Learning &
Adaptive Systems

ETH zürich

Improving and Evaluating LLM Alignment

For Semester Project or Master Thesis

Project Description

We are looking for motivated students to work on research projects related to improving and evaluating the alignment of large language models (LLMs), that is, how to ensure that models behave in accordance with human values, safety constraints, and domain-specific norms (e.g., legal or institutional guidelines).

Evaluation of LLM Alignment. Evaluating how well language models align with a user's preferences or guidelines is a challenging but crucial problem. Potential projects are concerned with designing benchmarks to evaluate LLM alignment with domain-specific, non-verifiable objectives and values, e.g., the Swiss AI Charter. This will involve (a) methods for synthetic generation of domain-specific evaluation data, and (b) research on how to build and train reliable LLM-based preference models ("LLM-as-a-judge") to rank model alignment. Depending on the exact layout of the project, this work has the opportunity to be attributed to and credited by the ongoing Swiss AI Apertus model development.

Improving LLM Alignment. Beyond evaluation, we also offer projects on *improving* LLM alignment. The goal is to develop training and post-training methods (e.g., preference optimization, RLHF/RLAIF-style approaches, or alternative feedback mechanisms) that make models better aligned under realistic resource constraints. Typical challenges include sample efficiency (high-quality prompts or human labels are scarce and expensive), reliance on noisy or biased synthetic labels, and maintaining robustness and generalization when training from synthetic data only. We offer several projects in this space, ranging from experimental/empirical work (e.g., large-scale training and ablations) to more theory-oriented projects.

Ideal Candidate

- Prior experience in machine learning, illustrated by previous projects or coursework, preferably in language models or reinforcement learning.
- Strong interest in AI alignment, reinforcement learning, and/or LLM evaluation.
- Strong track record of academic performance and/or past research/internship experience.
- Strong coding skills; experience with large-scale training and modern machine learning or language modeling frameworks is a plus for the experimental projects.

The projects will be supervised by Thomas Kleine Buening and potentially co-supervised by other people from the LAS group or the ETH AI Center. The project/thesis would be registered with Prof. Andreas Krause.

Contact

Send an email to thomas.kleinebuening@ai.ethz.ch to discuss these projects in more detail. Please attach your CV, transcripts, and briefly explain why you are interested in this project.