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

Thesis proposal

Topic: **Aligning Natural Language Explanations with Token-level Attributions in Instruction-tuned Language Models**

Supervisor: Yuetian Lu

Examiner: Prof. Dr. Hinrich Schütze

Level: MSc

Summary: Instruction-tuned language models can produce natural language explanations (rationales, chain-of-thought style explanations), yet it remains unclear to what extent such explanations are faithful to the model's underlying decision process. This thesis will study the relationship between generated explanations and token-level attributions (e.g., Integrated Gradients and attention flow/rollout). The student will design and validate metrics that quantify *explanation–attribution consistency*, and benchmark them across several NLU tasks (e.g., NLI and sentiment analysis; optionally multilingual via XNLI). Building on these metrics, the thesis will explore lightweight interventions to improve faithfulness, such as explanation-based regularization during fine-tuning and constrained decoding strategies that bias explanations toward attributed evidence while maintaining task performance. The expected outcome is a well-documented experimental framework, thorough analysis of failure cases, and insights suitable for a research paper.

Requirements: Strong Python skills; good knowledge of deep learning for NLP; practical experience with PyTorch and Transformer/LLM toolchains (e.g., HuggingFace); ability to design careful evaluations and manage experiments; strong English writing skills; interest in interpretability and model evaluation.

References:

- Long Ouyang et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
- Jason Wei et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
- Tao Lei, Regina Barzilay, and Tommi Jaakkola (Nov. 2016). "Rationalizing Neural Predictions". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, pp. 107–117. DOI: 10.18653/v1/D16-1011. URL: https://aclanthology.org/D16-1011/
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan (Aug. 6–11, 2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html
- Samira Abnar and Willem Zuidema (July 2020). "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://aclanthology.org/2020.acl-main.385/
- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez (2017). "Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2662–2670. DOI: 10.24963/ijcai.2017/371. URL: https://doi.org/10.24963/ijcai.2017/371

- Chris Hokamp and Qun Liu (July 2017). "Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 1535–1546. DOI: 10.18653/v1/P17-1141. URL: https://aclanthology.org/P17-1141/
- Jay DeYoung et al. (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408/