Learning & Adaptive Systems

ETH zürich

Project for LLM Safety and Alignment

Project Proposal for Master Thesis or Semester Project

Project description

Recent advances in large language models (LLMs) have demonstrated unprecedented capabilities, but also revealed critical alignment challenges that could pose catastrophic risks as AI systems become more powerful. We look for highly motivated and ambitious students to address some fundamental problems in AI safety: reward hacking, and AI control and scalable oversight.

Reward Hacking Current LLMs often exhibit undesirable behaviors like sycophancy (1) (agreeing with users regardless of truth), deception (2), and subtle manipulation of human evaluators. These behaviors emerge because models optimize for reward signals that imperfectly capture human values. Students will work on developing novel techniques to detect and mitigate reward hacking, including:

- Developing training methods that discourage gaming of reward signals.
- Building principled and interpretable tools to understand when and why models engage in reward hacking.

AI Control and Scalable Oversight As AI systems tackle increasingly complex tasks, human oversight becomes both more critical and more challenging. We need methods to ensure AI systems remain aligned even when their capabilities exceed human ability to directly evaluate their outputs. Research directions include:

- Developing automated red-teaming methods to discover failure modes before deployment.
- Developing systems or theories for monitoring and constraining AI's outputs to detect and prevent unsafe actions.
- Scalable safety and control analysis in model activations, chain of thought reasoning, and inter-model communication.

Under exceptional cases, a student can propose their most interested other LLM safety topics, but it has to be a high-quality proposal that falls into these research areas.

Most of the projects will be jointly supervised by Xin Chen (Cynthia) (at Prof. Andreas Krause's group) and Lukas Fluri (at Prof. Florian Tramer's group). Depending on the project's focus, you may register your project/thesis under Andreas or Florian.

Ideal Candidate

We seek motivated students with:

- Existing experiences in machine learning, illustrated by previous projects, preferably in language models, reinforcement learning, or adversarial attack/defense.
- Strong interests in AI safety and alignment, and its impact on humanity in the long term.
- Strong track record of academic performance or past research/internship experiences.

Contact

If you are interested in this project, please send an email with two recipients: xin.chen@inf.ethz.ch and lukas.fluri@inf.ethz.ch. Please attach your resume, transcript of records, and briefly state why you are interested in this project.

References

[1] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston *et al.*, "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[2] T. Hagendorff, "Deception abilities emerged in large language models," *Proceedings of the National Academy of Sciences*, vol. 121, no. 24, p. e2317967121, 2024.