
# COALA: Convex Optimization for Alignment and Preference Learning on a Single GPU

MIRIA FENG AND MERT PILANCI

43rd Conference on International Conference on Machine Learning (ICML 2026)

# Background

1. **Preference Alignment of LLMs**
   - Paradigm: pre-training -> SFT -> preference alignment (RLHF, DPO)
   - RLHF is resource and data expensive, also requires Humans in the loop. DPO utilizes the Bradley-Terry ranking objective, with a reference model for stability. Many other heuristics driven methods!

2. **Convex Neural Networks**
   - Pilanci and Ergen[1] developed exact representations of training two-layer ReLU NN with a single convex program
   - CRONOS[2] developed a framework for applying convex neural networks on LLMs, on the scale of GPT-2

[1] Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks (2020)
[2] Cronos: Enhancing deep learning with scalable gpu accelerated convex neural networks (2024)

# **COALA**: **C**onvex **O**ptimization for **A**lignment and Preference **L**earning **A**lgorithm

Derived from a reformulation of the preference objective to a convex formulation.

The key idea: attach a convex neural network on top of frozen LLM representations and optimize it using convex methods such as CRONOS.

**Algorithm 1** Convex Preference Optimization (COALA)

**input** Dataset $(x, y_w, y_l)$, Pre-trained model $f_{\theta_{\text{pre}}}(x)$, offset parameter $\gamma$, penalty parameter $\rho > 0$

**Phase I**: **Train the policy network**

Train $\pi_\theta^{\text{cvx}}$ to obtain $(\Theta_1, \theta_2)$ by solving (4) using CRONOS($\rho$) (Algorithm 2).

**Phase II**: **Finetuning**

Freeze weights of the first layer $\Theta_1$.

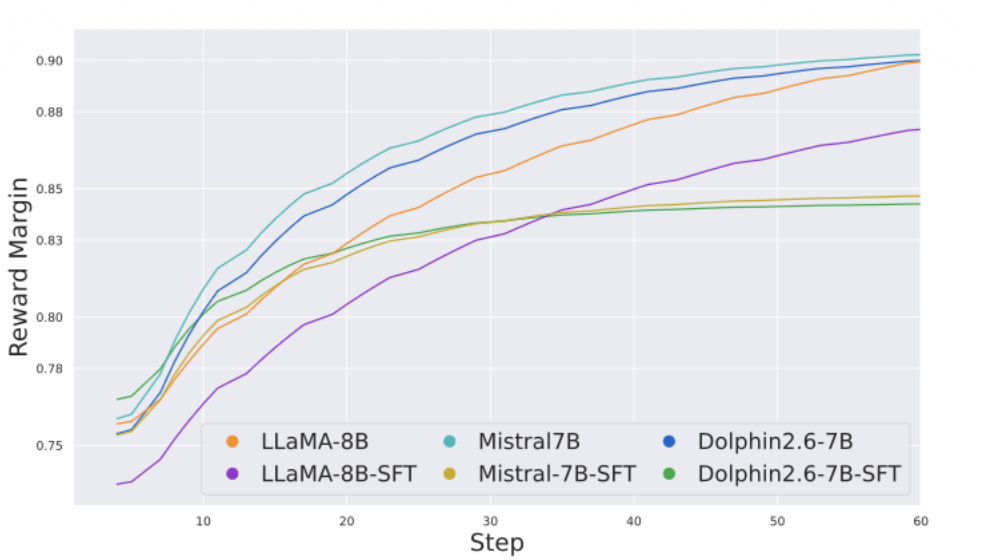
Finetune weights of second layer $\theta_2$ by solving the convex minimization problem (9):

$$\min_{\theta_2} \; L_{\text{COALA}}(\pi_{\theta_2}^{\text{cvx}}) \quad \{\text{Solve via AdamW}\}$$

**output** $(\Theta_1, \theta_2)$.

# Four Datasets x Six Models x Four Methods

Metrics include: ArenaHard, AlpacaEval2, MT-Bench, and a 107 double-blind human sample experiment.




*(a)* COALA

# Alternating Population Strategy for Preference Datasets

- Maximum value from real world dialog data.
- No synthetic model generated rejected responses.
- No reward model required.
- https://huggingface.co/datasets/miria0/EduFeedback

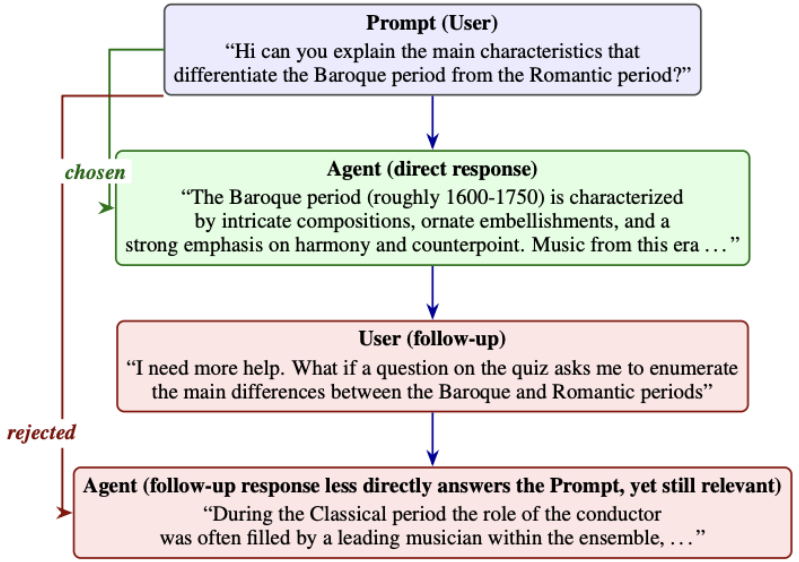



*Figure 2.* Alternating Population Method for creating preference datasets. One conversation yields multiple (prompt, chosen, rejected) preference triplets, without requiring external LLMs to generate matching responses in chosen-rejected pairs.

# Contributions and Conclusions

- We prove COALA's convergence guarantees in Section 4, to show under mild theoretical conditions we can train COALA loss to global optimality.
- This results in stable reward margin gains, and paradigm shift towards more interpretable preference alignment.
- COALA's design intentionally combines the convex architectural optimality (which we guarantee) and model-wide expressiveness (which we trade for efficiency).
- Limitations and practical applications: personal assistants, where speed and data privacy are key.
- Future work: generalizing the COALA method to more advanced algorithms, such as GRPO.

Please see codebase at https://github.com/pilancilab/COALA