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

Thesis proposal

Topic: **Empowering Minority Voices in Text Generation: Creating a Contrastive Dataset for Activation Steering**

Supervisor: Lea Hirlimann

Examiner: Hinrich Schuetze

Level: BSc / MSc

Summary: Large language models (LLMs) often produce fluent text but fail to equitably represent diverse cultural and minority perspectives. Prior work on cultural adaptability shows that models can exhibit biased or limited representation of cultural values and linguistic variation, highlighting the need for better evaluation and control methods in generative systems (Rao et al., 2024). **Activation steering** is an emerging technique that modifies a model's internal activations at inference time to push generation toward specific semantic directions, often identified through contrastive example pairs (e.g., positive vs. negative sentiment) to create steering vectors (Rimsky et al., 2024). While activation steering has been studied for bias mitigation and behavioral control, there is no existing datasets for finding such steering directions for minority narratives or inclusive representations.
The goal of this project will be to create a publicly available dataset and an evaluation framework that supports more **inclusive generative storytelling**. The contrastive data and derived steering vectors could serve as tools for both research and practical applications where equitable representation of linguistic and cultural diversity is critical.
The individual components of this project are:

- Build a **contrastive dataset of narratives** written from the perspectives of one or more underrepresented or minority groups of your choice. Each contrastive pair will consist of a majority-centric narrative and a corresponding narrative that foregrounds minority experience, values, or linguistic traits.
- (MSc) Use these contrastive pairs to identify **latent activation directions** within a target LLM that correlate with minority-oriented generation. Such steering vectors could then be used to bias generation toward diverse storytelling outputs without full model fine-tuning.
- Carefully assess the **quality** and **composition** of the dataset and if applicable the steered outputs.

Requirements: Motivation and Curiosity

References:

- Abhinav Rao et al. "Normad: A framework for measuring the cultural adaptability of large language models". In: *arXiv preprint arXiv:2404.12464* (2024)
- Nina Rimsky et al. "Steering llama 2 via contrastive activation addition". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 15504–15522