LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Multilingual Factual and Linguistic Knowledge Probing**

Supervisor: Ahmad Dawar Hakimi

Examiner: Prof. Hinrich Schütze

Level: BSc / MSc

Summary: The goal of this thesis is to create a multilingual dataset for evaluating language models' abilities to handle factual knowledge and linguistic tasks. The dataset will include factual relationships across diverse domains and linguistic challenges such as identifying antonyms, comparatives, superlatives, verb tenses, and word patterns like first- or last-letter similarities (Hernandez et al. 2023). By incorporating both high-resource and low-resource languages, the dataset will enable a detailed investigation into how models manage diverse linguistic structures and cultural nuances.

The evaluation will focus on zero-shot and few-shot learning scenarios, examining how well language models adapt to new languages or tasks with minimal data. In particular, the thesis will explore how few-shot learning improves a model's ability to solve specific tasks (Li et al. 2024). Additionally, the thesis will address tokenization challenges, such as errors in processing compound words and complex morphological structures, and their impact on model performance.

Supervision can be provided in either German or English.

Requirements:

- Enthusiasm in NLP research
- Knowledge of Python & Transformer-based language models
- Ideally knowledge of language model interpretability

References:

- Evan Hernandez et al. (2023). "Linearity of relation decoding in transformer language models". In: *arXiv preprint arXiv:2308.09124*
- Daoyang Li et al. (2024). "Exploring multilingual probing in large language models: A cross-language analysis". In: *arXiv preprint arXiv:2409.14459*