Facts Matter — Investigating Domain Knowledge in Large Language Models

Master Project Summer 2024 – AI and Intelligent Systems Group

Abstract

Large Language Models (LLMs) and Natural Language Processing (NLP) have seen significant advancements since ChatGPT and similar tools emerged in 2022. LLMs excel in Natural Language Generation (NLG) due to their ability to generate text and react to textual and multimodal input. They are particularly effective in creative situations like writing stories, brainstorming, and casual chats.
However, when reproducing factually correct text, particularly in domain-specific knowledge problems, these models tend to "hallucinate", i.e., to produce text that conveys information that may sound convincing but is incorrect. Despite the immense progress in this field of research, several unanswered questions remain in this area.
In this project, we aim to delve deeply into the factuality of LLMs, particularly in specialised knowledge domains. Our focus is to explore how well these models handle domain-specific information across various fields such as science, medicine, engineering, etc. We also plan to investigate the limitations of LLMs in handling up-to-date and specialised domain knowledge, assess the impact of training data on their domain expertise, and explore approaches to enhance their accuracy and reliability in those areas. Moreover, we will examine the challenges in evaluating the factuality of LLMs, especially considering the dynamic nature of information in the real world. This project intends to contribute to the ongoing discourse on improving the factuality of LLMs in diverse fields.

Research Questions

During this project, we aim to set up and conduct various LLM experiments along several of the following research topics:

1. **Evaluation of LLM factuality**
    a. How well do LLMs perform across different knowledge domains (e.g., science, history, medicine)?
    b. How comprehensive and effective are current benchmarks in evaluating the domain-specific factuality of LLMs?
    c. Can we use larger LLMs such as GPT-4 to evaluate LLMs across domains and specifically their factual knowledge?

2. **Developing factual LLMs**
    a. How effective are different techniques to provide LLMs with knowledge (e.g., pre-training, supervised fine-tuning, alignment with RL, augmentation with RAG)?
    b. How comprehensive and effective are current benchmarks in evaluating the domain-specific factuality of LLMs?
    c. How can we make LLMs better understand specific domains?
3. **Further questions (optional)**:
    a. What are the dangers of LLMs making mistakes in critical, knowledge-dependent applications?
    b. How do the different LLMs' writing styles differ and how does this difference affect their evaluation? E.g., with the same level of factuality, are shorter or longer / simpler or more complex texts rated as better?
    c. How do LLMs deal with domain-specific knowledge across multiple languages? (this would focus on Wikipedia due to data availability)
    d. How do LLMs deal with new and changing information?

Project Outline

This project will consist of multiple work packages that will be worked on during the semester.

1. **Introductory Meeting**: Initial discussions on research approaches, available data and relevant literature
2. **Literature Review**: Review the references provided by the supervisors and extend them if required. We emphasise scientific work to ensure our project is grounded in relevant literature.
3. **Dataset curation**: We will provide multiple datasets that need to be cleaned and filtered to curate high-quality datasets. These include Reddit data from relevant subreddits, a Wikipedia dataset, and potentially other domain-specific data sources.
4. **Factuality benchmark development**: Assess existing benchmarks and their limitations, and develop a new, complementary evaluation approach across the selected domains and different evaluation metrics.
5. **LLM Experiments**: Design, train and test LLM models to apply evaluation techniques. First with pre-trained models out-of-the-box, then with models that are fine-tuned/aligned/augmented with external knowledge by the project team.
6. **Analysis and Results Evaluation**: Evaluate the LLM test results from the benchmarks, discuss results and iterate experiments if required.
7. **Discussion and Conclusion**: Conclude the project with a discussion of the key findings and results, and a summary of the contributions and achievements of the project. We aim to publish the results of this work at a top-tier conference.

This is a general outline, and the actual project approach may be adjusted depending on the project team's progress and intermediate results.

Related Work

Evaluation:

- Huang, Lei et al. “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.” *ArXiv abs/2311.05232 (2023).* https://arxiv.org/pdf/2305.11747.pdf
- Chen, Shiqi et al. “FELM: Benchmarking Factuality Evaluation of Large Language Models.” *ArXiv* abs/2310.00741 (2023). https://arxiv.org/abs/2310.00741
- Liu, Yang, et al. "G-Eval: NLG evaluation using GPT-4 with better human alignment." *arXiv preprint arXiv:2303.16634 6* (2023). https://arxiv.org/pdf/2303.16634.pdf
- Lin, Stephanie C. et al. “TruthfulQA: Measuring How Models Mimic Human Falsehoods.” *Annual Meeting of the Association for Computational Linguistics* (2021). https://aclanthology.org/2022.acl-long.229/
- Zheng, Lianmin, et al. "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena." *arXiv preprint arXiv:2306.05685* (2023). https://arxiv.org/pdf/2306.05685
- Sathe, Aalok, et al. "Automated fact-checking of claims from Wikipedia." *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020. https://aclanthology.org/2020.lrec-1.849/

Fine-tuning, alignment & augmentation:

- Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *arXiv preprint arXiv:2305.18290* (2023). https://arxiv.org/abs/2305.18290
- Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
- Pride, David, Matteo Cancellieri, and Petr Knoth. "CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering." *International Conference on Theory and Practice of Digital Libraries*. 2023. https://arxiv.org/abs/2307.04683
- Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." *arXiv preprint arXiv:2306.08302 (2023).* https://arxiv.org/pdf/2306.08302.pdf
- Su, Weihang, et al. "Wikiformer: Pre-training with Structured Information of Wikipedia for Ad-hoc Retrieval." *arXiv preprint arXiv:2312.10661* (2023). https://arxiv.org/pdf/2312.10661.pdf

LLM hallucination:

- Xu, Silei, et al. "Fine-tuned llms know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over wikidata." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023. https://arxiv.org/pdf/2305.14202.pdf
- Semnani, Sina, et al. "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia." *Findings of the Association for*

Computational Linguistics: EMNLP 2023. 2023. https://aclanthology.org/2023.findings-emnlp.157.pdf
- Sclar, Melanie, et al. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." *arXiv preprint arXiv:2310.11324 (2023).* https://arxiv.org/abs/2310.11324

Prior knowledge

Participants need to have experience with PyTorch and training and testing machine learning models. Ideally, you would also have experience with NLP use cases such as training LLMs or BERT-based models and with Huggingface, and Docker.

Contact

gerard.demelo@hpi.de, lucie-aimee.kaffee@hpi.de, russa.biswas@hpi.de, yindong.wang@hpi.de, tolga.buz@hpi.de, konstantin.dobler@hpi.de, margarita.bugueno@hpi.de
This project will be supervised by the HPI Chair for Artificial Intelligence and Intelligent Systems (Prof. Dr Gerard de Melo, Dr Lucie-Aimée Kaffee, Dr Russa Biswas, Yindong Wang, Tolga Buz, Konstantin Dobler, Margarita Bugueño)