LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Investigating the Impact of Knowledge Editing on Large Language Models**

Supervisor: Mingyang Wang

Examiner: Hinrich Schütze

Level: MSc

Description: Large language models (LLMs) have been shown to implicitly store vast amounts of factual knowledge within their parameters. However, as the world evolves, some of this knowledge may become outdated or incorrect. Recently, knowledge editing has emerged as a promising way to update or correct specific factual information in LLMs without retraining the entire model.
However, the impact of various knowledge editing methods on the internal mechanisms of language models remains largely unexplored. A deeper understanding how different editing techniques affect the underlying recall mechanisms in transformer models could lead to more effective methods for knowledge editing and offer valuable insights for better controlling LLM behavior.
In summary, this project aims to investigate how different knowledge editing approaches affect the internal factual recall mechanisms in transformer-based language models and to identify potential strategies for optimizing these editing techniques.

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:

- Yunzhi Yao et al. (2023). "Editing Large Language Models: Problems, Methods, and Opportunities". In: URL: https://aclanthology.org/2023.emnlp-main.632
- Javier Ferrando and Elena Voita (2024). "Information flow routes: Automatically interpreting language models at scale". In: URL: https://arxiv.org/abs/2403.00824