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

Thesis proposal

Topic: **Analyzing the processing of idiomatic phrases in pretrained (and finetuned) transformer models**

Supervisor: Lea Hirlimann

Examiner: Hinrich Schuetze

Level: MSc

Summary: Idiomatic phrases convey meanings that go beyond the literal interpretation of their individual words, posing unique challenges for language models. In recent years researchers have been able to trace different functionalities, knowledge and operators through the transformer architecture, linking them to distinct submodules, such as layers, attention heads, or neurons. Understanding how transformers process these idiomatic phrases advances their interpretability and might offer insights on cultural knowledge of the model. This thesis will include the sourcing of a idiomatic data and the creation of altered counterparts as input to a transformer model to analyze patterns across individual outputs and activations correlated to the valid idiom, which contribute to the association of the phrase with its non-literal meaning.
Following steps would include the visualization of individual phrase pairs and architectural activation patterns and a comprehensive comparison with existing knowledge on the functionalities within attention heads and layers.

Requirements: (recommended) good programming skills, enthusiasm to learn

References:

- Ye Tian, Isobel James, and Hye Son (2023). "How Are Idioms Processed Inside Transformer Language Models?" In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pp. 174–179
- Prateek Saxena and Soma Paul (2020). "Epie dataset: A corpus for possible idiomatic expressions". In: *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*. Springer, pp. 87–94
- nostalgebraist (n.d.). *interpreting GPT: the logit lens*. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, lesswrong. Accessed: 2025-01-08
- Kevin Ro Wang et al. (2023). "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small". In: *The Eleventh International Conference on Learning Representations*
- Amit Elhelo and Mor Geva (2024). "Inferring Functionality of Attention Heads from their Parameters". In: *arXiv preprint arXiv:2412.11965*
- Mor Geva et al. (2021). "Transformer Feed-Forward Layers Are Key-Value Memories". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495