LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Exploring Strategies to Defend Editing Attacks in Large Language Models**

Supervisor: Mingyang Wang

Examiner: Hinrich Schütze

Level: BSc / MSc

Description: Knowledge editing has demonstrated significant potential in correcting false or outdated information within Large Language Models (LLMs). However, this capability also presents a significant vulnerability: editing attacks. Such attacks involve injecting malicious or biased information into LLMs, often integrated alongside factual updates, posing risks to the model's integrity and trustworthiness.

This project aims to investigate whether LLMs possess an inherent ability to distinguish between legitimate factual modifications and injected biased or false information through in-context knowledge editing. Understanding this capability could pave the way for developing defenses against editing attacks by optimizing knowledge selection mechanisms to resist misinformation and ensure the reliability of updated knowledge.

The research builds on recent studies that highlight the risks posed by editing attacks and delve into the internal processes of LLMs when managing conflicting or inconsistent knowledge. By exploring these aspects, the project seeks to contribute to the development of more robust and secure LLMs.

Requirements:

- Enthusiasm in NLP research
- Fluent in English
- Experience with programming in Python
- Basic knowledge of transformers-based language models
- Ideally knowledge of language model interpretability

References:

- Canyu Chen et al. (n.d.). "Can Editing LLMs Inject Harm?" In: *ICML 2024 Next Generation of AI Safety Workshop*. URL: https://arxiv.org/abs/2407.20224
- Yu Zhao et al. (2024). "Analysing the Residual Stream of Language Models Under Knowledge Conflicts". In: *arXiv preprint arXiv:2410.16090*. URL: https://arxiv.org/abs/2410.16090
- Andy Arditi et al. (2024). "Refusal in language models is mediated by a single direction". In: *arXiv preprint arXiv:2406.11717*. URL: https://arxiv.org/abs/2406.11717