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

Thesis proposal

Topic: **Investigating Media Bias in Language Models**

Supervisor: Molly Kennedy

Examiner: Hinrich Schütze

Level: MSc

Summary: As large language models (LLMs) become integral to applications that interact with diverse user groups, addressing biases that may distort factual representation or perpetuate stereotypes is crucial Bender et al. 2021. The attached paper: here predominantly explores biases linked to gender or occupation Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017, less attention has been given to other types of bias such as media bias—how language models may reflect or amplify the political, ideological, or cultural slants present in their training data.

Aims:

- Collect datasets of various bias types and use their labels as a Ground truth for what the conclusion of the explanation is Kulshrestha et al. 2017.
- Investigate the correlation between media bias in pre-training datasets and its influence on LLM outputs Garg et al. 2018.
- Examine how factors like political ideologies, regional biases, and source credibility affect model responses Gentzkow and Shapiro 2010.
- Employ relevant input-based explanation methods (i.e. which tokens are responsible for producing a certain output in a model) from XAI methods and tools Lipton 2018.
- Evaluate and Mitigate Bias in Model Development: Investigate potential biases that may emerge during the model creation process and propose methods to mitigate them, ensuring the robustness and fairness of the final model. For example, having an equal number of samples for each collective group/viewpoint.
- Build on prior research that has assessed bias in generative models using encoder models.
- Identify prompt features that contribute to or mitigate bias Jiang et al. 2019.

Requirements: Programming skills in Python, Machine Learning Basics, Natural Language Processing (NLP) knowledge, Data Handling (Pandas, Numpy), Bias Measurement Techniques, Data Science and Statistics, Knowledge of Explainable AI (XAI) and Model Interpretability methods.

References:

- Tolga Bolukbasi et al. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* 29
- Emily M Bender et al. (2021). "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186
- Juhi Kulshrestha et al. (2017). "Quantifying search bias: Investigating sources of bias for political searches in social media". In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 417–432

- Nikhil Garg et al. (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644
- Matthew Gentzkow and Jesse M Shapiro (2010). "What drives media slant? Evidence from US daily newspapers". In: *Econometrica* 78.1, pp. 35–71
- Zachary C Lipton (2018). "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57
- Haoming Jiang et al. (2019). "Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization". In: *arXiv preprint arXiv:1911.03437*