Mitigating Hallucinations in Large Language Models via Semantic Enrichment of Prompts: Insights from BioBERT and Ontological Integration

Stanislav Penkov

Sofia University “St. Kliment Ohridski”

spenkov101@gmail.com

Abstract

The advent of Large Language Models (LLMs) has been transformative for natural language processing, yet their tendency to produce "hallucinations"—outputs that are factually incorrect or entirely fabricated—remains a significant hurdle. This paper introduces a proactive methodology for reducing hallucinations by strategically enriching LLM prompts. This involves identifying key entities and contextual cues from varied domains and integrating this information into the LLM prompts to guide the model towards more accurate and relevant responses. Leveraging examples from BioBERT for biomedical entity recognition and ChEBI for chemical ontology, we illustrate a broader approach that encompasses semantic prompt enrichment as a versatile tool for enhancing LLM output accuracy. By examining the potential of semantic and ontological enrichment in diverse contexts, we aim to present a scalable strategy for improving the reliability of AI-generated content, thereby contributing to the ongoing efforts to refine LLMs for a wide range of applications.

Keywords: Large Language Models (LLMs), Semantic Prompt Enrichment, Hallucination Mitigation, Domain-Specific Ontologies, BioBERT Entity Recognition

1. Introduction

Large Language Models (LLMs) have revolutionized numerous sectors by enabling machines to parse, understand, and generate human-like text, thus becoming cornerstones of modern AI applications. However, despite their sophistication, LLMs face a critical challenge that threatens their reliability and ethical application: the generation of "hallucinations"— “the creation of factually erroneous information spanning a multitude of subjects” (Rawte et al., 2023). This issue, while technical, carries profound ethical implications, particularly when these models are deployed in information-sensitive areas such as healthcare, law, or education (Martino et al., 2023; Rawte et al., 2023; Feldman et al., 2023; Wan et al., 2024; Peng et al., 2023). It raises questions about the trustworthiness and dependability of AI-generated information, emphasizing the need for a solution that ensures LLM outputs are not just coherent but also factually accurate.

To address this, our paper introduces a novel, comprehensive methodology that pre-emptively mitigates these hallucinations by integrating domain-specific knowledge and semantic information directly into LLM prompts. This approach indirectly complements existing techniques like Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) and prompt engineering. Drawing on tools like BioBERT[1] for biomedical entity recognition and the ChEBI [2](Chemical Entities of Biological Interest) database for chemical ontologies, we propose a versatile approach to enhancing the accuracy of LLM outputs. While the current submission focuses on the theoretical framework and initial findings, we outline a detailed plan for future empirical evaluation to rigorously validate our methodology and provide concrete evidence of its effectiveness.

[1] GitHub - dmis-lab/biobert: Bioinformatics'2020: BioBERT: a pre-trained biomedical language representation model for biomedical text mining

[2] Chemical Entities of Biological Interest (ChEBI)

2. Background

The versatility of general-purpose LLMs in handling tasks across various languages and domains introduces significant complexities in effectively evaluating and mitigating hallucinations (Zhang et al., 2023). These models, while robust, often display limitations in dynamic or culturally diverse contexts where nuances of language and factual accuracy are paramount. Current strategies such as dataset curation (Wan et al., 2024), model fine-tuning (Martino et al., 2023), and the integration of external knowledge bases (Peng et al., 2023) typically address these issues only after inaccuracies have been identified, which is not sufficient in preventing the initial occurrence of hallucinations.

Our methodology addresses these critical gaps by leveraging cutting-edge tools like BioBERT (Lee et al., 2019) and ChEBI, which anchor LLM outputs in verifiable facts, thereby not only reacting to but pre-emptively correcting potential inaccuracies. By embedding deeper semantic understanding directly into LLM prompts, our approach extends the ethos behind RAG, enhancing its capability to improve LLM reliability. This proactive integration, aligned with the principles demonstrated by Brown et al. (2020) in “Language Models are Few-Shot Learners”, sets a new benchmark for developing robust and ethical AI systems. The innovative approach ensures that LLMs operate within ethical bounds, enhancing their reliability across diverse applications and reducing the risk of misinforming users.

3. Proposed Methodology

To pre-emptively address the challenge of hallucinations in LLM outputs, our methodology employs a multi-layered approach, uniquely combining BioBERT for entity recognition, ChEBI for structured chemical ontologies, and direct API interactions with LLMs to guide the generation of accurate and relevant textual responses.

3.1 Technical Setup and Frameworks

3.1.1 BioBERT[3]: In our methodology, BioBERT is not merely used for its strong biomedical entity recognition capabilities. Instead, it serves a crucial role in the initial phase of our semantic enrichment process, where it precisely identifies and categorizes biomedical entities within LLM prompts. This specificity is vital as it ensures that the subsequent enrichment steps are accurately informed, targeting the most relevant semantic contexts required for each prompt. This targeted recognition is critical for minimizing errors in the generated text, especially in complex scenarios involving medical terminologies and contexts.

3.1.2 libChEBI API[4]: Similarly, the ChEBI database is utilized through the libChEBI API not just as a repository of chemical ontologies but as an integral component of our semantic enrichment framework. By programmatically accessing detailed ontological data, we enrich LLM prompts with deep semantic information that comprehensively describes chemical entities and their interactions. This enrichment goes beyond basic ontology integration by dynamically adjusting the context of the prompts to reflect current and precise chemical knowledge, significantly reducing the likelihood of generating inaccurate chemical data in LLM outputs.

3.1.3 LLM API: Interaction with LLMs, such as OpenAI's[5] GPT models, is achieved through their respective APIs.

3.2 Step-by-Step Implementation

3.2.1 Setting Up the Environment: The implementation begins with configuring the environment for BioBERT and setting up the API interactions with ChEBI and LLMs. This preparation ensures that all components are ready for the subsequent steps of entity recognition and prompt enrichment.

3.2.2 Pre-processing and Entity Recognition with BioBERT: Using BioBERT, we pre-process and analyze the input text to identify key biomedical entities. This identification is crucial for determining the specific entities around which the prompt enrichment revolves.

[3] https://huggingface.co/dmis-lab/biobert-v1.1

[4] https://github.com/libChEBI/libChEBIpy

[5] OpenAI API | OpenAI

3.2.3 Retrieving Ontological Data from ChEBI: Once entities are identified, we query the ChEBI database to fetch ontological information related to these entities. Such information provides a layer of semantic depth to the prompt, grounding the LLM's output in factual accuracy.

3.2.4 Enriching LLM Prompts: The core of our methodology lies in the enrichment of LLM prompts. By integrating the entity recognition from BioBERT with the ontological data from ChEBI, we craft prompts that are rich in domain-specific knowledge and context. This integration is pivotal for guiding the LLM towards generating responses that are not only relevant but also anchored in factual accuracy.

3.2.5 Generating Responses with LLM API: With enriched prompts in hand, we interact with the LLM via its API, feeding it the prompts and analyzing the generated responses. This phase tests the effectiveness of our semantic prompt enrichment, with the expectation that the model's outputs will exhibit a marked reduction in hallucinations.

3.3 Planned Empirical Evaluation

Due to time and resource constraints, we were unable to conduct a full empirical evaluation for this submission. However, we outline our planned evaluation framework to be conducted in future work to rigorously validate our methodology.

3.3.1 Data Collection: We plan to collect a diverse dataset of biomedical queries and responses to ensure comprehensive evaluation.

3.3.2 Experimental Design: Our future work will implement a controlled study comparing the outputs of LLMs using original versus enriched prompts. This will help us understand the impact of semantic enrichment on reducing hallucinations.

3.3.3 Expert Review: Domain experts will be involved to review and score the responses based on accuracy and relevance. This qualitative assessment will provide valuable insights into the effectiveness of our approach.

3.3.4 Statistical Analysis: We will analyze the collected data to determine the statistical significance of improvements in accuracy and reduction in hallucinations. Metrics such as precision, recall, and F1-score will be used to quantify the benefits of our methodology.

By conducting this empirical study in future work, we aim to provide concrete evidence of the effectiveness of our methodology. This planned evaluation will not only validate our approach but also contribute valuable insights to the ongoing efforts to enhance the reliability of AI-generated content.

3.4 Integration Flow Example

Example: Describing the interaction between Aspirin and blood pressure.

Input Text: “Describe the interaction between Aspirin and blood pressure.”

1. BioBERT identifies entities

- Output: “Aspirin”, “blood pressure”

2. Retrieve ontological data from ChEBI

- Output: “Aspirin (acetylsalicylic acid, C9H8O4)”

3. Enrich LLM prompt

- Output: “Describe the interaction between Aspirin (acetylsalicylic acid, C9H8O4) and blood pressure, considering its properties as an anti-inflammatory agent and its effects on blood coagulation.”

4. Generate LLM response

- Output: Detailed and accurate response based on enriched prompt

The integration flow example illustrates the practical application of combining BioBERT for biomedical entity recognition with ChEBI for chemical ontology, facilitated by direct interactions with LLM APIs. This strategic combination not only enhances the input to the models but also guides them towards generating outputs that are both accurate and contextually relevant. By embedding such enriched prompts, we pave the way for the development of LLMs capable of producing more accurate and trustworthy AI-generated content across a broad spectrum of domains. Our methodology demonstrates a scalable and versatile approach to improving the fidelity of LLM outputs, addressing the critical challenge of hallucinations and setting a new standard for the reliability of AI in sensitive and information-intensive fields.

4. Pre-emptive Implementation and Comparative Analysis

A pivotal innovation of our approach is the introduction of a pre-emptive layer of semantic depth, a concept designed to mitigate potential hallucinations even before they occur. Unlike traditional methods that react to inaccuracies post-generation, our methodology proactively incorporates domain-specific ontologies and entity recognition into the enrichment of LLM prompts. This pre-emptive strategy is fundamental in setting our work apart from existing techniques such as the Retrieval-Augmented Generation (RAG) system highlighted by Kang et al. (2023). While RAG enhances LLM responses by integrating external knowledge before generation, our approach extends these capabilities by embedding a deeper layer of semantic understanding, offering a novel solution to the challenge of hallucinations in LLM outputs.

Our technical framework aligns with broader hallucination mitigation tools, drawing parallels with initiatives like LLM-Augmenter and FreshPrompt (Peng et al., 2023; Vu et al., 2023). By incorporating BioBERT and ChEBI for structured prompt enrichment, we not only align with but also advance the principles of augmenting LLMs with external knowledge and feedback mechanisms. This theoretical exploration sets a new benchmark for mitigating hallucinations and opens avenues for future research to explore pre-emptive measures over reactive ones.

5. Discussion and Broader Implications

Our proactive methodology not only shifts the paradigm of semantic enhancements (Zhang et al., 2023) but also sets a new standard for reliability in AI-generated content. By incorporating empirical validations across various domains, we aim to substantiate our methodology's robustness. Exploring additional domain-specific ontologies will refine and expand our approach, enhancing the adaptability of LLMs. This is crucial as it allows LLMs to operate effectively within diverse fields, pushing the boundaries of current AI capabilities. Future research will customize and enhance the semantic accuracy of LLM outputs, ensuring AI-generated content is reliable and contextually appropriate across disciplines. This holistic approach mitigates the risks associated with hallucinations and advances the development of ethically sound and universally dependable AI systems.

6. Conclusion

This paper details an innovative, proactive methodology designed to tackle the challenge of hallucinations in LLM outputs from the ground up, marking a significant stride towards the ethical deployment and operational reliability of these advanced models. While our current findings are based on theoretical frameworks, we have outlined a comprehensive plan for empirical evaluation in future work. This planned study will rigorously validate our approach, providing the concrete evidence needed to establish the effectiveness of semantic prompt enrichment in improving LLM output accuracy and reliability. By establishing this new standard for the development of AI systems, we underscore our methodology's potential to significantly reduce hallucinations and enhance the factual integrity of AI-generated content.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Feldman, P., Foulds, J. R., & Pan, S. (2023). Trapping LLM Hallucinations Using Tagged Context Prompts. https://doi.org/10.48550/arXiv.2306.06085

Kang, H., Ni, J., & Yao, H. (2024). Ever: Mitigating Hallucination in Large Language Models through

Real-Time Verification and Rectification. https://arxiv.org/abs/2311.09114

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://arxiv.org/abs/2005.11401

Martino, A., Iannelli, M., & Truong, C. (2023). Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In Proceedings of the European Semantic Web Conference (ESWC). Yext New York NY.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. https://doi.org/10.48550/arXiv.2302.12813

Rawte, V., Priya, P., Islam Tonmoy, S. M., Zaman, S. M. M., Sheth, A., & Das, A. (2023). Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability Formality and Concreteness. AI Institute University of South Carolina USA. https://doi.org/10.48550/arXiv.2309.11064

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., & Luong, T. (2023). FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. https://doi.org/10.48550/arXiv.2310.03214

Wan, F., Huang, X., Cui, L., Quan, X., Bi, W., & Shi, S. (2024). Mitigating Hallucinations of Large Language Models via Knowledge Consistent Alignment. https://doi.org/10.48550/arXiv.2401.10768

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. https://doi.org/10.48550/arXiv.2309.01219