East African Journal of Information Technology
eajit.eanso.org
Volume 7, Issue 1, 2024
Print ISSN: 2707-5346 | Online ISSN: 2707-5354
Title DOI: https://doi.org/10.37284/2707-5354

EANSO
EAST AFRICAN NATURE & SCIENCE ORGANIZATION

Original Article

Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs

Tibakanya Joseph[1*] *& Male Henry Keneth*[1]

[1] Makerere University, P. O. Box 7062, Kampala, Uganda.
* Correspondence ORCID ID: https://orcid.org/0000-0001-7368-3808; Email: jtibakanya@gmail.com

Article DOI: https://doi.org/10.37284/eajit.7.1.2111

Date Published:

15 August 2024

Keywords:

Large Language Models (LLMs), LLM Hallucination, Grammar-Aware Prompt Engineering (GAPE), Formal Methods (FMs).

ABSTRACT

Recent advancements in Artificial Intelligence (AI), particularly in the advanced machine learning for the Natural Language Processing (NLP) paradigm, have led to the development of powerful Large Language Models (LLMs) capable of impressive feats in tasks like translation, text summarisation, text generation and code generation. However, a critical challenge hindering their real-world deployment is their susceptibility to hallucinations, where they generate plausible looking but factually incorrect outputs. These limitations come with adverse effects, such as the propagation of misinformation and reducing user trustworthiness in the related technologies, even when they possess transformative potential in various sectors. This study aims to enhance the performance of LLMs by presenting a new strategy that combines grammar-aware prompt engineering (GAPE) and formal methods (FMs) to leverage their synergy in the LLM process logic. We argue that by combining linguistic principles using GAPE and constructing a basis of formal structures using FMs, we could improve the LLM's ability to analyse language, decrease ambiguity in prompts, improve consistency in output, and eventually, greatly diminish LLM hallucinations. To do this, we propose a collaboration between linguists and AI experts while also providing specialised training for LLMs that emphasises linguistic precision. Additionally, we suggest implementing iterative design and development procedures for LLMs that use GAPE and FM principles to continuously enhance the performance of LLMs. By following these techniques, we may create a future in which LLMs are more trustworthy for a wide range of users and use cases with reliable LLM technologies and efficient advancements in practical situations.

APA CITATION

Joseph, T. & Keneth, M. H. (2024). Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs. *East African Journal of Information Technology*, *7*(1), 188-201. https://doi.org/10.37284/eajit.7.1.2111

CHICAGO CITATION

Joseph, Tibakanya and Male Henry Keneth. 2024. "Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs". *East African Journal of Information Technology* 7 (1), 188-201. https://doi.org/10.37284/eajit.7.1.2111.

HARVARD CITATION

Joseph, T. & Keneth, M. H. (2024) "Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs", *East African Journal of Information Technology*, 7(1), pp. 188-201. doi: 10.37284/eajit.7.1.2111.

IEEE CITATION

T. Joseph & M. H. Keneth "Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs", *EAJIT*, vol. 7, no. 1, pp. 188-201, Aug. 2024.

MLA CITATION

Joseph, Tibakanya & Male Henry Keneth "Exploring the Synergy of Grammar-Aware Prompt Engineering and Formal Methods for Mitigating Hallucinations in LLMs". *East African Journal of Information Technology*, Vol. 7, no. 1, Aug. 2024, pp. 188-201, doi:10.37284/eajit.7.1.2111.

INTRODUCTION

The ubiquity of pre-trained Large Language Models (LLMs) has revolutionised natural language processing tasks, demonstrating remarkable forms of applications in various sectors, including automated natural language translation (ANLT), automatic text summarisation (ATS), classification, medical care, and computer code generation. LLMs are hundred or even billion-parameter models, resource-intensive natural language processing models that are by far easy to use (de Wynter et al., 2023). The extensive number of parameters make LLMs (Bang et al., 2023). The most prevalent LLMs include GPT-3, ChatGPT, LLaMA, GPT-4, and Bard (Zhou et al., 2023). These are trained using a large corpus of text data and can create coherent and contextually appropriate replies to various questions and assertions (Huo et al., 2023). This massive corpus contains text documents from the internet, including books, news, research publications, public code repositories, and websites. The training data is preprocessed to exclude non-textual components before input into a transformer architecture trained as an unsupervised auto-regressive generative model (Jha et al., 2023). They are neural networks trained to assign probabilities to a series of texts to predict the next most likely word (Sartori & Orrù, 2023). LLMs have revolutionised human-computer interaction in artificial intelligence (AI) by enabling machines to produce text and content that progressively emulates communication authored by humans. They can create human-like text, pictures, audio, and other formats. Prompts are utilised to explore the LLMs and achieve the desired output (Bang et al., 2023). Consequently, LLMs have tremendously impacted various sectors of human life, thus expanding the limits of what computerised machines can do.

LLM solutions can traverse simple job completion to provide more exciting, contextually rich interactions that improve human-computer interaction experiences. Also, LLMs influence digital content generation, as they are a valuable AI tool for automating content creation across domains such as education, news stories, digital marketing materials, and narratives. LLMs are critical in language translation to break down language barriers and increase global communication and understanding. Furthermore, in the healthcare arena, they assist in the study of medical texts, as well as the development of medical reports, research papers, and the summarising of lengthy patient records (Athavale et al., 2023). In education, LLMs have been identified with the ability to give highly personalised and accessible no- or low-cost dynamic learning. They are poised to transform this dynamic and, as a result, help build a more diverse future in adaptive learning (Bommineni et al., 2023). Finally, e-commerce and recommendation systems employ LLMs to provide personalised content, increasing user engagement and personalising product suggestions. LLMs' importance may be seen in their adaptability, which permeates many facets of everyday life and industrial sectors.

Although these LLMs exhibit substantial potential for a wide range of applications, they are not devoid of obstacles and constraints (Li et al., 2023). One notable issue persistently giving rise to doubts regarding LLMs' dependability, security, and trust is hallucinations. Hallucinations refer to creating fictitious material unsupported by factual facts or unsuitable in context (Bang et al., 2023). Other scholars call it LLM confabulation (Rawte, Chakraborty, et al., 2023b) or, even LLM fabrications (Zhan et al., 2023), LLM falsification (Emsley, 2023) or simply LLM misinformation (Alberts et al., 2023). Hallucinations are said to be caused by poor quality of the data as well as the models used in the training process (Dziri et al., 2022), limited contextual understanding, a poor description of the user's intention and repetition (Sartori & Orrù, 2023), and poor optimisation techniques.

Hallucinations provide a significant barrier for AI developers and consumers since they may result in ethical and legal challenges (Athaluri et al., 2023), affecting the dependability and trustworthiness of LLM-generated material. For instance, in a study of the prevalence of AI hallucination in research proposals written exclusively by ChatGPT, out of the 178 references examined, 69 did not have a Digital Object Identifier (DOI), and 28 did not appear in a Google search or had an existing DOI (Athaluri et al., 2023). In another fascinating incident, a lawyer called Schwartz defended a client in a personal injury action against Avianca, the airline, when the latest event occurred (*My "Hallucinating" Experience with ChatGPT*, 2023). He employed ChatGPT to look out for legal precedents that might support his client's argument, and the chatbot produced many examples that Schwartz mentioned in his court submission. However, the instances created by ChatGPT were fabricated and contained false information (Giray, 2023). When the opposing counsel questioned the citations, Schwartz acknowledged using ChatGPT. Hallucinating erroneous responses may have negative monetary repercussions, such as when Google Bard hallucination cost the firm $100 billion during its debut, as well as societal ramifications, such as when ChatGPT mistakenly accused a professor of sexually abusing students (Ahmad et al., 2023). These LLMs can automatically generate disinformation and difficult-to-detect because of recent breakthroughs in large-scale pre-trained models (e.g., BERT, GPT-3, GPT-4) and adversarial learning (Islam et al., 2020). We thoroughly investigated two foundational pillars for tackling these concerns within this framework: integrating FMs and GAPE. Therefore, hallucination is a critical issue that must be mitigated to ensure the trustworthiness and reliability of LLMs (Jha et al., 2023).

Several measures have been suggested to solve the hallucination conundrum in LLMs. On the one hand, general measures aim to solve this problem mainly with the human-in-the-loop, algorithmic correction, fine-tuning, and application of sophisticated prompt engineering (Sartori & Orrù, 2023). On the other hand, there are more specialised approaches. For instance, Nie et al. (2019) offer a strategy combining a language understanding module for data refinement with self-training iterations to induce substantial equivalence between the input data and the matched text and minimise hallucination by more than 50%. MixCL by Sun et al. (Sun et al., 2023) involves a contrastive learning scheme that performs similarly to other state-of-the-art approaches but displays high efficacy and scalability. The other is the Chain of Natural Language Inference (CoNLI), a plug-and-play hierarchical framework for detecting and mitigating hallucination without fine-tuning or domain-specific prompt engineering (Lei et al., 2023). In addition, SELF-FAMILIARITY is yet another unique pre-detection self-evaluation technique that focuses on assessing the model's familiarity with the context provided in the user prompt instruction and deferring response production in the event of unknown concepts (Luo et al., 2023a). Finally, another unique solution called "WikiChat", which is grounded in Wikipedia and is one of the most extensive hand-curated text corpora available for the public achieves 97.3% factual accuracy (Semnani et al., n.d.). However, it is essential to note that none of these tackles the technical causes of hallucination related to the user prompt and user language. In other words, they may not handle the ambiguities caused by the user's language in the prompts.

In this position paper, we propose a complementary line of action, arguing that a combined integration of GAPE and FMs in LLMs provides a substantial approach to addressing the hallucinations by improving contextual understanding, improving language parsing and processing, resolving ambiguities, enabling more efficient error detection and correction. The primary objective of this paper is to illuminate the struggle to eradicate hallucinations in LLMs by analysing the diverse and complex contributions of a combination of FMs and GAPE to the advancement of LLMs, thereby presenting a formidable resolution to the issue of LLM hallucinations and the misinformation problems that come it.

RELATED WORKS

The issue of hallucinations is not an entirely new problem; some scholars have attempted to provide solutions to it. Most of these employ techniques and approaches in searching for reliable and hallucination-free LLMs. The LLM robustification strategies can be classified as general or specific, addressing specific LLM challenges that lead to hallucination (Y. Li et al., 2023). On the one hand, the available interventions are Human-in-the-loop, algorithmic correction, fine-tuning and application of sophisticated domain-specific prompt engineering (Sartori & Orrù, 2023). On the other hand, specific interventions address the problem from unique perspectives. For instance, to reduce hallucinations, Sun et al. (2023) propose a contrastive learning scheme called Mixed Contrastive Learning (MixCL), a unique solution that overtly optimises the implicit knowledge generation process LLMs. Mixed contrastive objective and negative sampling are employed to reinforce the solution. The solution’s efficacy is based on experimental evidence conducted on Wizard-of-Wikipedia, a public, open-domain knowledge-grounded dialogue benchmark and human assessment. It returns excellent results, reduces hallucinations, and displays unique benefits in efficacy and scalability.

In another endeavour, Lei et al. (2023) developed a framework to detect and tackle hallucinations using a plug-and-play framework named Chain of Natural Language Inference (CoNLI) through post-editing. The solution achieves high-level performance without fine-tuning or any domain-specific prompt engineering. Nie et al. (Nie et al., 2019) proposed another unique proposal in which a language understanding module is used for refining data. Iterative self-training would create a substantial equivalence between the input facts and the paired text. The solution reduces hallucination by over 50% of the original data-text pairings' relative unaligned noise. The authors note a need to add features for lexical choices. Luo et al.(2023b) tackle the hallucination issue from a fascinating perspective involving self-evaluation techniques. Their solution, SELF-FAMILIARITY, is a pre-detection self-evaluation strategy that focuses on assessing the model's familiarity with the ideas in the input instruction and deferring response production in the event of unknown concepts. The solution brings a new trend towards proactive hallucination reduction measures in LLMs, promising reliability, applicability, and interpretability increases. According to (Jones et al., 2023), optimising LLMs to hallucinate less is complex since hallucination is challenging to measure correctly at each optimisation phase. The authors use SynTra to develop a synthetic task that may decrease hallucination on real-world downstream tasks. SynTra creates an artificial task where hallucinations are simple to generate and assess. It then optimises the LLM's system message on the artificial job via prefix-tuning and ultimately transfers the system message to actual, difficult-to-optimize workloads. They show that the techniques can address the issue even when better optimisation techniques (like LoRA) are employed. Lastly, by bringing humans in the loop, (Semnani et al., n.d.) present WikiChat, which achieves over 97.3 percent factual accuracy, outperforming even fine-tuned models. This only emphasises the incorporation of a human in the loop. From this analysis, none of the measures address the hallucination issue more comprehensively.

METHODOLOGY

This position paper uses a literature review methodology to analyse the difficulties associated with hallucinations in Large Language Models (LLMs) and provide possible solutions to mitigate them. The research approach included the following stages: Initially, we conducted an extensive exploration of academic databases, industry reports, and pertinent articles by using specific keywords such as 'Grammer-aware prompting', 'LLM hallucinations', 'LLMs', 'Prompt Engineering', and 'Formal Methods'. Furthermore, the identified sources underwent a meticulous assessment, considering their pertinence, reliability, and thoroughness of analysis. The emphasis was on scholarly works that have undergone peer review, conference papers, and industry reports from reliable sources. Furthermore, a thorough analysis was conducted on the chosen literature to extract significant discoveries, highlight issues related to hallucinations, and investigate current approaches for mitigating these challenges. The analysis also emphasised using user input to improve the quality of output from LLMs. Finally, after collecting data, a thorough analysis was performed to provide a logical case in favour of a particular technique involving the combined effects of grammar-aware prompting and Formal approaches to reduce hallucinations in LLMs.

LLM Hallucinations

LLMs may inadvertently generate plausible-looking information that is full of errors (Dhuliawala et al., 2023). This phenomenon is termed hallucination. It involves the generation of outputs that stray from contemporary factual reality to incorporate fabricated information (Rawte, Sheth, et al., 2023a). Such outputs may include fictional claims, misinformation, or fabrications rather than the presentation of reliable and truthful information. Generally, hallucinations in natural language generation and processing may be classified into two: intrinsic hallucinations and extrinsic hallucinations (Huang et al., 2023a). All these can happen unintentionally and may result from various factors. Common factors include biases in the training data, the model's lack of access to real-time or up-to-date information, and the model's inherent limitations in comprehending and generating contextually accurate responses. In other words, Hallucinations are identified to have multifaceted causes, covering the entire spectrum of the models' capability development process (Huang et al., 2023a). Tackling the hallucination challenge may be challenging due to the large volume of data used in the training process, the imperceptibility of LLM hallucination by humans and the versatile nature of LLMs as general-purpose technologies (Zhang et al., 2023).

Types of LLM Hallucinations

Currently, we can categorise LLM hallucinations in three ways as follows:

- Input-conflicting hallucination, in which LLMs generate content that is not intermittent with the source input provided by users.
- Context-conflicting hallucination, in which LLMs generate outputs that contradict earlier generated information and
- Fact-conflicting hallucination, in which LLMs produce outputs incongruent with established global knowledge.

Causes of LLM Hallucinations

Some of the leading causes of hallucinations in LLMs include.

- **Poor quality data used to train the models.** Potential data-related causes include faulty sources and inefficient utilisation, poor training strategies that may cause hallucinations during pre-training and alignment (Huang et al., 2023a), and those resulting from the stochastic nature of decoding strategies and imperfect representations during the inference process (Huang et al., 2023b).
- **Insufficiency of the user prompt used to generate outputs**. LLM's results depend on the quality of the user prompt. Well-prepared

prompts always generate reliable outcomes, while poor ones yield false knowledge and misinformation (Huang et al., 2023a) mainly due to the user's poor language (Heston & Khun, 2023). LLMs employ a variety of innovative prompting tactics to tackle specific difficulties and get the desired results. Three main approaches stand out in particular: the Input-Output (I/O) method, the Chain of Thought (CoT), and the Tree of Thoughts (ToT) prompts (Abedi et al., 2023).

- **LLM self-contradiction**. Lack of self-consciousness and sequenced reasoning always result in the LLM's self-contradiction (Ahmad et al., 2023). They tend not to tolerate output critical evaluation consistently and lack reasoning and argumentative capacity, generating hallucinations (Heston & Khun, 2023)**.**
- **Output generation techniques based on probability**. Using probabilistic methods to generate outputs allows for the possibility of recombining very reliable information to produce plausible-looking information (Huang et al., 2023a).
- **Using biased data in training the model.** This can lead to suffering from the stale information issue, as demonstrated by recent research, in which the model outputs are based on obsolete or incorrect knowledge owing to bias in the dataset or the model's failure to keep up with developing understanding (Jha et al., 2023).

Prompt Engineering

Nowadays, prompt engineering is one of the most significant skills one must have if one is to engage in interaction with LLMs (White et al., 2023). Prompts augment LLMs with task-specific cues, adapting them to new tasks (Gu et al., 2023). They are instructions presented to LLMs to ensure specific content qualities are generated in the output (White et al., 2023). Prompt engineering is the process of crafting, iteratively refining, and optimising prompts to define the user's intention for the LLM (Ekin, 2023). Users can guide the LLMs in bypassing the limitations and restrictions by meticulously designing and refining prompts (Liu et al., 2024). This means it is a way to program the LLM for user-specific outputs (Wang et al., 2024). Prompt engineering provides several benefits over conventional engineering. First, adapting a pre-trained model to new tasks requires only a few labelled data, reducing human supervision and computer resources for fine-tuning. Second, prompt engineering allows a pre-trained model to forecast new tasks based on the prompt without altering its parameters, enabling it to serve many downstream jobs (Gu et al., 2023). This allows real-world use of huge pre-trained models. The LLM instructions have also been referred to as discrete prompts or complex prompts, while the internal vector representations are called continuous prompts or soft prompts (Liu et al., 2024). While the LLM's efficacy significantly relies on algorithms and training data, it depends so much on prompt quality (Bozkurt & Sharma, 2023). Since the quality of the prompts directly affects the quality of the output generated, understanding the nuances of user prompt engineering is critical for creating compelling and meaningful interfaces with LLMs.

FORMAL METHODS (FMS) AND GRAMMAR-AWARE PROMPT ENGINEERING (GAPE)

Formal methods are rigorous techniques and tools for specifying, designing, and verifying hardware and software based on mathematics (Patterson, 2013). On the other hand, GAPE is a method of imbuing AI models with linguistic rules and structures (Ssanyu et al., 2021), allowing them to negotiate the complex terrain of human language with accuracy and coherence. It aims to deliberately construct prompts with a careful study of grammatical structures to improve a language model's performance or behaviour. A robust approach to alleviate hallucination difficulties in LLMs exists at the intersection of GAPE and FMs. GAPE focuses on language structure refinement and context awareness, while FMs give a systematic, rule-based approach to rigorous analysis. Combining these techniques yields a complete strategy: GAPE enhances contextual understanding by refining language components (Ssanyu et al., 2021), whereas FMs provide a systematic framework for effective mistake detection and repair. They form a synergistic alliance to combat LLM hallucinations via complex language parsing, increased contextual understanding, and robust error mitigation.

Critiques and Considerations for GAPE and FMs in LLMs

Some scholars may argue that integrating GAPE and FMs brings about computational complexities. Adhering to strict grammar rules, for example, may drastically increase the computing resource needs, thereby increasing the processing time required to generate an output from a user prompt. In turn, this can dramatically slow down real-time human-computer interactions, thus spoiling the benefits of LLMs in certain situations. However, Patterson (Patterson, 2013) states that formal methods can be applied in more specific instances through simpler notations in lightweight formal methods. Lightweight formal techniques may be used in a system focused on certain features of that component, without complete mathematical representation, to discover flaws rather than gain mathematical proof. Nonetheless, from the ethical point of view, the trade-off between computing complexity and LLM linguistic correctness is so fundamental that we can’t ignore it. AI developers can apply optimisation techniques to balance accuracy and efficiency (Patterson, 2013). For instance, pre-processing, parallel processing, and LLM optimisations can be employed to reduce the computational burdens from escalating. AI developers can harness advancements in algorithms and hardware to improve efficiency (Franceschelli & Musolesi, 2023) and thus incorporate formal methods without jeopardising real-time human-computer interactions in practical situations and scenarios.

Similarly, others may argue that strict adherence to grammatical rules and formal methods concepts may limit the LLMs’ capacity to develop creative content or novel solutions. Also, this may reduce the power of these models to produce varied and surprising content, thus inhibiting the generation of innovative ideas. However, integration of these concepts may not limit the LLM's capacity to act creatively but rather act as a scaffold for their creativity. LLMs use advanced machine learning techniques, which means the models can be trained to apply the rules flexibly, allowing them to behave creatively. For instance, integrating the GAPE in LLMs does not limit creativity; instead, it may provide a framework within which the models could deviate flexibly. Further, LLMs train with a large corpus of text data, exposing them to various language styles and creative phrases, allowing creativity within well-established language rules.

Others may argue that overemphasis on formal language principles may lead to an overreliance on rule-based frameworks, causing the AI to miss out on identifying colloquialisms, idiomatic expressions, or differences in language use. This may result in sterile or rigid responses lacking the flow and richness of natural language conversations. However, integrating grammar rules doesn’t do away with comprehension of popular expressions or idiomatic language. The fact that LLMs are trained on various datasets means they include different morphological styles and phrases. Including GAPE does not imply rigorous structural observance but mindfulness and applying rules in context. LLMs may be furnished with systems for contextual interpretation and response, allowing for flexibility in language usage while retaining grammatical correctness.

The Role of Grammar-aware Prompts and FMs in LLMs

As mentioned earlier, LLM hallucinations offer several critical impediments when using the models in various life sectors. However, these issues are manageable and thus can be overcome. We posit that combining GAPE and FMs can solve the hallucination challenges allied to the LLMs’ prompting by users. This can be done in several ways:

Prompt context misinterpretation

The quality of the user prompt is as good as its output (Ji et al., 2023). Hallucinations and language parsing problems are often caused by a lack of contextual comprehension, mainly because of the user language. We posit that following grammatical structures and formal language rules can be reduced. When prompted with syntax-preserving but differently worded prompts and different-syntax but comparable semantics prompts, LLMs yield inconsistent results, suggesting that LLMs are inadequate for reliably extracting factual information and that prompt syntax is essential. While some scholars argue that we cannot completely do away with hallucinations in LLMs because of how tokens are generated [18], integrating GAPE and FMs in LLMs prompts used in the generating outputs serves as a solution to the problem. This reduces the risk of producing meaningless or contextually incorrect replies. Failure of the decoder component of the LLMs to understand the context of the user requirements heavily relies on the grammar and related syntax in the user prompt. The role of GAPE is to meticulously refine the language structure of the user prompt, thus fostering grammatical accuracy. Although LLMs already capture some syntactic information, utilising syntactic information when training sophisticated models for knowledge extraction can boost performance (Dietze et al., n.d.). Similarly, using FMs provides an efficient modelling framework enabling extremely exact representation. In other words, a combination of GAPE and FMs in LLM prompt structures enhances prompt context interpretation with high levels of accuracy, considering the subtle nuances (Patterson, 2013) Therefore, accurate, prompt context interpretation generates plausible content, eliminating hallucinations in LLMs.

Precision in Language Parsing and Processing

The quality of the output from the user question depends on the precision of the LLM language parsing and processing. Higher language parsing and processing precision means quality and contextual output (Ssanyu et al., 2021). Consequently, LLMs cannot hallucinate erroneous and plausible-looking results that are irrelevant or factually incorrect with precise language parsing and processing. Language comprehension reduces grammatical ambiguities, which might lead to hallucinations or misinterpretations. For example, a better comprehension of subject-verb agreements, phrase patterns, and pronoun references minimises mistakes and improves the naturalness of LLM replies. Integrating GAPE means improved grammatical accuracy (Patterson, 2013). In other words, user prompts are checked to ensure they adhere to linguistic rules. On the other hand, incorporating FMs gives AI developers a systematic and formalised framework for language modelling. This way, formal techniques can support the identification of inconsistencies during iterative prompting testing until it converges to a proper answer acceptable to the formal verifier (Jha et al., 2023). This synergy enables LLMs to parse and process language with meticulous attention to linguistic rules, ensuring a more accurate representation. Therefore, achieving precision in language parsing and procession is critical in preventing errors that might contribute to hallucinations, fostering a more accurate representation of user intent.

Resolution of Linguistic Ambiguities

Language ambiguities also play a significant role in causing hallucinations. The decoder can create appropriate tokens only if there are no ambiguities in the prompt, thus leading to misinterpretation. It is founded on the premise that LLMs use prompts entered by human language, which follows syntactic structures, grammatical rules, and sophisticated semantics (Perzylo et al., 2015). By incorporating these criteria into the architecture of LLMs, we enable them to read and create language that closely matches human linguistic expectations. For instance, for each noun and verb in an excellent prompt phrase existing in a parsed sentence, a matching item from the knowledge base is sought (Perzylo et al., 2015) This means that using proper grammar in these robust AI systems gives them the capacity to cope with and appropriately resolve ambiguities.

If an LLM can improve at recognising and resolving ambiguities by following grammatical rules, it gets more dependable at providing correct and contextually relevant outputs. FMs offer a formal foundation for disambiguation. Formal approaches seek to expose system ambiguity, incompleteness, and inconsistency. This would boost our three quality indicators (accuracy, completeness, and usability). This means that using formal language rules can aid in the disambiguation of statements and reduce possible mistakes caused by different interpretations (Patterson, 2013). Further, the synergy between GAPE and FMs contributes to resolving linguistic ambiguities in LLMs. Together, they address linguistic ambiguities, which are crucial in mitigating hallucinations. By clarifying linguistic complexities, the integrated methodology ensures LLMs can reduce the risk of generating plausible-looking content, contributing to more accurate and reliable language processing.

Efficient Error Detection and Correction

No matter how wrong a user prompt may be linguistically, LLMs may never return a message to the user but instead interpret them and produce any form of output based on the decoder's interpretation. In other words, no form of feedback is provided to the user to alert them of lingual mistakes. AI developers and users could adopt syntax-aware prompt engineering to overcome this (Patterson, 2013). The use of grammatical rules is not limited to linguistic correctness; it also works with fact-checking methods to confirm the factual accuracy of created information (Ssanyu et al., 2021). If LLM-generated text breaks linguistic standards, it may trigger a review of its objective correctness, improving information dependability. Inspired by the Temporal Logic of Action (TLA) and other formal approaches, formal language testing may be used for AI models to find and rectify language faults. Combined with FMs, the technique could offer a more efficient LLM error detection and correction. This streamlines the identification and correction of LLM-related errors. In addition, combining GAPE and FMs in GAPE could provide a repair technique. Providing a well-defined set of rules means the LLM can detect deviation of the user prompt from the standard grammatical structure and linguistic principles, allowing real-time error detection and correction. This capability could substantially reduce language parsing errors and solve hallucination challenges in LLMs. In short, efficient error detection and correction are pivotal in improving reliability in LLMs, promptly identifying and rectifying likely causes of hallucination. Ultimately, this can result in a more accurate and trustworthy output from LLMs.

DISCUSSION AND RECOMMENDATIONS

This work examined the transformative potential of grammar-aware prompt engineering (GAPE) in combination with formal methods (FMs) to address hallucinatory behaviour in LLMs. We found that user queries and prompts significantly influence the quality of LLM output. Therefore, we propose an integrative approach leveraging GAPE's capabilities to enhance language parsing and FMs to establish a solid foundation for more robust and reliable LLM outputs. Our investigation yielded several critical results. First, integrating GAPE and FMs could significantly improve the quality of LLM outputs by mitigating LLM hallucinations. This suggests that including linguistic principles and formal structures in the user query or prompt design could steer LLMs towards producing more objective and verifiable outcomes. Second, the recommendations outlined in this work provide an overview of the significance of collaboration among various stakeholders in the LLM development process. Using this background, we propose the following recommendations for future research endeavours:

- **Interdisciplinary Collaboration During LLM Development:** We support the intensification of collaboration among AI developers, ethicists, and linguists. These efforts can foster a deeper understanding of linguistic concepts within AI language models, thereby promoting the development of more robust, responsible and trustworthy LLMs.

- **Linguistically Rigorous Datasets for Model Training:** Data scientists and engineers should prioritise creating open-source datasets incorporating ethics and linguistic principles. These would empower NLP developers to train LLMs on ethical and linguistic rigour. Moreover, experimentation with such curated corpora could pave the way for the adoption of these concepts globally.

- **Iterative LLM Refining and Improvement:** We encourage AI developers and prompt engineers to sustain investment by exploring the transformative potential of GAPE and FM concepts for iterative LLM refining and improvements. By viewing LLM development as a continuous process rather than a one-time endeavour, we can ensure ongoing advancements in LLM capabilities.

- **Responsible LLM Development Considerations:** Practitioners and developers could establish a framework for responsible and ethical oversight in LLM development. This framework should ensure the incorporation of linguistic principles and responsible AI components that align with ethical considerations. For instance, tackling potential biases, misinformation, transparency, security and privacy, and societal consequences of LLM development should be the central tenets of this framework.

- **Training and Resources:** We recommend the development of training programs and resources specifically designed to equip AI developers and academics with the necessary knowledge and skills to apply GAPE and FM concepts, particularly formal language principles. Effective implementation of these principles necessitates a well-trained workforce.

- **Standardization of Language Rules:** Standardization efforts for language rules and principles within the AI development community are essential. This collaborative effort can establish a common framework for LLM development, promoting consistency and interpretability of AI models.

- **Sharing Best Practices in LLM endevours:** We encourage disseminating case studies and best practices from organisations and academic institutions that have successfully implemented GAPE and FM principles in their LLMs. Sharing real-world examples can provide valuable insights for the broader AI development community.

- **Evaluation Frameworks:** It is crucial to develop standardised frameworks for testing and evaluating LLMs. These frameworks should assess language quality, parsing accuracy, and the effectiveness of hallucination-reduction techniques. Standardised benchmarks will facilitate comparisons between different LLM models.

- **AI Policy and Legislation:** We advocate developing AI policies and legislation that promote the use of GAPE and formal language principles in AI systems. Policymakers should be informed about these approaches' potential benefits and drawbacks to ensure responsible AI development.

CONCLUSION

Conclusively, this paper examined the feasibility of integrating GAPE and FMs in developing and using LLMs. The paper’s fundamental premise stresses how these concepts combined have transformative potential in substantially changing LLMs landscape, improving language processing, reducing language parsing errors, leading to better decoding, less prompt ambiguity, more consistency, and more accurate and contextually appropriate outputs. These then can solve the hallucination issues in the LLMs. These benefits may extend to natural language processing, where language quality, coherence and reliability are crucial. AI developers can create a solid base through these guidelines. This can result in more reliable, coherent, and trustworthy LLM-generated content. The benefits of this guidance extrapolate to AI ethics because solving the hallucination issues also means the potential for misinformation is solved. As AI appears to be shaping all sectors of human life, from mere conversational tools to content production, the way users perceive it is crucial. This means dependability and trust will remain critical aspects. Addressing hallucination issues via the application of GAPE and FMs, in one way, contributes to the ongoing outcry to ensure that AI systems are responsible. LLMs will infiltrate all significant sectors; thus, adhering to responsibility principles becomes increasingly critical. In sum, while the AI era progresses, the proposed methodology serves as a beacon of hope, showing the future in which responsible AI-driven language is more coherent, trustworthy, and reliable with human linguistic expectations. Further investigation into GAPE and FMs in LLMs carries profound importance in advancing our understanding of this trending domain.

REFERENCES

Abedi, M., Alshybani, I., Shahadat, M. R. B., & Murillo, M. (2023). Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education. *Qeios*. https://www.qeios.com/read/MD04B0

Ahmad, M. A., Yaramis, I., & Roy, T. D. (2023). *Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI* (arXiv:2311.01463). arXiv. http://arxiv.org/abs/2311.01463

Alberts, I. L., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., & Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: What will the impact on nuclear medicine be? *European Journal of Nuclear Medicine and Molecular Imaging*, *50*(6), 1549–1552. https://doi.org/10.1007/s00259-023-06172-w

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., Duddumpudi, R. T. S., Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, *15*(4). https://doi.org/10.7759/cureus.37432

Athavale, A., Baier, J., Ross, E., & Fukaya, E. (2023). The potential of chatbots in chronic venous disease patient management. *JVS-Vascular Insights*, *1*, 100019. https://doi.org/10.1016/j.jvsvi.2023.100019

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity* (arXiv:2302.04023). arXiv. http://arxiv.org/abs/2302.04023

Bommineni, V. L., Bhagwagar, S., Balcarcel, D., Davatzikos, C., & Boyer, D. (2023). Performance of ChatGPT on the MCAT: The road to personalized and equitable premedical learning. *MedRxiv*, 2023–03.

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education, 18*(2), i–vii.

de Wynter, A., Wang, X., Sokolov, A., Gu, Q., & Chen, S.-Q. (2023). An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, *4*, 100024. https://doi.org/10.1016/j.nlp.2023.100024

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). *Chain-of-Verification Reduces Hallucination in Large Language Models* (arXiv:2309.11495). arXiv. http://arxiv.org/abs/2309.11495

Dietze, S., Jabeen, H., Kallmeyer, L., & Linzbach, S. (n.d.). *Towards syntax-aware pretraining and prompt engineering for knowledge retrieval from large language models*.

Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5271– 5285. https://doi.org/10.18653/v1/2022.naacl-main.387

Ekin, S. (2023). *Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices*. https://doi.org/10.36227/techrxiv.22683919.v2

Emsley, R. (2023). ChatGPT: These are not hallucinations – they're fabrications and falsifications. *Schizophrenia*, *9*(1), 52, s41537- 023- 00379– 4. https://doi.org/10.1038/s41537-023-00379-4

Franceschelli, G., & Musolesi, M. (2023). On the creativity of large language models. *arXiv Preprint arXiv:2304.00008*. https://arxiv.org/abs/2304.00008

Giray, L. (2023). Authors should be held responsible for artificial intelligence hallucinations and mistakes in their papers. *Journal of the Practice of Cardiovascular Sciences*, *9*(2), 161. https://doi.org/10.4103/jpcs.jpcs_45_23

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., & Torr, P. (2023). *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models* (arXiv:2307.12980). arXiv. http://arxiv.org/abs/2307.12980

Heston, T. F., & Khun, C. (2023). Prompt Engineering in Medical Education. *International Medical Education*, *2*(3), 198–205. https://doi.org/10.3390/ime2030019

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023a). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions* (arXiv:2311.05232). arXiv. http://arxiv.org/abs/2311.05232

Huo, S., Arabzadeh, N., & Clarke, C. L. A. (2023). *Retrieving Supporting Evidence for Generative Question Answering*. https://doi.org/10.1145/3624918.3625336

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining*, *10*(1), 82. https://doi.org/10.1007/s13278-020-00696-x

Jha, S., Jha, S. K., Lincoln, P., Bastian, N. D., Velasquez, A., & Neema, S. (2023). Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. *2023 IEEE International Conference on Assured Autonomy (ICAA)*, 149– 152. https://doi.org/10.1109/ICAA58325.2023.00029

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1– 38. https://doi.org/10.1145/3571730

Jones, E., Palangi, H., Simões, C., Chandrasekaran, V., Mukherjee, S., Mitra, A., Awadallah, A., & Kamar, E. (2023). *Teaching Language Models to Hallucinate Less with Synthetic Tasks* (arXiv:2310.06827). arXiv. http://arxiv.org/abs/2310.06827

Lei, D., Li, Y., Hu, M., Wang, M., Yun, V., Ching, E., & Kamal, E. (2023). *Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations* (arXiv:2310.03951). arXiv. http://arxiv.org/abs/2310.03951

Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 6449–6464.

https://doi.org/10.18653/v1/2023.emnlp-main.397

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating Object Hallucination in Large Vision-Language Models (arXiv:2305.10355). arXiv. http://arxiv.org/abs/2305.10355Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., & Liu, Y. (2024). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study* (arXiv:2305.13860). arXiv. http://arxiv.org/abs/2305.13860

Luo, J., Xiao, C., & Ma, F. (2023). *Zero-Resource Hallucination Prevention for Large Language Models* (arXiv:2309.02654). arXiv. http://arxiv.org/abs/2309.02654

Nie, F., Yao, J.-G., Wang, J., Pan, R., & Lin, C.-Y. (2019). A simple recipe towards reducing hallucination in neural surface realisation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2673–2679. https://aclanthology.org/P19-1256/

Patterson, J. L. (2013). *Parsing of Natural Language Requirements* [California Polytechnic State University]. https://doi.org/10.15368/theses.2013.227

Perzylo, A., Griffiths, S., Lafrenz, R., & Knoll, A. (2015). Generating grammars for natural language understanding from knowledge about actions and objects. *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2008– 2013. https://doi.org/10.1109/ROBIO.2015.7419068

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A. P., & Das, A. (2023b). *The Troubling Emergence of Hallucination in Large Language Models—An Extensive Definition, Quantification, and Prescriptive Remediations* (arXiv:2310.04988). arXiv. http://arxiv.org/abs/2310.04988

Rawte, V., Sheth, A., & Das, A. (2023a). *A Survey of Hallucination in Large Foundation Models* (arXiv:2309.05922). arXiv. http://arxiv.org/abs/2309.05922

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *14*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10629494/

Semnani, S. J., Yao, V. Z., Zhang, H. C., & Lam, M. S. (n.d.). *WikiChat: Combating Hallucination of Large Language Models by Few-Shot Grounding on Wikipedia*.

Ssanyu, J., Bainomugisha, E., & Kanagwa, B. (2021). PAMOJA: A component framework for grammar-aware engineering. *Science of Computer Programming*, *211*, 102703. https://doi.org/10.1016/j.scico.2021.102703

Sun, W., Shi, Z., Gao, S., Ren, P., de Rijke, M., & Ren, Z. (2023). Contrastive learning reduces hallucination in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(11), 13618–13626. https://ojs.aaai.org/index.php/AAAI/article/view/26596

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Pan, Y., Liu, Z., Sun, L., Li, X., Ge, B., Jiang, X., … Zhang, S. (2024). *Prompt Engineering for Healthcare: Methodologies and Applications* (arXiv:2304.14670). arXiv. http://arxiv.org/abs/2304.14670

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (arXiv:2302.11382). arXiv. http://arxiv.org/abs/2302.11382

Zhan, X., Xu, Y., & Sarkadi, S. (2023). Deceptive AI Ecosystems: The Case of ChatGPT. Proceedings of the 5th International Conference on Conversational User Interfaces, 1– 6. https://doi.org/10.1145/3571884.3603754

Article DOI: https://doi.org/10.37284/eajit.7.1.2111

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models (arXiv:2309.01219). arXiv. http://arxiv.org/abs/2309.01219

Zhou, Z., Li, L., Chen, X., & Li, A. (2023). *Mini-Giants: "Small" Language Models and Open Source Win-Win* (arXiv:2307.08189). arXiv. http://arxiv.org/abs/2307.08189