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

# Thesis proposal

**Topic:** **A Prompt Stress-Test Suite for LLM Media-Bias Judgments**

**Supervisor:** Molly Kennedy

**Examiner:** Hinrich Schütze

**Level:** BSc

**Summary:** The student will implement a compact "prompt perturbation" benchmark for media-bias judgments (e.g., role vs neutral prompts, alternative bias/framing definitions, and output formats such as label vs rationale vs scalar). Using a fixed dataset slice and 1–2 models, they will quantify stability (agreement/variance) and produce a small taxonomy of failure modes. Deliverable: a reproducible evaluation harness + short report.

**References:**

- Mikhail Seleznyov et al. (2025). "When punctuation matters: a large-scale comparison of prompt robustness methods for llms". In: *arXiv preprint arXiv:2508.11383*
- Zekun Li et al. (2024). "Evaluating the instruction-following robustness of large language models to prompt injection". In: *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 557–568
- Kaijie Zhu et al. (2023). "Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts". In: *Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis*, pp. 57–68