Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Multi-Agent Collaboration for Safe and Robust RLHF

Reinforcement learning from human feedback (RLHF) has become the de facto approach to align large language models. However, RLHF pipelines are fragile: *preferences are noisy or adversarial*, and *exploration can surface unsafe or harmful outputs*. A single reward model or majority-vote aggregation is often insufficient to guarantee safety and robustness.

Recent advances in multi-agent systems show that panels of specialized critics, debaters, or verifiers can outperform single models on alignment tasks. Yet, it remains unclear how to best use *multi-agent collaboration* to (i) robustly aggregate preferences under noise/adversaries, and (ii) pre-filter unsafe generations before they reach human annotators. Both directions are critical to improve safety while respecting tight budgets of human labels and tokens.

This project investigates whether *multi-agent collaboration*—in the form of critic councils and robust preference aggregation—can improve the safety and robustness of RLHF. The student will build a reproducible RLHF sandbox, implement multi-agent safety filters and robust aggregation methods, and evaluate them against strong baselines. The exact datasets, safety metrics, and aggregation schemes will be finalized together to match the student's background and compute resources.

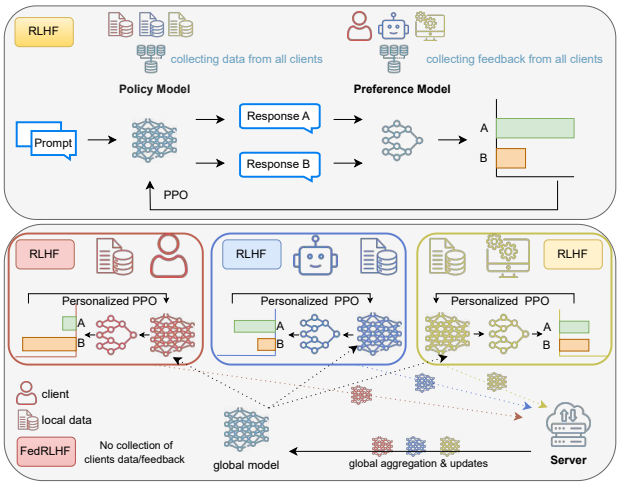

Figure 1: Illustration of the FedRLHF framework [1].

Interested? Please contact us for more details!

Contact

- Flint Xiaofeng Fan: xiafan@ethz.ch, ETZ G97

References

[1] Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. Fedrlhf: A convergence-guaranteed federated framework for privacy-preserving and personalized rlhf. In *AAMAS*, 2025.