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

Stanislav Penkov, Sofia University "St. Kliment Ohridski"

INTRODUCTION

Large Language Models (LLMs) have revolutionized text generation and understanding, becoming cornerstones in fields like healthcare and law. However, they often produce 'hallucinations'—outputs that are factually incorrect or fabricated. These hallucinations pose a significant challenge to the reliability and ethical application of AI systems. This research introduces a proactive methodology to reduce hallucinations by enriching LLM prompts with domain-specific semantic information, guiding models toward more accurate outputs.

BACKGROUND

LLM hallucinations are problematic as they undermine the reliability of AI systems, especially in sensitive areas like healthcare, law, and education. Existing methods like Retrieval-Augmented Generation (RAG) and fine-tuning address inaccuracies only after they occur. A more proactive approach is needed to ensure factually accurate outputs from the outset.

PROPOSED METHODOLOGY

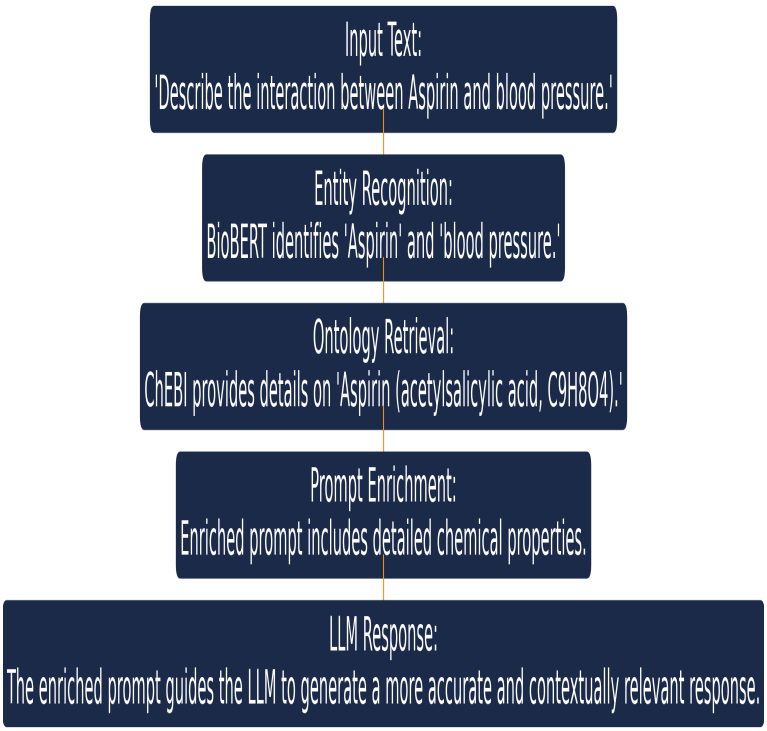
- BioBERT for Entity Recognition: BioBERT is used to identify and categorize biomedical entities within prompts, ensuring relevance and accuracy of LLM responses.
- ChEBI Ontologies for Semantic Enrichment: The ChEBI database provides chemical ontology data to enrich prompts with deep semantic information, improving the factual accuracy of the LLM output.
- LLM API Integration: Enriched prompts are integrated with LLM APIs (e.g., OpenAI's GPT models) to generate more accurate and contextually relevant outputs.

IMPLEMENTATION STEPS

- Setup and Pre-processing: Configure BioBERT and ChEBI API, and set up the LLM environment for prompt enrichment.
- Entity Recognition and Ontology Retrieval: Use BioBERT to identify key entities in the input text and retrieve related ontological data from ChEBI.
- Prompt Enrichment and Response Generation: Enrich prompts with semantic information from BioBERT and ChEBI, then use LLM APIs to generate responses.

INTEGRATION FLOW EXAMPLE

An example of applying the methodology to understand the interaction between Aspirin and blood pressure:



PLANNED EMPIRICAL EVALUATION

- Data Collection: Diverse biomedical datasets will be collected to test the approach.
- Experimental Design: Compare LLM outputs using original versus enriched prompts to measure the impact of semantic enrichment.
- Expert Review and Statistical Analysis: Evaluate accuracy using domain experts and metrics like precision, recall, and F1-score.

DISCUSSION AND BROADER IMPLICATIONS

The proposed methodology improves LLM reliability by reducing hallucinations, making AI outputs more trustworthy in information-sensitive applications. Future research will explore additional domain-specific ontologies and refine the approach for broader adaptability across various fields.

CONCLUSION

This research introduces a proactive method to mitigate hallucinations in Large Language Models (LLMs) by enriching prompts with domain-specific semantic information. By combining BioBERT for precise entity recognition and ChEBI for ontology-based semantic enrichment, the approach ensures that LLMs generate outputs that are both accurate and contextually relevant.

The integration of enriched prompts with LLM APIs, such as OpenAI's GPT models, significantly enhances the reliability of AI-generated responses, particularly in fields like healthcare and law, where factual accuracy is crucial. This methodology provides a scalable framework for improving LLM performance and minimizing the risks of misinformation in sensitive applications.

Future work will focus on empirically validating the effectiveness of this approach across diverse domains and refining the integration of additional domain-specific ontologies to further improve the factual integrity of AI outputs.

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

Contact: spenkov101@gmail.com