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

Thesis proposal

Topic: **Language-Specific Concepts in LLMs**

Supervisor: Yihong Liu

Examiner: Prof. Hinrich Schütze

Level: MSc

Summary: Some recent studies show that LLMs, e.g., Llama, tend to "translate" the input text into a specific latent language, e.g., English, and "think" in that latent language in the middle layers (we could roughly classify the layers in LLMs to early layers, middle layers, and final layers based on the ordered layer sequence) (Wendler et al. 2024; Zhong et al. 2024). The concepts in the input language are therefore always represented in the latent language. For example, the hypothesis of how LLMs perform in-context translation is as follows: concept and language are represented independently. When doing the translation, the model first detects the target language from the context, and then identifies the concept C. In the last layers, the model then maps C to tokens that correspond to the concept C in the target language (Dumas et al. 2024). However, there is an underlying assumption: the concept is language-irrelevant, or, there are always equivalent words in that latent language to represent the concept. In this project, we aim to "challenge" this assumption. We want to collect language-specific concepts – the concepts that are almost unique in one specific language, or at least, the concepts that are hard to express in other languages. Then we move further to investigate how these concepts are represented in LLMs. Are they still be "mapped" to the latent language or they just remain what they are in the original language?

Requirements: enthusiasm, good mathematical background and programming background (preferably Python), good knowledge of NLP, a good command of DL framework (preferably PyTorch)

References:

- Chris Wendler et al. (Aug. 2024). "Do Llamas Work in English? On the Latent Language of Multilingual Transformers". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 15366–15394. DOI: 10.18653/v1/2024.acl-long.820. URL: https://aclanthology.org/2024.acl-long.820
- Chengzhi Zhong et al. (2024). "Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in?" In: *arXiv preprint arXiv:2408.10811*
- Clément Dumas et al. (2024). "How do Llamas process multilingual text? A latent exploration through activation patching". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=0ku2hIm4BS