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

Thesis proposal

Topic: **Faithfulness / plausibility of logit lens across models**

Supervisor: Sebastian Gerstner

Examiner: Hinrich Schütze

Level: MSc

Summary: When researchers try to interpret the hidden states of LLMs, they often resort to a simple but powerful technique called "logit lens" (nostalgebraist 2020): applying the unembedding matrix to the hidden state at hand (ignoring all the intermediate layers), thus translating a hidden state into scores for tokens. The result can often be interpreted as an intermediate guess about the next token (nostalgebraist 2020). However, logit lens is not equally plausible across models (nostalgebraist 2021; Belrose et al. 2023). In particular, Belrose et al. 2023 found that logit lens is usually only a biased predictor of the model's actual next-token prediction. Other authors have proposed more sophisticated interpretations of logit lens results: For example, Wendler et al. 2024 found intermediate representations that resemble English tokens while processing text in other languages (e.g. "door" when the next token should be German "Tür"); they interpreted them as representing the meaning of the next token independent of its language.
In your thesis you will investigate how and why this variation across models happens. In particular, you will:

- Review literature on logit lens and related methods.
- More closely examine results of nostalgebraist 2021; Belrose et al. 2023. Is the variation across models somehow predictable? For example, does it correlate with architecture details, model size, or multilinguality?
- If you do find such a correlation: can you find a plausible explanation for it?
- Possibly, extend the investigations on logit lens faithfulness to other models, such as instruction-tuned models, or different training checkpoints of a given Pythia or OLMo model. This step would require GPU resources.
- Possibly, come up with a measure for semantic plausibility (as opposed to mere faithfulness), that would account for the type of phenomena found by Wendler et al. 2024 and others. You can then measure this across models as well. This step would require GPU resources.

Requirements: Good command of Python; basic knowledge of the Transformer architecture.

References:

Belrose, Nora et al. (2023). *Eliciting latent predictions from transformers with the tuned lens*. URL: https://arxiv.org/pdf/2303.08112.

nostalgebraist (2020). *Interpreting GPT: The logit lens*. URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

— (2021). *logit lens on non-gpt2 models + extensions*. Colab. URL: https://colab.research.google.com/drive/1MjdfK2srcerLrAJDRaJQKO0sUiZ-hQtA?usp=sharing.

Wendler, Chris et al. (2024). "Do Llamas work in English? On the latent language of multilingual transformers". In: *arXiv*. URL: https://arxiv.org/pdf/2402.10588.