STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 12, Issue 1, 2025
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

Unmasking Hallucinations in Large Language Models Using Analysis of the LLAMA 2 Model and RAG Intervention

Atharva Patil*, Arohi Paigavan, Amarti Dhamele,
Abbas Merchant, Aditya Kasar

Abstract

The study describes the creation of a chatbot for financial trading called "TradeBot" and how it uses Retrieval Augmented Generation (RAG) to overcome the problem of producing false or unverifiable information, sometimes known as hallucinations. RAG allows the chatbot to refer to an external data source in addition to its taught knowledge, which increases the accuracy of its responses. The NCFM (NSE's Certification in Financial Markets) book was integrated as an external data source by the authors of this study, who employed the Llama 2 Model for the chatbot and RAG implementation. We examine the ways in which RAG improves the accuracy of LLAMA 2 in financial trading scenarios by including the NCFM (NSE's Certification in Financial Markets) book as an external data source. The results of our investigation demonstrate that the addition of RAG significantly lowers the frequency of hallucinations and enhances reaction reliability when LLAMA 2 is used with and without RAG. The study compared the chatbot's responses to when RAG was used and when it was not, to show how RAG helps avoid hallucinations and guarantee that the chatbot delivers more accurate and trustworthy information.

Keywords: LLMs, hallucinations, RAG, llama 2, generative AI, hallucinations in LLMs, retrieval augmented generation

INTRODUCTION

Deep learning models that have been pre-trained on substantial data sets are known as large language models (LLM). The transformer consists of neural networks with self-attention capabilities, including encoders and decoders. The encoder and decoder process text to extract meaning and connect words and phrases. Large language models are highly adaptable. A model is capable of handling multiple functions, including question answering, document summarization, language translation, and sentence completion. Even though they are not flawless, LLMs show a remarkable capacity for making predictions from a limited set of cues or inputs. Using human language input, generative AI (artificial intelligence) may generate material using LLMs.

LLMs have showed promise for generative and knowledge-intensive activities that necessitate

***Author for Correspondence**
Atharva Patil
E-mail: atharvamp@gmail.com

Students, Department of Electronics and Telecommunication Engineering, Shri Vile Parle Kelavani Mandal's Narsee Monjee Institute of Management Studies, School of Technology, Management & Engineering, Navi Mumbai, Maharashtra, India

Received Date: July 11, 2024
Accepted Date: November 05, 2024
Published Date: November 08, 2024

Citation: Atharva Patil, Arohi Paigavan, Amarti Dhamele, Abbas Merchant, Aditya Kasar. Unmasking Hallucinations in Large Language Models Using Analysis of the LLAMA 2 Model and RAG Intervention. Journal of Artificial Intelligence Research & Advances. 2025; 12(1): 76–86p.

global or domain knowledge. The way in which LLMs represent words is critical to their operation. Each word was represented by a numerical table in earlier machine learning models.

However, this representation was unable to identify word links, such as those with similar meanings. This limitation was lifted by expressing words using multidimensional vectors, also known as word embeddings, in a way that places words with relevant contextual meanings or other associations close together in vector space. Transformers can use word embeddings to pre-process text into numerical representations through the encoder. Although large language models (LLMs) are effective artificial intelligence tools across a range of areas [1–3], these are prone to producing erroneous or faked outputs crafted to be believable, or hallucinations [5]. Serious repercussions might arise from hallucinations [5, 1], particularly in fields where it is relied upon heavily for referencing and learning. This study investigates the origins and consequences of hallucinations and suggests a potential remedy based on Retrieval Augmented Generation (RAG). RAG is a method that directs the creation of LLMs by utilizing outside information sources [6]. The concept of RAG is further explained in the paper in the beginning of the proposed methodology section (section 3). An open-source large language model (LLM) developed by Meta AI is called Llama 2 [7]. It is a family of transformer-based autoregressive causal language models. It is trained with a reinforcement learning approach to produce non-toxic and family-friendly output. A set of pretrained and refined generative text models with scales ranging from 7 billion to 70 billion parameters is called Llama 2. The Llama-2-Chat LLMs have been fine-tuned and are optimized for dialogue use cases. On the majority of benchmarks, Llama-2-Chat models perform better than open-source chat models [7], and they are comparable to several well-known closed-source models like ChatGPT and PaLM [8] when evaluated by humans for helpfulness and safety.

In the world of conversational AI, ChatGPT is a leading frontier. From its mere contextual capabilities provided by GPT 3.5, to now also accepting image as in input, introduced in the fourth model (GPT 4) in its series of the GPT foundational models. It too is prone to AI hallucinations. This study further discusses about the susceptibility of ChatGPT to hallucination, based on the instances of GPT 3.5 hallucinations. Using RAG, the authors of this study have utilized a custom embedding model for contextualization. Using the Llama 2 Model, RAG, and an external data source consisting of a single pdf document encompassing of content for learning how to trade using technical analysis, the study builds and tests a chatbot called TradeBot to illustrate the efficacy of RAG. The purpose of TradeBot is to skill individuals on leveraging technical analysis strategies to master trading. The pdf document in the external data source is the NSE's Certification in Financial Markets (NCFM) book [9] by National Stock Exchange of India. It is an excellent source for learning technical analysis and is followed by many universities nationally in their pedagogy.

LITERATURE REVIEW

Within the burgeoning field of conversational AI, the reliability of Large Language Models (LLMs) is a concern largely due to the rise of the phenomenon of hallucinations which are responses deviating from factual or contextually accurate information. This literature review surveys existing research on hallucinations in LLMs, homing in on their implications in conversational AI applications. In this study, the authors explore the nature and causes of hallucinations, generalize current mitigation strategies, and particularly investigate the potential of Retrieval Augmented Generation (RAG) in improving response accuracy.

Hallucination in Large Language Models

This is a behavior in which the model crafts responses with false knowledge as if it is accurate. Hallucinations mostly happen when mixing facts with language-based context, and they are mainly brought on by skewed training data, unclear prompts, and erroneous large language model parameters [10]. AI hallucinations might be defined differently, based on context. In general, AI hallucinations occur when LLM outputs are contextually unreasonable, contradictory with reality, or unreliable to the input. Some academics feel that the term "hallucination" is misleading and that AI hallucinations are

more accurately described as fabrications. On the surface, hallucinations result from a lack of contextual knowledge since the model must abstract the training data and the prompt, which may cause some information to be lost. Additionally, noise in the training set could produce a distorted statistical pattern that causes the model to behave differently than one would have predicted.

Hallucinations could be seen as a characteristic of large language models. If you want the models to be creative, you want to see them hallucinate. For instance, if you ask ChatGPT or other large language models to provide you with a fantasy story plot, you want it to create a fresh character, scene, and storyline rather than copying an already existing one. Temperature is a parameter which can be used to control such a characteristic. Temperature determines how "creative" the model can be with its responses i.e. selecting the less probable tokens along with the most probable ones. The value of the temperature ranges from 0 to 1. A temperature closer to 1 indicates a more hallucinated response while a temperature closer to 0 indicates zero ability of the model to generate innovative responses. Even outside the contextual scope, hallucinations can be utilized to improve the quality of the generated result of large foundational models. This has motivated recent work on vision-based machine translation which aims to improve machine translation systems by leveraging hallucinations [11]. Addressing Gen-AI hallucinations is imperative, as it not only compromises the reliability of LLMs but also poses the risk of disseminating misinformation.

ChatGPT

Despite its prowess, ChatGPT is not immune to hallucination. Instances abound where the GPT 3.5 model generates responses that, while sounding plausible, are factually inaccurate or contextually inappropriate. In a much recent study, ChatGPT was tested with the HaluEval benchmark to examine the rate of hallucinations in responses [12]. HaluEval is a comprehensive benchmark meant to assess hallucinations in LLMs. The empirical results showed that 19.5% of the responses on specific topics provided by ChatGPT were fabricated responses. Graver situations of hallucination can be found in research studies where the application of Generative AI in scientific writing is explored. One such study proposed the 'double-edged sword' nature of ChatGPT in generating pathogenesis (the cause and development of a disease) of homocystinuria-associated osteoporosis and late-onset Pompe disease (LOPD) [1]. In that paper, the authors, Alkaissi and McFarlane, first prompted ChatGPT to provide a literature review of the contents of the diseases to which ChatGPT replied accurately with the right kind of scientific explanation. The authors then prompted ChatGPT to explain the findings further with references to fact check the presumed literature review. ChatGPT then provided five references dating to the early 2000s. In reality, none of the papers exist.

Many financial firms such as JP Morgan, Bank of America, Citigroup, Deutsche Bank, Goldman Sachs and Wells Fargo had forbidden their employees from using ChatGPT or any conversational AI for that matter [3]. This rooted from the mere prevalence of the concern of the potential for generative AI to produce inaccurate or misleading information which, in the finance sector is far more damaging and edging to the risk of users' monetary information, than in other sectors. This is a result of the LLM's capability to not retrieve relevant information. The Retrieval Augmented Generation (RAG), discussed in this study, however, can help refine this querying ability of any LLM. The methodology proposed in the study does not make use of any monetary data or real-time financial data.

Mitigating Hallucinations

When misleading or inaccurate information is generated or seen, it can have major effects in mission-critical fields like medical [13], banking, and finance settings. In these fields, accuracy and dependability are critical, and any kind of distortion in data, analysis, or decision-making can have a big negative impact on results and operations. Thus, in these high-stakes areas, strong policies and procedures are necessary to reduce and even eliminate hallucinations [14]. If the hallucination was brought on by tainted training data, you can purge the data and retrain the model. On the other hand, the majority of models are too big to train on your own. On commodity technology, it might not even

be able to fine-tune an existing model. Human intervention in the outcome [10], and requesting the model to regenerate if something went horribly wrong could be two of the finest mitigating strategies.

Numerous strategies and tactics exist to reduce hallucinations from the perspective of the user (prompt engineering) and the LLM (model development). Tonmoy *et al.* [15] offers a taxonomy of strategies for mitigating hallucinations in LLMs, emphasizing popular approaches that include model building and prompting tactics. Techniques to remove hallucinations can be further simplified into three branches, including knowledge retrieval based approach (introducing external source of data), redefinitions to the vector databases (training data source), and supervised fine-tuning during model development. In this study, the authors explored the branch of knowledge retrieval and implemented Retrieval Augmented Generation (RAG) method on their Llama 2 chatbot to prevent it from creating hallucinated responses pertaining to data on learning financial trading.

The foundation for a thorough examination of the impact of generative AI hallucinations is laid by this study of the literature. By synthesizing existing knowledge on the nature, impact and prevention of these hallucinations, researchers can develop a nuanced understanding of the implications for data reliability. The understanding of the RAG approach further enhances the relevance of the research, contributing valuable insights to all users. With the domanial problems of ChatGPT put forth, one must be fully aware of the limitations of their chosen LLM [5]. ChatGPT for instance, does not possess any data after September 2021 nor does it support Realtime data. Hence, prompting to reference the latest articles or research papers will become an impossible task for ChatGPT if it will not have access to the material. So, it is less about which LLM to use and more about using an LLM after understanding all its strengths and weaknesses.

PROPOSED METHODOLOGY

Retrieval Augmented Generation (RAG)

RAG is an AI framework that takes care of these two problems. It was first proposed in 2020 by Meta (then known as Facebook) [6] to give LLMs access to data beyond their training data. There are two main problems suffered by LLMs: no access to relevant data sources, and the data source being static in nature. The traditional architecture of generative LLMs consists of the user sending prompts to the LLM and receiving a response back, having users' queries resolved. Here, the responses sent by LLMs are often not evidenced or questioned by the user on validity as LLM always claims to be confident in its reply. The model solely uses the training set of data. So, the RAG framework instructs the LLM to retrieve data from an additional set of external data, make changes to the response and then display it to the user. This data contains multiple content files and can be open source like the internet or closed source (a fixed set of documents and files). In this way, the RAG approach causes the LLM to hallucinate less without leaking confidential data and produces more reliable responses. The retriever that queries or indexes information from the external data source needs to be highly accurate in order to assist the LLM. To highlight the key differences in the responses between a non-RAG implemented LLM and a RAG implemented LLM, the authors of this study developed two variations of a chatbot called "TradeBot". TradeBot comes with open-source code available at https://github.com/Progpr/TradeBot. One variant is a chatbot, not supported by the RAG framework (Figure 1). This variant depicts the standard user-LLM architecture where the user first enters the input, sends the prompt, the prompt is received by the LLAMA 2 model, the model then sends responses, the responses are recorded in the chatbot's session and finally displayed in the chat UI. The second variant is supported by the RAG framework (Figure 2). This variant introduces a retrieval engine that the model leverages before replying to the user. The user first enters the input, sends the prompt, the prompt is received by the LLAMA 2 model, the model then uses the retrieval engine to retrieve relevant content from the pdf document, the retrieved information is converted to text, the text is then used to generate the response which is then displayed back to the user in the chat UI. The Llama 2 model was trained on an external data source that comprised of a single pdf document (the NCFM book [9]). The NCFM

book consists of information about learning technical analysis for trading. For this experiment, the authors categorized the QnA prompts in three categories of financial trading concepts: chart patterns, trading strategies, and trading psychology. One question per category was asked to the chatbot and the responses were recorded.

Components

Llama 2 7B Chat

To develop the chatbot, the authors of this study chose the GGUF (GPT-Generated Unified Format) quantized Llama 2 7B chat model by Meta. The model is available publicly at this repository, https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGUF. The model was run on Google Colab's T4 GPU resource of the free tier, and it showcased sufficient utilization of system resources (Figure 3). The authors set the maximum number of tokens to be used in a response as 256.

Embedding Model

Embedding models are used to convert text or tokens to continuous vector representations. They are used by LLMs to contextualize information. The authors have chosen the GTE (General Text Embedding) model over plain sentence transformer embedding models. According to the text embeddings leaderboard proposed by Muennighoff *et al*. [16], the GTE model performs better than most sentence transformer embedding models across various tasks such as retrieval, summarization, classification and many more.

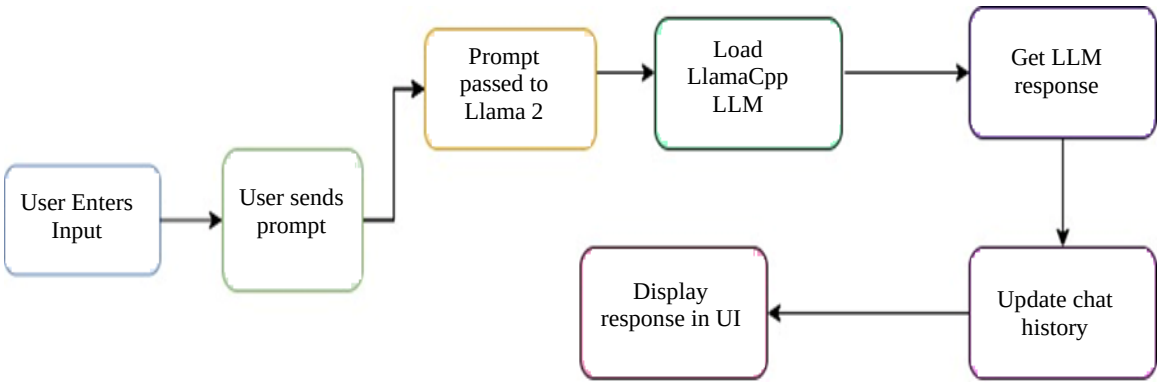

Figure 1. Solution architecture without Retrieval Augmented Generation (RAG).

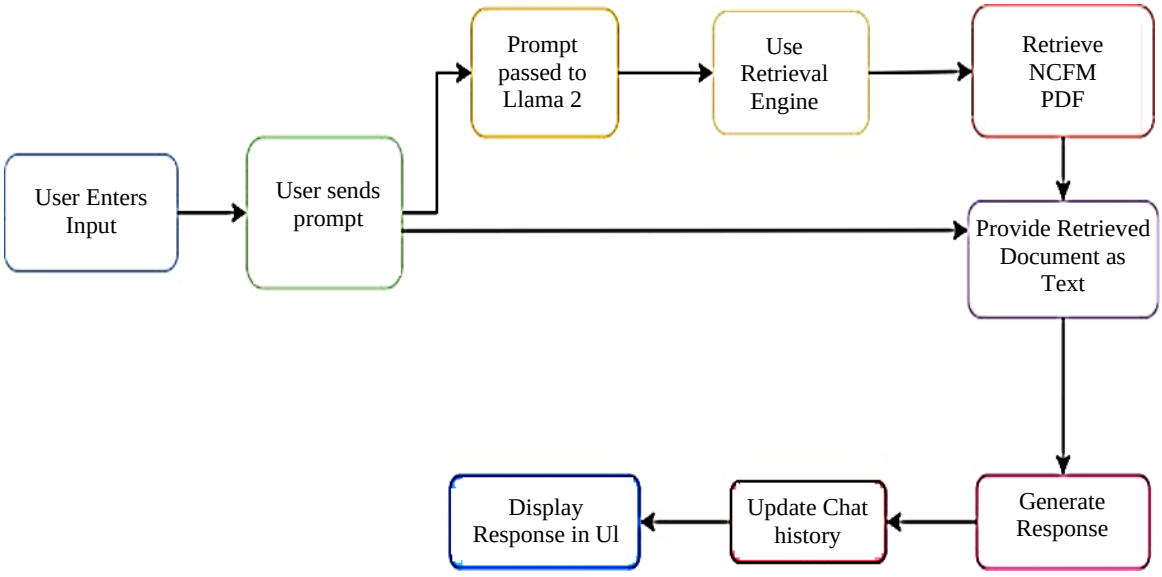

Figure 2. Solution architecture with Retrieval Augmented Generation (RAG).

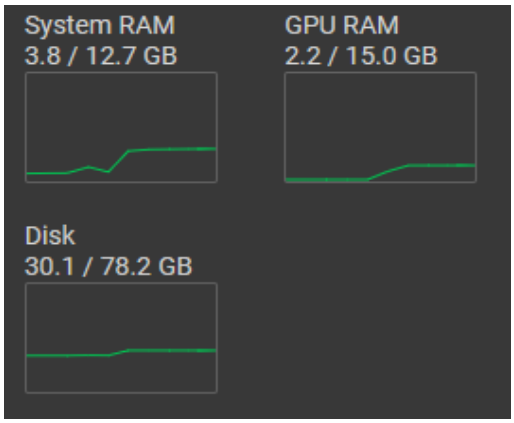

Figure 3. Resource usage on running the Llama 2 model with RAG implementation using the T4 GPU.

RAG with LlamaIndex

A data framework for LLM-based apps is called LlamaIndex. It makes use of Retrieval Augmented Generation (RAG) systems, which fuse a private knowledge base with massive language models. The indexing stage and the querying stage are the two stages that make up RAG in general. Your data is put into RAG and "indexed", or ready for queries. The index is impacted by user searches, which narrow down your data to the most pertinent context. This framework has been used to create a vector index of the NCFM book [9] so that it can be queried and converted to text to be sent in the response to the user [17].

Chatbot

Streamlit, a python framework, was used as front end means to create the chatbot. The chatbot stores the user prompts and the LLM responses as key value pairs in a maintained chat session history using Streamlit's session package [18–20].

RESULTS AND DISCUSSION

Questions prompted to TradeBot from each of the three categories were: from chart patterns, the authors asked "What is a shooting star candle stick pattern?"; from trading strategies, the authors asked "Under Dow Theory, a reversal from a bull market to a bear market or vice versa is not signaled until and unless both indexes, i.e., the Industrial Index and the Transportation Index, confirm the reversal. True or False?"; and lastly from trading psychology, the authors asked, "Why do trends never turn on a dime?". Degree of hallucination as to how far off the chatbot's responses are from the factual and conceptually correct information, about technical analysis is measured. The results showed that the Llama model had just the surface level knowledge (basic familiarity with terms) about technical analysis, so it generated contextually correct but conceptually inaccurate information with no in-depth insights and hence poor justifying ability.

The chatbot was prompted for two iterations. The first iteration consists of hallucinated/inaccurate responses. The second iteration consists of RAG enabled responses. The figures ahead are snapshots of the chatbot UI that showcase the generated responses.

First Iteration of Responses

For the first question in the first iteration (Figure 4), the model unusually referenced a German article with false information in their response which was neither mentioned in the user prompt nor implied otherwise. The response then continues to incorrectly define the shooting star candlestick pattern as a bullish reversal pattern when factually, it is a bearish reversal pattern. The actual pattern is formed during the uptrend of the market price but the response states that it is formed during the downtrend. The model manages to get one fact right, which is the upper shadow of the body being longer than the

body as a candle stick pattern has an inverted hammer present in it with its upper shadow being twice as big as its body. For the second question in the first iteration (Figure 5), the response stated by the model is not entirely accurate according to Dow Theory. According to Dow Theory, confirming indications for a trend reversal should be provided by the Dow Jones Transportation Average (DJTA) as well as the Dow Jones Industrial Average (DJIA).

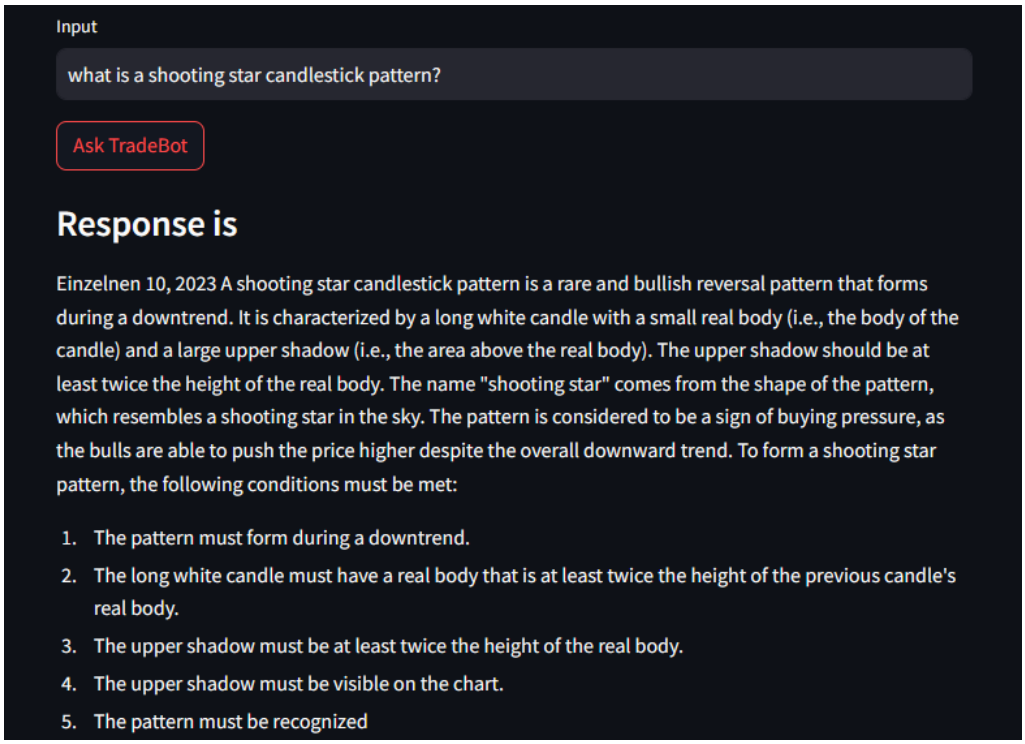

Figure 4. Response to the question, "What is a shooting star candlestick pattern?".

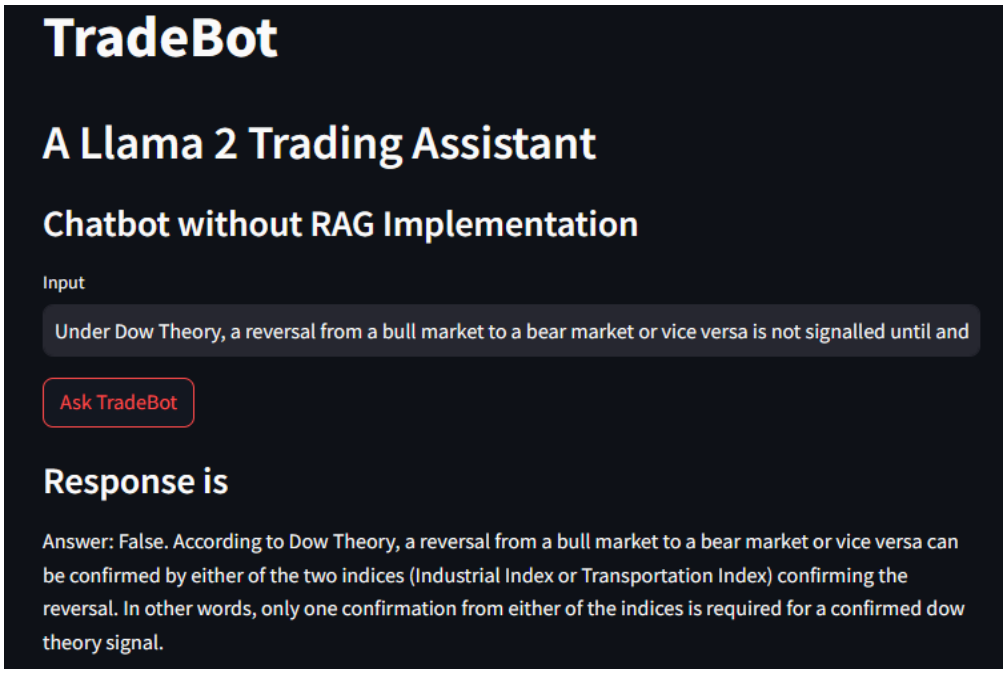

Figure 5. Response to the question, "Under Dow Theory, a reversal from a bull market to a bear market or vice versa is not signaled until the Industrial Index and the Transportation Index, confirm the reversal. True or False?".

In Dow theory, the Industrial Average represents the production sector, and the Transportation Average represents the transportation sector. The idea is that for a healthy economy, both sectors must be in sync. Therefore, if one index shows a trend reversal, the other should follow suit to confirm the change in market conditions. So, according to Dow Theory, it is not sufficient for only one of the indices to confirm a reversal. The expression "turn on a dime" is an idiom that means to change direction quickly and sharply, similar to the way a coin can pivot rapidly when flipped. In the context of trends, it suggests the uncertain nature of trends and how trends can shift patterns or continue in their ongoing nature of the market. In the response (Figure 6), the model however fails to understand this and deviates into associating trends with the fashion industry. While the information in the referenced article seems plausible, the response is contextually unreliable. The model again references a German article which goes to show the model is sitting on flawed data.

Second Iteration of Responses

In the second iteration of responses, for the same number of tokens, the model is able to answer responses, both factually correct and contextually correct. For the first question (Figure 7), the model answers precisely about what a shooting star candlestick pattern depicts, when it is found, and its structure. The pattern is explained correctly by the model without stating extra information that is not relevant to the content.

For the second question (Figure 8), the model replies to the point by simply stating the answer as True and does not justify a hallucinated reply unlike in the first iteration. More importantly, it does not deliberately try to leverage more false information to create a believable response. The model now understands the principles of Dow Theory and how it affects the economy. If asked to explain further, the model will guarantee, query the right information about Dow's theory from the book to elaborate since it now has access. For the third question (Figure 9), the model finally interprets the idiom rightfully and associates it with trading to explain how market participants go against the trend (buy in bearish market and sell in a bullish market). Hence implying the abrupt change in the market conditions and the uncertain nature of trends.

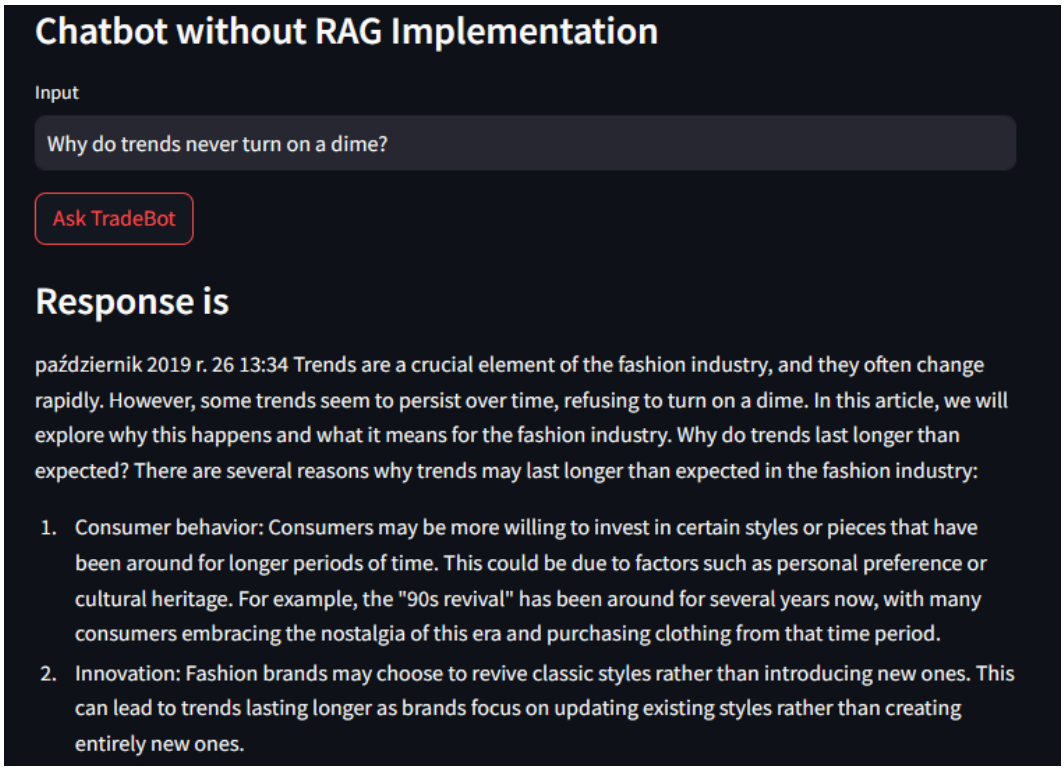

Figure 6. Response to the question, "Why do trends never turn on a dime?".

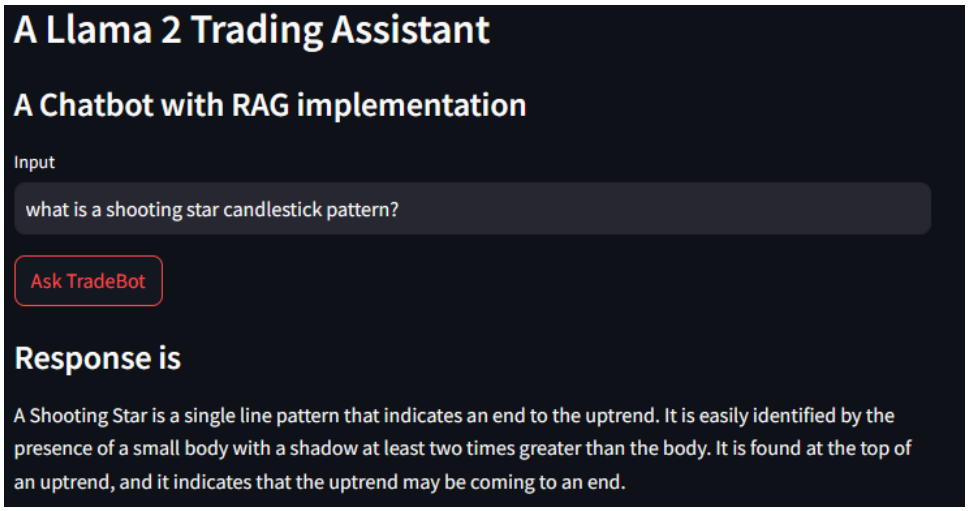

Figure 7. Shooting star candlestick pattern.

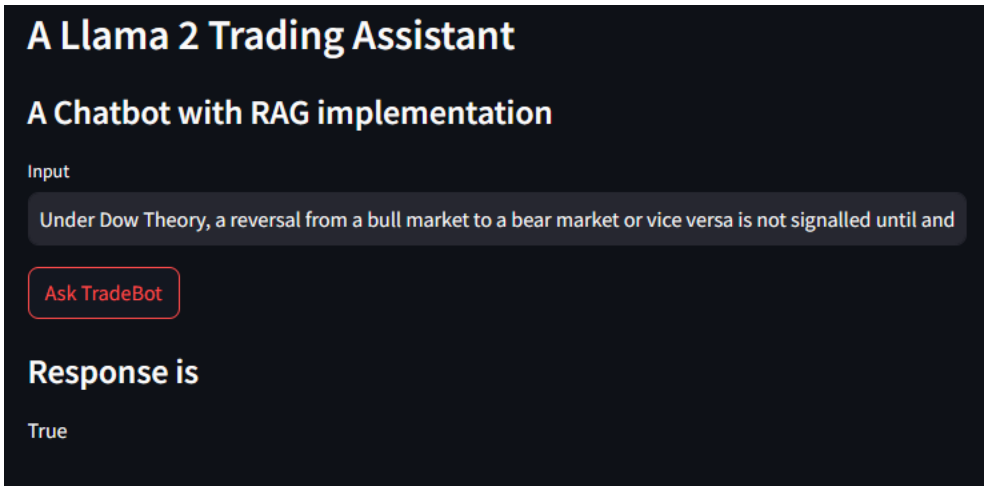

Figure 8. Under Dow Theory, a market reversal is confirmed when both the Industrial and Transportation Indexes agree. True or False?.

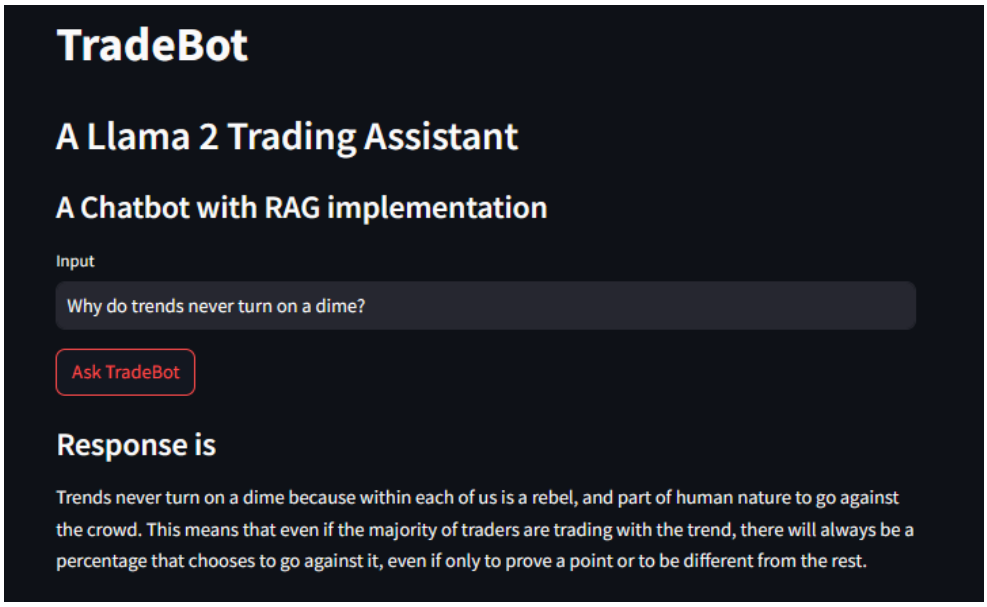

Figure 9. Why do trends not change abruptly?.

CONCLUSION

Large Language Models (LLMs) and their perils co-exist. LLM hallucinations is one such characteristic which can add to or deteriorate the performance of the desired functionality that the model is being used for. Hallucinations in generation tasks provide serious problems for AI's reliability and accountability. However, asking for suggestions or seeking diversity are two more situations where hallucinating might be desirable. Optimizing the output of a big language model by referencing a reliable knowledge base outside of its training data sources before producing a response is known as Retrieval Augmented Generation, or RAG. Because Llama 2 is built for assistant-like chat use cases and can be customized for a range of natural language generating jobs, it was the best option for our research. RAG is not designed to eliminate hallucinations from LLMs, but it serves the purpose of adding more functionality to the LLM. This is preferred in instances where LLMs are utilized for more personalized applications which according to this study, was a chatbot that helps individuals to learn about technical analysis for financial trading.

Limitations and Future Scope

In this study, the authors experimented with a GGUF quantized Llama 2 model by taking into account, the requirements of the computing system on which the experiment was performed.

There is pace for leveraging other quantization techniques including, GPTQ, and AWQ quantized Llama 2 models to explore the full potential of a system's CPU and GPU. The Llama 2 chat model comes in two other parameterized versions, one having 13 billion and the other having 70 billion, both of which can be utilized. Modern model evaluation benchmarks can be used to evaluate the accuracy of the model. A greater number of tokens (more than 256) can be used for more accurate responses and for a stronger test against hallucinated responses.

REFERENCES

1. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023 Feb; 15(2): e35179.
2. Sharun K, Banu SA, Pawde AM, Kumar R, Akash S, Dhama K, Pal A. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references–a preliminary study. Ann Med Surg. 2023 Oct 1; 85(10): 5275–8.
3. Krause D. Mitigating risks for financial firms using generative AI tools. Available at SSRN 4452600. 2023 May 18.
4. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. Crit Care. 2023 May 10; 27(1): 180.
5. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023 Mar 21; 27(1): 120.
6. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst. 2020; 33: 9459–74.
7. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023 Jul 18.
8. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, Schuh P. Palm: Scaling language modeling with pathways. J Mach Learn Res. 2023; 24(240): 1–13.
9. Block P, Tower R, Ring I. The National Stock Exchange of India Limited. 2023. Available from https://announcement.acesphere.com/Annoucement/20240209/c85b6347-29c8-4ee5-87fc-ae66ad22fd3c.pdf
10. Tam A. (2023 Jul 20). A gentle introduction to hallucinations in large language models. [Online]. Machine Learning Mastery. https://machinelearningmastery.com/a-gentle-introduction-to-hallucinations-in-large-language-models/

11. Li Y, Panda R, Kim Y, Chen CF, Feris RS, Cox D, Vasconcelos N. Valhalla: Visual hallucination for machine translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; 5216–5226.
12. Li J, Cheng X, Zhao WX, Nie JY, Wen JR. Halueval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:2305.11747. 2023 May 19.
13. Aurangzeb Ahmad M, Yaramis I, Dutta Roy T. Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. arXiv e-prints arXiv: 2311. 2023 Sep.
14. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922. 2023 Sep 12.
15. Tonmoy SM, Zaman SM, Jain V, Rani A, Rawte V, Chadha A, Das A. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313. 2024 Jan 2.
16. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316. 2022 Oct 13.
17. Frantar E, Ashkboos S, Hoefler T, Alistarh D. Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323. 2022 Oct 31.
18. Lin J, Tang J, Tang H, Yang S, Chen WM, Wang WC, Xiao G, Dang X, Gan C, Han S. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. Proceedings of Machine Learning and Systems. 2024 May 29; 6: 87–100.
19. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing E, Zhang H. Judging llm-as-a-judge with mt-bench and chatbot arena. Adv Neural Inf Process Syst. 2023 Dec 15; 36: 46595–623.
20. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S. Superglue: A stickier benchmark for general-purpose language understanding systems. Adv Neural Inf Process Syst. 2019; 3266–3280.

