



Projects: Safety and Alignment Science of Frontier Models

PROJECT PROPOSAL FOR SEMESTER PROJECT OR MASTER'S THESIS

Motivation

In practice, frontier models are rarely used as released. To remain performant on narrow, domain-specific tasks, models typically undergo prolonged training via LoRA, supervised fine-tuning (SFT), preference optimisation, or SDPO/ self-distillation [1, 2]. Problematically, recent studies show that such parameter updates can reopen or amplify misalignment [3, 4, 5] that was previously mitigated during post-training. These *anthropomorphic misalignment risks* (AMR) include:

- sycophancy (agreeing to user opinions regardless of truth)
- deception or manipulation
- evaluation awareness
- shutdown resistance

and many more [6]. Open research questions include *when* AMR temporally emerges during narrow fine-tuning, *how* their representations geometrically evolve, and *whether* AMR can be mitigated actively using mechanistic interpretability techniques in-training.

What we work on

- **Misalignment science.** How safety-relevant representations form and degrade across pretraining, fine-tuning, and post-deployment adaptation stages
- **Active mechanistic interpretability.** Steering vectors, circuit analysis and probe-guided detection employed as intervention proactively while training
- **Open benchmarks & evaluations.** Tools, datasets, and reproducible benchmarks on emergent misalignment so other researchers can build on top
- **Harness safety.** How tool calling, memory, and multi-agent coordination in harness systems create or compound anthropomorphised misalignment risks

Ideal Candidate

We look for highly motivated students with:

- Strong background in LLMs, post-training methods (RLHF, DPO, SFT), and ML frameworks such as PyTorch

- Strong academic performance or prior research or internship experience
- Strong commitment to advancing AI safety, coupled with high agency and an independent, structured working style

Most of the projects will be jointly supervised by [Anna Hedström](#) (at Prof. Andreas Krause's and Prof. Menna El-Assady's group) and [Xin Chen \(Cynthia\)](#) (at Prof. Andreas Krause's group), [Lukas Fluri](#) (at Prof. Florian Tramèr's group) and [Yannick Metz](#) (at Prof. Menna El-Assady's group). Depending on the project's focus, you may register your project/thesis under Andreas, Florian or Menna.

Contact

If you are interested in this project, please submit your application via [this form](#).

Thank you!

References

- [1] Thomas Kleine Buening, Jonas Hübotter, Barna Pásztor, Idan Shenfeld, Giorgia Ramponi, and Andreas Krause. Aligning language models from user interactions. *arXiv preprint arXiv:2603.12273*, 2026.
- [2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.
- [3] Muhammed Ustaomeroglu and Guannan Qu. Block-em: Preventing emergent misalignment by blocking causal features, 2026.
- [4] Jeremiah Giordani. Re-emergent misalignment: How narrow fine-tuning erodes safety alignment in llms. *arXiv preprint arXiv:2507.03662*, 2025.
- [5] Ann-Kathrin Dombrowski, Dillon Bowen, Adam Gleave, and Chris Cundy. The safety gap toolkit: Evaluating hidden dangers of open-source models. *arXiv preprint arXiv:2507.11544*, 2025.
- [6] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.