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

Thesis proposal

Topic: **Query-Level Uncertainty for Ability-Aware Routing and Triage in Active Learning**

Supervisor: Ahmad Dawar Hakimi

Examiner: Prof. Hinrich Schütze

Level: MSc

Summary: Large Language Models (LLMs) are increasingly used as cost-effective annotators in text classification pipelines, yet their reliability varies widely across instances. Some examples can be labeled accurately by small models, while others require escalation to stronger models or human experts. Existing Active Learning (AL) methods focus on *what* to label, but largely overlook *who* should label each instance and at what cost. This thesis proposes an Ability-Aware Active Learning system that jointly optimizes three coupled decisions: **Acquisition:** which unlabeled examples maximize learning value; **Routing:** which oracle (small LLM, larger LLM, or human) should label each selected example; and **Adaptation:** how selective fine-tuning on human-routed examples can expand the small model's capabilities across AL rounds.

This work leverages Query-Level Uncertainty (QLU), specifically the Internal Confidence metric from Chen and Varoquaux 2025, which estimates whether an LLM can confidently handle a query before generating any tokens. Unlike existing triage systems (Jung et al. 2025, Rouzegar and Makrehchi 2024) that compute confidence after expensive generation, QLU operates on prefill only (parallel input processing), enabling pool-scale triage of unlabeled examples efficiently. The thesis integrates QLU into a unified framework that jointly optimizes acquisition, routing, and adaptation while maintaining calibration throughout the pipeline. The system is evaluated on anti-immigrant content classification in German TikTok political comments, a challenging domain with implicit rhetoric, dynamic, and multi-party political context.

Research questions:

1. How well does pre-generation Internal Confidence (QLU) correlate with classification accuracy compared to post-hoc uncertainty signals (output probability, consistency)?
2. Under a fixed budget (total human labels + LLM tokens), does ability-aware routing + QLU-based acquisition outperform: Random Sampling, Uncertainty Sampling with post-hoc signals, and Triage-only without active acquisition.
3. Does pre-generation Internal Confidence correlate with the quality and faithfulness of post-hoc explanations?

Supervision can be provided in either German or English.

Requirements:

- Strong interest in NLP, text classification, and reliable human-in-the-loop annotation
- Solid Python skills and experience with deep learning frameworks (PyTorch, Hugging Face Transformers)
- Basic familiarity with Active Learning concepts (or willingness to learn quickly)

References:

- Lihu Chen and Gaël Varoquaux (2025). "Query-Level Uncertainty in Large Language Models". In: *arXiv preprint arXiv:2506.09669*. URL: https://arxiv.org/abs/2506.09669

- Yu Xia et al. (July 2025). "From Selection to Generation: A Survey of LLM-based Active Learning". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 14552–14569. DOI: 10.18653/v1/2025.acl-long.708. URL: https://aclanthology.org/2025.acl-long.708/
- Hayoung Jung et al. (2025). "MythTriage: Scalable Detection of Opioid Use Disorder Myths on a Video-Sharing Platform". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. To appear
- Hamidreza Rouzegar and Masoud Makrehchi (Mar. 2024). "Enhancing Text Classification through LLM-Driven Active Learning and Human Annotation". In: *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*. St. Julians, Malta: Association for Computational Linguistics, pp. 98–111. URL: https://aclanthology.org/2024.law-1.10/
- Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza (2024). "Language Model-Driven Data Pruning Enables Efficient Active Learning". In: *arXiv preprint arXiv:2410.04275*. URL: https://arxiv.org/abs/2410.04275
- Markus Bayer, Justin Lutz, and Christian Reuter (2024). "ActiveLLM: Large Language Model-based Active Learning for Textual Few-Shot Scenarios". In: *arXiv preprint arXiv:2405.10808*. URL: https://arxiv.org/abs/2405.10808