LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Morality-related directions in the activation space of multilingual models**

Supervisor: Leonor Veloso

Examiner: Prof. Hinrich Schütze

Level: MSc

Summary: Recent works have studied the linear representation of model features across a broad spectrum of concepts such as truth, humor, and factual knowledge. Notably, Arditi et al. (2024) use a set of contrastive pairs of harmful vs harmless instructions to identify and steer a single direction that mediates refusal of user requests in LLMs. Steering techniques have also been applied to the moral values of a model (Tlaie 2024) Additionally, it has been shown that multilingual models contain language-dependent moral variability (Aksoy 2024). Having these works in mind, this project raises the question: **how does morality-steering impact multilingual models?**
The project would entail:

- Reviewing frameworks for defining moral values (as described by Graham et al. (2008), for example) and their application to current NLP research;
- Applying existing methods for steering directions in the activation space (such as the one introduced by Arditi et al. (2024)) to multilingual models;
- Analyzing the moral framework of multilingual models pre and post steering.

Requirements:

- Enthusiasm!
- Good command of Python, Pytorch and HuggingFace's `transformers` library
- Basic knowledge of ML and NLP concepts (particularly modern LLM architectures, such as the Transformer architecture)

References:

- Jesse Graham et al. (2008). "Moral foundations questionnaire". In: *Journal of Personality and Social Psychology*
- Andy Arditi et al. (2024). "Refusal in language models is mediated by a single direction". In: *arXiv preprint arXiv:2406.11717*
- Meltem Aksoy (2024). "Whose Morality Do They Speak? Unraveling Cultural Bias in Multilingual Language Models". In: *arXiv preprint arXiv:2412.18863*
- Alejandro Tlaie (2024). "Exploring and steering the moral compass of Large Language Models". In: *arXiv preprint arXiv:2405.17345*
- Shaoyang Xu et al. (2024). "Exploring Multilingual Human Value Concepts in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?" In: *arXiv preprint arXiv:2402.18120*