LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Effects of Pretraining on Bias in LLMs**

Supervisor: Leonor Veloso

Examiner: Prof. Hinrich Schütze

Level: MSc

Summary:

Efforts in the field of mechanistic interpretability have lead to insights on how factual knowledge and bias are internally represented in LLMs. A recent set of works has focused on mechanisms of knowledge acquisition – concretely, on the evolution of factual recall throughout continual pre-training. Ou et al. 2025; Zucchet et al. 2025 find that the evolution of knowledge circuits follows a deep-to-shallow pattern, where deeper layers extract the relevant information for the task and shallow layers enrich their knowledge representation with more specific knowledge. This is accompanied by evolutions in the role of model components, which become more specialized as training progresses (Hakimi et al. 2025). Previous work has hinted that bias may be represented in LLMs differently than facts – Kirsten et al. 2025 show that different inference acceleration techniques affect biased and factual predictions differently. This body of work begs the question: how do continual pretraining dynamics affect **socially biased** representations and associations in the model?
A pipeline for the development of this project can look like:

- Choosing a dimension of social bias to explore (dimensions that have been previously studied in mechanistic interpretability include gender bias, demographic bias, and racial bias);
- Identify bias-related circuits across checkpoints of a small open-weight LLMs (`OLMo-1b`) with an efficient circuit analysis method (such as Information Flow Routes, proposed by Ferrando and Voita 2024);
- Analyze the evolution of bias-related circuits throughout training with assessments of performance, topology, and the role of individual components.

Requirements:

- Enthusiasm!
- Good command of Python, Pytorch and HuggingFace's `transformers` library
- Solid knowledge of ML and NLP concepts (particularly the Transformer architecture)

References:

- Javier Ferrando and Elena Voita (2024). "Information flow routes: Automatically interpreting language models at scale". In: *arXiv preprint arXiv:2403.00824*
- Yixin Ou et al. (2025). "How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training". In: *arXiv preprint arXiv:2502.11196*
- Nicolas Zucchet et al. (2025). "How do language models learn facts? Dynamics, curricula and hallucinations". In: *arXiv preprint arXiv:2503.21676*
- Ahmad Dawar Hakimi et al. (2025). "Time Course MechInterp: Analyzing the Evolution of Components and Knowledge in Large Language Models". In: *arXiv preprint arXiv:2506.03434*
- Elisabeth Kirsten et al. (2025). "The impact of inference acceleration on bias of llms". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1834–1853