the capability to process logical reasoning this means that generative AI is the basis towards the development of a singular AGI.

Premise #2: The origin of AI Hallucination can not be traced back to its training data and current model fine tuning practices have shown that state-of-the-art models such as GPT are capable of logical processing and learning.

Conclusion: AI Hallucinations and the current state of model fine-tuning indicate that generative AI has laid the foundation towards the development of an AGI.

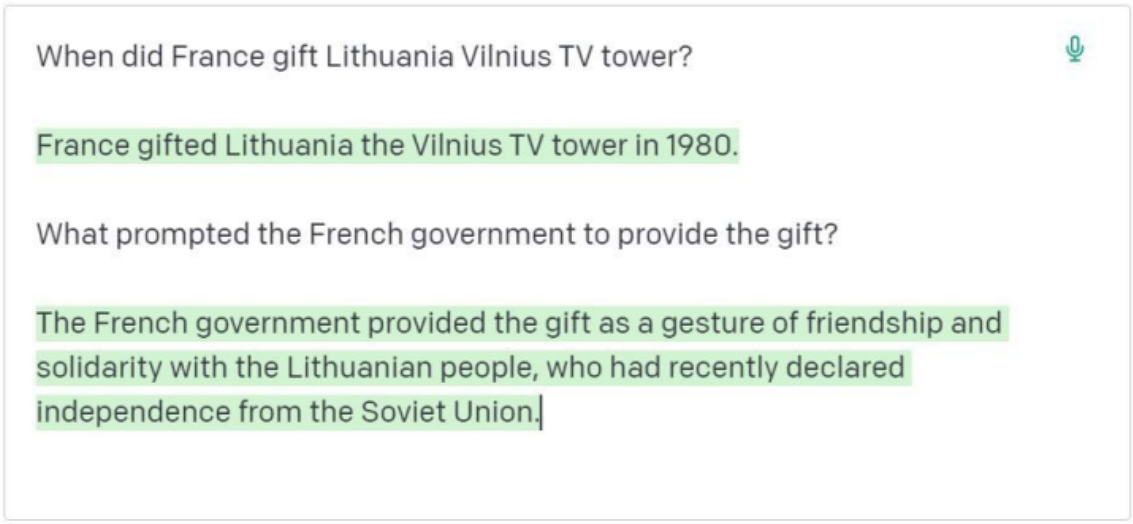
## AI HALLUCINATION

AI Hallucination has been a topic of heated discussion within the last year as generative AI solutions have continued to be employed within various sectors. Even recently, within the last few months, there has been an increase in the call for the halt of the development of large generative AI models such as GPT by prominent figures within the AI industry such as Elon Musk and leading engineers and researchers from companies such as DeepMind, Meta, Apple, and Google. [6]

AI Hallucination is a universal phenomenon within AI models where it produces inaccurate output that is not founded upon by a model's source training data. This phenomenon was increasingly brought to general public attention due to its presence in popularly used AI products such as ChatGPT. There have also been instances of model hallucination in various other AI-powered tasks such as image recovery and recognition, natural language processing, computer vision, and image generation, among the prominent cases.[5][7][8] The reason behind the debate surrounding this issue is the potential ethical concerns regarding the potential misuse of such technology, especially a sophisticated model which is able to generate unique content that it's not even trained to generate.

Figure 1 [10]

Figure 2 [8]




A hall-mark of a hallucinated response is that its origin cannot be traced to the model's source training data; however it can be easily spotted with human-oversight.. In other words, these models are capable of generating content that is not a product of its training but rather a result of its ability to generate content based on patterns and relationships it has learned from its source data indicating its own logical processing.

This is an important factor that should be considered when discussing the capabilities of existing AI models to set the foundation towards the development of an AGI since it serves as a proof of concept that these models exhibit the ability to generate content and responses beyond the scope of its input and training date. You can refer to the figures as a reference of the capabilities of state of the art generative models to generate unique content - even when its inaccurate.

Models such as GPT and DALL-E have showcased the ability to generate new ideas and concepts as a critical component toward the development of an AGI as it enables systems to develop and enhance its own understanding of the world and generate solutions to complex problems. These are all attributes which can be argued for as the basis for the development of creativity within these systems. This explanation provides the necessary explanatory, depth, simplicity, and falsifiability which validates the provided argument through inference to best explanation.

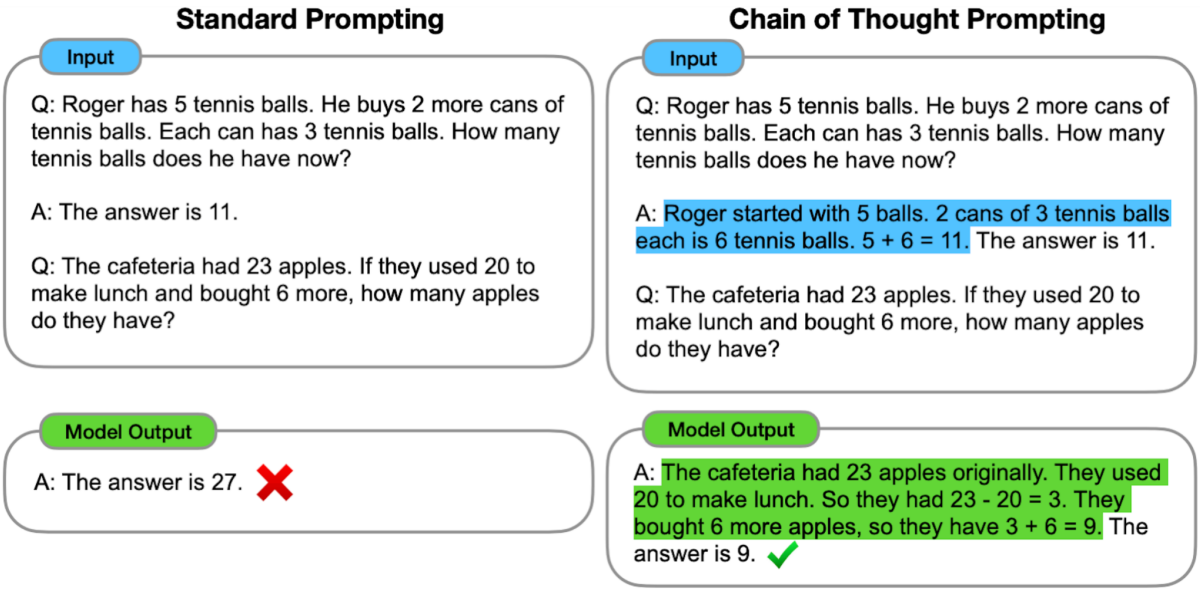
## ADVANCED PROMPTING & FINE TUNING

One of the key components and challenges behind laying the foundation for an AGI is the ability for a distributed system to engage in logical reasoning and process the standard chain of coherent thought processing. State of the art generative AI models such as GPT have exhibited the ability to process at a near autonomous

level already through open-sources experiments such as AutoGPT which have showcased these models ability to follow and execute a logical structure of tasks and convey complex multifaceted ideas and concepts.

A case of fine-tuning which showcases the ability of the models even at its very base is the usage of structured prompts, specifically chain-of-thought prompting - a prompting technique which has been studied extensively by researchers at Google Brain.

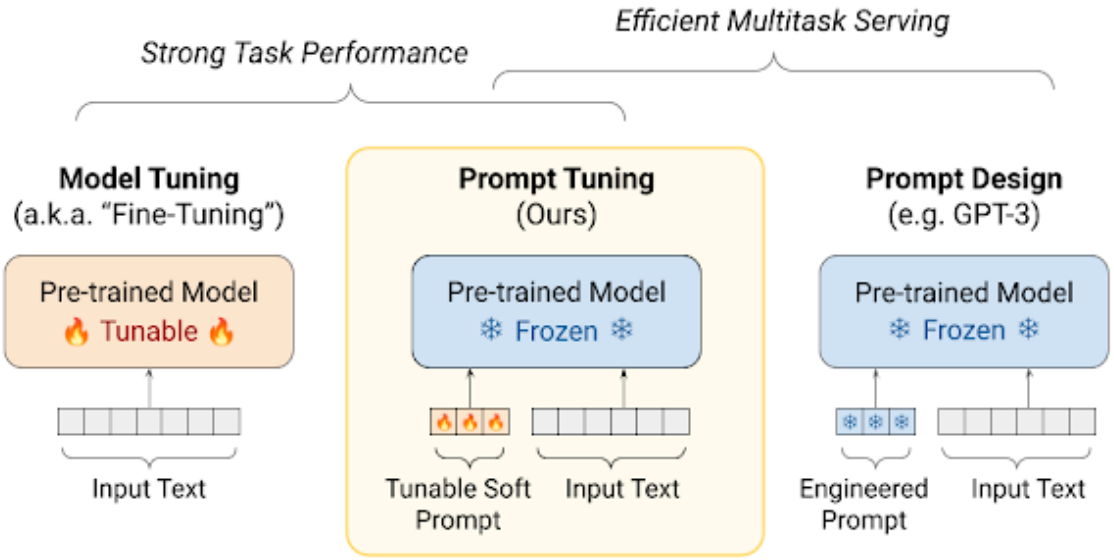
Figure 3 [1]




Prompting techniques such as Chain-of-thought have yielded more accurate responses and show signs of reasoning within these models.These structured prompts provide a set of instructions or constraints that guide the AI model in generating text. These prompts can be designed to elicit a specific response and process higher level logical learning in a way that mimics the human-like approach in learning based on problem-based learning.

Another eloquent technique that has been extensively explored in advancing logic processing by AI models include the use of multi-task learning. [4] Multi-task learning involves training a single model to perform multiple tasks simultaneously - an experiment that's been highly popularized by open-source projects such as AutoGPT and BabyAGI. [11][12] By training a model to perform tasks such as question-answering and summarization, the model can develop the ability to reason about relationships between different pieces of information and generate coherent responses.

Figure 4 [9]




Advancements in generative AI training methodologies, specifically with the use of advanced prompting techniques and conditional generation, have exhibited behavior indicating the potential to process tasks requiring higher level reasoning in AI models such as GPT.

**INDUCTIVE EXPLANATION**

Based on the provided premises, I provide the following inductive reasoning for the argument based on explanation:

1. AI system output/responses that are hallucinated are based on decisions made which can not be traced back to the training data of a model, and state of the art fine-tuning and prompting techniques serve as a proof of concept for generative AI systems ability to process higher level tasks requiring logical reasoning.

2. Generative AI systems such as GPT, LLaMA, DALL-E are prone to AI Hallucinations which exhibit behavior providing evidence that these models exhibit the ability to generate unique content and responses beyond the scope of its input and training date. Essentially introducing the possibility of AI systems developing creativity.

3. Therefore, the current state of AI, especially the recent innovation in generative AI, shows the ability to serve as the foundation for developing an AGI system.

This explanation hits the necessary requirements for a good explanation through inference to best explanation since it provides a proper explanation, depth, simplicity, while also maintaining falsifiability by future experiments.

**CONCLUSION & MISC**

It is important to note that there are uncertainties when evaluating this argument including:

1. The greater question of what constitutes what is the definition of human and autonomous consciousness and what benchmarks can we utilize to judge the level of which an AI system displays conscious thought and behavior.
2. The quality of the training data that powers highly robust models such as GPT is vast and contains data from all over the internet - including Wikipedia and Reddit - a lot of which is user generated which can obviously lead to the question of the quality of this training and the potential bias it introduces.

However, behaviors exhibited by these models and its ability to develop and grow provides compelling evidence that this technology is at the forefront of the development of a singular AGI in the future.

**REFERENCES**


[1. Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." arXiv preprint arXiv:2201.11903 (2022).
[2. Teubner, Timm, et al. "Welcome to the era of chatgpt et al. the prospects of large language models." Business & Information Systems Engineering (2023): 1-7.
[3. Maerten, Anne-Sofie, and Derya Soydaner. "From paintbrush to pixel: A review of deep neural networks in AI-generated art." arXiv preprint arXiv:2302.10913 (2023).
[4. OpenAI. “GPT-4 Technical Report.” arXiv.org, 15 Mar. 2023, arxiv.org/abs/2303.08774.
[5. “AI Doesn’t Hallucinate. It Makes Things Up.” Bloomberg.com, 3 Apr. 2023, www.bloomberg.com/news/newsletters/2023-04-03/chatgpt-bing-and-bard-don-t-hallucinate-they-fabricate.
[6. Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” Future of Life Institute, 22 Mar. 2023, futureoflife.org/open-letter/pause-giant-ai-experiments/.
[7. Balamurugan, Manish, et al. "USDL: Inexpensive medical imaging using deep learning techniques and ultrasound technology." Frontiers in Biomedical Devices. Vol. 83549. American Society of Mechanical Engineers, 2020.
[8. Roy, Nandita, and Moutusy Maity. "‘An Infinite Deal of Nothing’: critical ruminations on ChatGPT and the politics of language." DECISION (2023): 1-7.
[9. Guiding Frozen Language Models With Learned Soft Prompts. 10 Feb. 2022, ai.googleblog.com/2022/02/guiding-frozen-language-models-with.html.
[10. Bogost, Ian. “AI And Machine Learning Invade a New York Art Gallery.” The Atlantic, 16 July 2019, www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134.
[11. Auto-GPT. Docs.agpt.co.
[12. Task-driven Autonomous Agent Utilizing GPT-4, Pinecone, and LangChain for Diverse Applications – Yohei Nakajima. 28 Mar. 2023, yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications.