Deep Dive into Multilinguality Analysis and Evaluation of Open-Source Large Language Models through Neural Representations

- ☐ **Supervisor:** Ercong Nie
- ☐ **Examiner:** Prof. Hinrich Schütze
- ☐ **BSc/MSc Open:** Msc preferred ; Bsc. with strong interest considered
- ☐ **Summary:**

Large language models (LLMs), such as ChatGPT, Llama-2, and GPT-4, have showcased exceptional performance across a diverse spectrum of NLP applications and beyond. Despite these advances, the exploration of multilingual capabilities and underlying mechanisms in LLMs remains an area of ongoing research. Predominant work[1] on the multilingual evaluation of LLMs have primarily employed prompting methods[2,3], where the task is reformulated into natural language descriptions, accompanied by the input sequence (the so-called prompt). The models take the prompt and generate the outputs which are directly used for the evaluation. Recent studies, however, indicate that prompting cannot fully replace the probability measurements in the LLM evaluation[4]. A significant limitation arises with popular proprietary LLMs, which often restrict access to their internal neural representations, such as token embeddings and output logits, through their APIs.

To circumvent this limitation, our project will utilize the open-source LLMs like Llama-2 and Vicuna, enabling a deeper exploration into their neural representations for enhanced multilinguality analysis and evaluation. The potential research avenues within this context include:

1. **Tokenizer Analysis:** While the intricacies of tokenizers in closed-source LLMs remains inaccessible, we can perform detailed multilingual analysis of the tokenizers for open-source models[6].
2. **Cross-lingual Token Embeddings:** Access to token embeddings of open-source LLMs allows for an in-depth interpretation study of cross-lingual understanding within LLMs[7].
3. **Layer-wise Hidden States:** Probing the hidden states of individual layers can yield richer insights into the models' multilingual capabilities[8].
4. **Output Logits and Probabilistic Evaluation:** Utilizing the output logits from open-source LLMs, we can conduct precise probabilistic measurements to evaluate their multilingual performances[5].

[1] Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

[2] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

[3] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

[4] Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

[5] Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Unleashing the Multilingual Encoder Potential: Boosting Zero-Shot Performance via Probability Calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15774–15782, Singapore. Association for Computational Linguistics.
[6] Ahmed Alajrami, Katerina Margatina, and Nikolaos Aletras. 2023. Understanding the Role of Input Token Characters in Language Models: How Does Information Loss Affect Performance?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9085–9108, Singapore. Association for Computational Linguistics.
[7] Andrea Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-Lingual Interpretability in Token Embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.
[8] Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797, Singapore. Association for Computational Linguistics.

- [ ] **Prerequisites:** Enthusiasm in NLP research; good programming background in Python; experience in NLP, DL and PyTorch; basic knowledge of transformers; preferably with some linguistic background.