Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

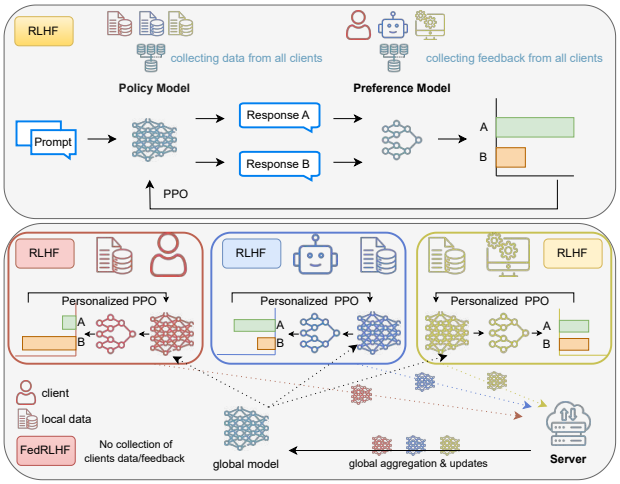
Distributed Computing

Prof. R. Wattenhofer

Topics on Decentralized RLHF

Reinforcement Learning from Human Feedback (RLHF) has become the cornerstone technique for aligning large language models with human preferences by fitting a reward model to human pairwise comparisons and then fine-tuning via Policy Optimization (e.g., PPO).

Yet the centralized form of RLHF remains hampered by prohibitive annotation expense, amplified evaluator bias, susceptibility to jail-break prompt attacks, reward mis-generalization that encourages spurious shortcuts, persistent hallucinations, and the sheer compute and bandwidth demanded by repeated policy-reward loops. Our recently introduced Federated RLHF framework [1] reframes several of these pain points: human comparisons are gathered locally so no monolithic data silo is required; each client trains its own reward model, preserving preference diversity while federated aggregation provably converges to a strong global policy. Despite its strengths, FedRLHF still grapples with heavy communication overhead from exchanging large model updates, client-drift caused by non-IID feedback, and privacy risks inherent in gradient sharing, etc.



You might tackle these (and many more) challenges of RLHF for AI alignment by designing more innovative decentralized RLHF frameworks with compression or sparsification techniques to curb bandwidth, heterogeneity-aware or meta-learning optimizers to stabilize convergence, or differential-privacy mechanisms to safeguard reward-update protocols, etc. Alternatively, you could build practical applications—such as privacy-preserving clinical chatbots, personalized edge-device assistants, or decentralized content-moderation tools—that demonstrate its real-world viability. Be creative!

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.