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

Thesis proposal

Topic: **Identifying Relation Neurons with Synthetic Data**

Supervisor: Yihong Liu

Examiner: Hinrich Schütze

Level: MSc

Summary: Current methods identify relation-specific neurons by ranking them based on their association with relation-labeled data. However, these approaches can be confounded by entity occurrences: neurons may respond not only to the relation itself but also to familiar entities seen during pretraining. To eliminate this confound, we propose generating fully synthetic data that expresses the target relation while using entities guaranteed not to appear in the model's pretraining corpus. We will then "fake-train" the model on this synthetic data and select candidate relation-neurons based on their gradient magnitudes. Additionally, we can compare the relation-specific neurons identified by synthetic data with the neurons identified by real relational factual data.

Requirements: enthusiasm, good programming background (preferably Python), good knowledge of NLP, a good command of PyTorch and HuggingFace.

References:

- Damai Dai et al. (2022). *Knowledge Neurons in Pretrained Transformers*. arXiv: 2104.08696 [cs.CL]. URL: https://arxiv.org/abs/2104.08696
- Evan Hernandez et al. (2024). *Linearity of Relation Decoding in Transformer Language Models*. arXiv: 2308.09124 [cs.CL]. URL: https://arxiv.org/abs/2308.09124
- Yihong Liu et al. (2025). *On Relation-Specific Neurons in Large Language Models*. arXiv: 2502.17355 [cs.CL]. URL: https://arxiv.org/abs/2502.17355