LLM Alignment: Behavior Steering of LLMs

Description

Large Language Models (LLMs) such as Llama, Qwen, and Gemma require careful alignment to ensure that their behavior follows human preferences, safety constraints, and task-specific requirements. This thesis explores modern alignment techniques including supervised fine-tuning, preference modeling, direct preference optimization (DPO), reinforcement learning from human or AI feedback (RLHF), and behavior steering through prompting or model post training. Student will investigate how to make open-source LLMs more controllable, predictable, and aligned across a variety of tasks. Related work:[1, 2, 3]

What we expect from you:

- Experience with deploying or fine-tuning open-source LLMs
- Basic understanding of reinforcement learning
- Interest in model safety, preference modeling, or human-centric AI
- High motivation to learn missing knowledge and work on challenging topics

What we provide:

- Computational resources
- Thesis supervision
- Possibility of scientific publication

For application please send me an email with title "Master Thesis Application: LLM Alignment ". Please also attach your resume and transcript of records in the email. An motivation letter is NOT required.

Tasks

- Conduct a literature review on LLM alignment methods
- Implement or adapt alignment techniques (e.g., SFT, DPO, RLHF, RLAIF)
- Build a small-scale human or automated feedback pipeline
- Fine-tune or evaluate aligned models on selected tasks
- Analyze behavioral changes and alignment effectiveness

References

[1] Fanchao Cao, Shuyang Xie, Yue Sun, Weikang Wang, Ruobing Xu, Ming Zeng, Jun Wang, Zeyang Min, Yiming Wang, Chengfei Li, et al. Defending against alignment-breaking attacks via robustly aligned llm. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024.

[2] Haoxiang Li, Xianjun Wang, Yutao Li, Wei Ye, Zhou Yu, and Houfeng Wang. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Technische Universität München

TUM School of Computation, Information and Technology

Lehrstuhl für Robotik, Künstliche Intelligenz und Echtzeitsysteme

---

Supervisor:
Prof. Dr.-Ing. Alois Knoll

Advisor:
Fengjunjie PAN, M.Sc.

Research project:
-

Type:
MA

Research area:
LLM, Post Training, Reinforcement Learning

Programming language:
Python

Required skills:
Python Transformers, LLM Training/Finetuning, Reinforcement Learning, Ubuntu, Virtualization

Language:
English or German

For more information please contact us:

E-Mail: f.pan@tum.de

Internet: www.ce.cit.tum.de/air

Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2025.

[3] Zeqi Lin, Ziniu Chen, Chenghao Shao, Tianwei Zhang, and Ce Zhang. A comprehensive survey of llm alignment: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024.

Technische Universität München

TUM School of Computation, Information and Technology

Lehrstuhl für Robotik, Künstliche Intelligenz und Echtzeitsysteme