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

Thesis proposal

Topic: **Survey on Sparse Autoencoders: Techniques, Applications, and Future Directions**

Supervisor: Ahmad Dawar Hakimi

Examiner: Prof. Hinrich Schütze

Level: BSc / MSc

Summary: Sparse Autoencoders (SAEs) (Ng et al. 2011) are neural networks designed to learn compact, meaningful representations of data by enforcing sparsity in the hidden units, ensuring that most remain inactive for any given input. This sparsity reduces polysemanticity (Bricken et al. 2023), a challenge where a single neuron encodes multiple unrelated features, resulting in more interpretable representations focused on essential patterns. This property makes SAEs valuable for applications such as dimensionality reduction, anomaly detection, and feature extraction.
This thesis aims to conduct a comprehensive survey on the current state of Sparse Autoencoders, including their tools, evaluation metrics, use cases, comparison to different methods, and a summary of current challenges and future directions.

Supervision can be provided in either German or English.

Requirements:

- Enthusiasm in NLP research
- Knowledge of Transformer-based language models
- Ideally knowledge of language model interpretability

References:

- Andrew Ng et al. (2011). "Sparse autoencoder". In: *CS294A Lecture notes* 72.2011, pp. 1–19
- T. Bricken et al. (2023). *Towards monosemanticity: Decomposing language models with dictionary learning*. Transformer Circuits Thread. URL: https://transformer-circuits.pub/2023/monosemantic-features/index.html
- Tom Lieberum et al. (2024). "Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2". In: *arXiv preprint arXiv:2408.05147*
- Adly Templeton (2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic
- Leo Gao et al. (2024). "Scaling and evaluating sparse autoencoders". In: *arXiv preprint arXiv:2406.04093*
- Hoagy Cunningham et al. (2023). "Sparse autoencoders find highly interpretable features in language models". In: *arXiv preprint arXiv:2309.08600*