ActiveUltraFeedback: Efficient Preference Data Generation using Active Learning

Davit Melikidze[1,*] Marian Schneider[1,*] Jessica Lam[2,*] Martin Wertich[1,*] Ido Hakimi[1,3] Barna Pásztor[1,3] Andreas Krause[1,3]

[1]ETH Zürich · [2]Institute of Neuroinformatics, University of Zürich & ETH Zürich · [3]ETH AI Center · [*]*Equal contribution*

ETH AI CENTER

Learning & Adaptive Systems

ETH zürich

Overview

THE PROBLEM

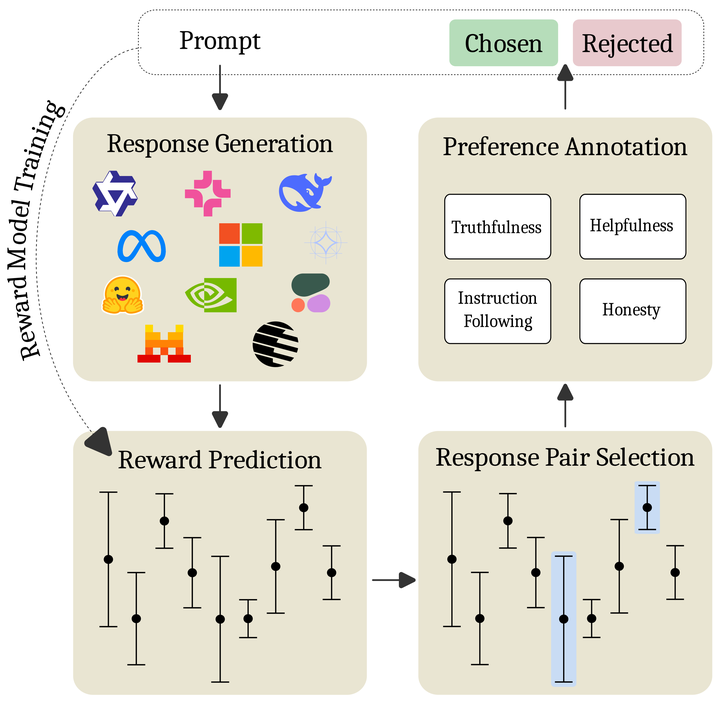
RLHF is the standard for aligning LLMs with human values, but acquiring preference data is **prohibitively expensive**, especially in low-resource and expert domains. The bottleneck is labeling: which response pairs are actually worth annotating?

CONTRIBUTIONS

- **ActiveUltraFeedback** — a scalable, modular active learning pipeline for preference data generation.
- **DRTS & DeltaUCB** — two novel response pair selection methods that target large quality differences for sample efficiency.
- The first systematic benchmarking of response pair selection methods across reward modeling and downstream tasks.
- Open-sourced code, preference datasets, and models.

The ActiveUltraFeedback Pipeline

For every prompt, we actively select the most informative response pair for annotation using the reward and uncertainty estimates of an Epistemic Neural Network, which is then iteratively trained on the accumulated preference labels.



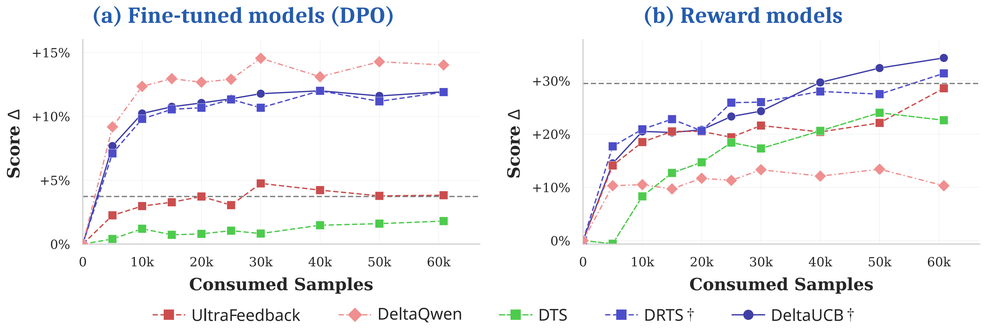
THE TAKEAWAY

Active Learning is the key to effective and data-efficient preference learning.

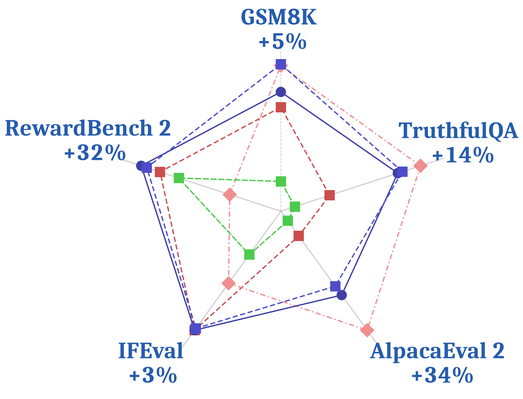
1/6 By actively selecting pairs with large predicted quality differences, ActiveUltraFeedback matches or beats static baselines while using as little as one-sixth of the annotated data across reward modeling and downstream tasks.

Sample Efficiency

DRTS and **DeltaUCB** lead to top performance with 5-10k samples, outpacing training on the full 60k samples in the original UltraFeedback dataset as well as the datasets produced by other selection methods.



Selection Method Comparison

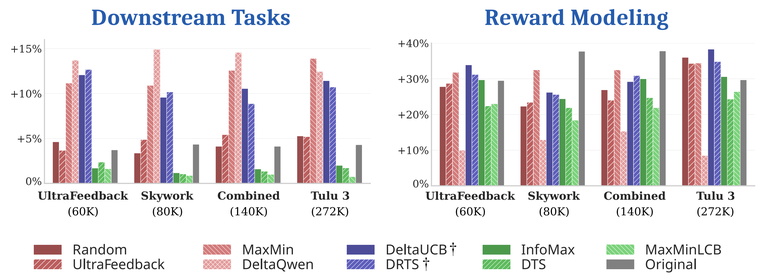


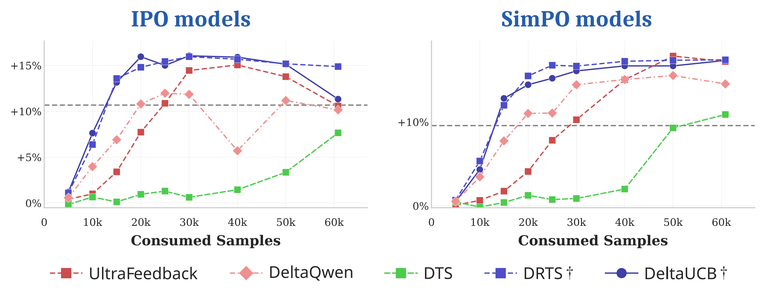
Improvement over the base model, averaged across four prompt datasets. **DRTS** and **DeltaUCB** dominate the full profile. **DeltaQwen** is better on certain downstream tasks but underperforms on reward modeling.

- UltraFeedback
- DeltaQwen
- DTS
- DRTS †
- DeltaUCB †

Generalization across Prompts & Algorithms

Sample efficiency and performance gains hold across four prompt datasets of increasing scale (UltraFeedback → Tulu 3, 60k → 272k) and across preference optimization algorithms.





Conclusions

TAKEAWAYS

- **Best ≠ informative.** A pair of excellent responses is less useful to label than a pair with a clear quality gap.
- **Sample efficiency.** DRTS & DeltaUCB maximize performance under strict annotation budgets.

FUTURE WORK

- **Active prompt selection.** Identify the most useful prompts based on reward and uncertainty estimates.
- **Active model selection.** Identify the two models per prompt for response generation, to reduce cost of using large model pools.

Acknowledgments

Supported by the Swiss National Supercomputing Centre (CSCS). B. Pásztor and I. Hakimi are supported by an ETH AI Center doctoral / postdoctoral fellowship. M. Schneider, J. Lam, and M. Wertich are supported by the G-Research Grant.