Evolution and Application of Reinforcement Learning (RL) in Large Language Models (LLMs)

Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
http://nice.rongpeng.info/

Aug. 12, 2025

Content

Development Outline

Overview

From Alignment to Advanced Reasoning

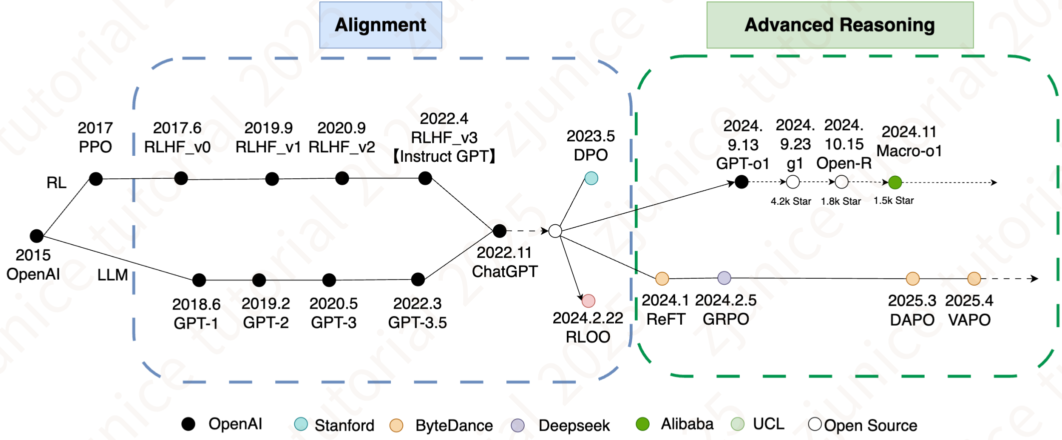
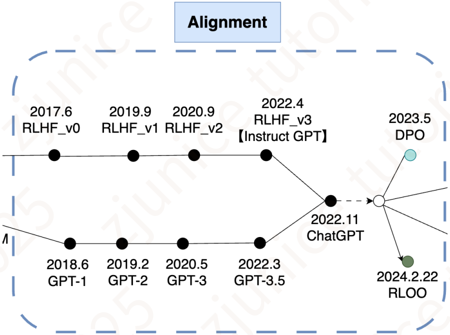


Figure: Roadmap of RL in LLM.

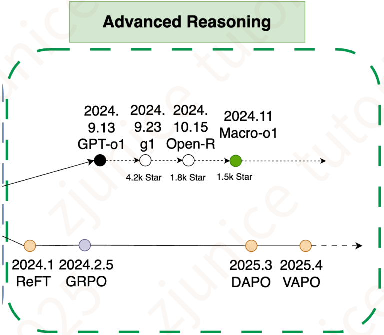
Emergence: Alignment Tuning

- **Goal:** Align the behavior of LLMs with **human preferences**.
  e.g., helpfulness, honesty, and harmlessness.
- **Motivation:** Reduce the reliance on costly human annotations.
- **Typical Approaches:**
  1. **With Reward Model (RM):**
     Train a separate RM to approximate human preferences; then fine-tune the LLM using PPO-based RLHF.
     e.g., InstructGPT.
  2. **Without RM:**
     Directly learn from preference pairs via implicit objective optimization.
     e.g., DPO.



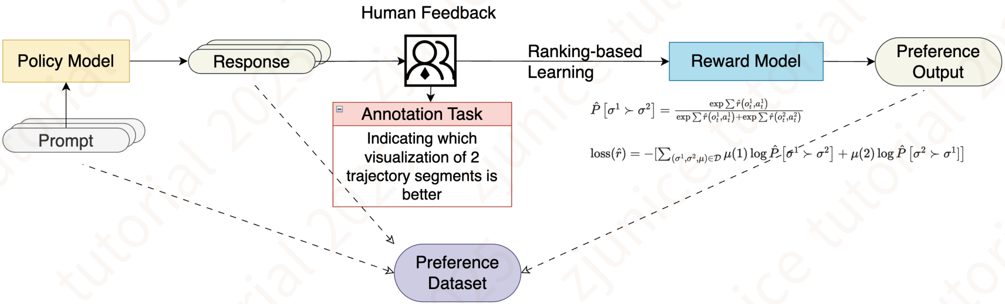
Development: Enhanced Reasoning

- **Goal:** Directly optimize final **task performance**.
  e.g., accuracy, usefulness, reasoning quality.
- **Motivation:** Overcome the limitations of Supervised Fine-Tuning (SFT), which is constrained by fixed reasoning paths and data quality.
- **Typical Approaches:**
  - **Without Training:**
    Reasoning path search methods.
    e.g., Monte Carlo Tree Search (MCTS) + Chain-of-Thought (CoT)/Chain-of-Action(CoA), Prompt Engineering.
  - **With Training:**
    Fine-tuning via RL with rewards.
    e.g., ReFT, GRPO.



Evolutionary History of the System Framework

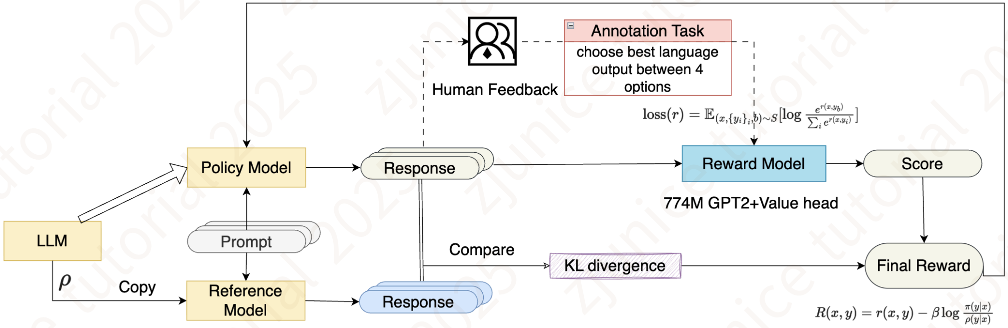
System Framework 1: RLHF_v0[1]



- **Task:** traditional RL benchmarks (e.g., Atari games).
- **Core Contribution:**
  - Preference-Based **Reward Modeling**;
  - Policy Optimization with Learned Rewards.

[1] P. F. Christiano et al., "Deep reinforcement learning from human preferences,", 2017.

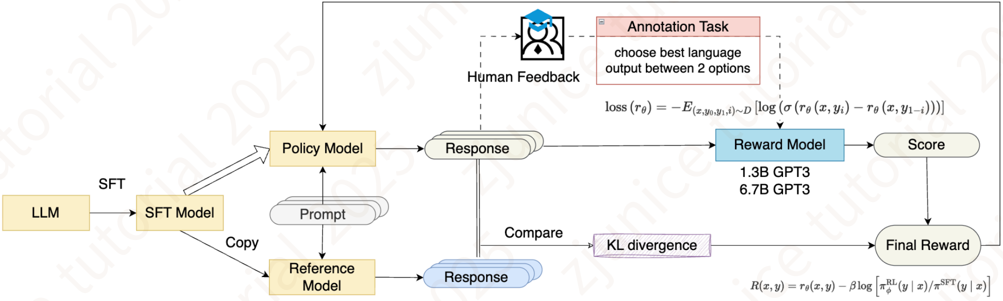
System Framework 2: RLHF_v1[2]



- **Task:** First RLHF for LLMs.
- **Core Contribution:**
  - Novel **Reward Model Initialization** and **KL Control**;
  - Online Data Collection to Mitigate Distribution Shift.

[2] D. M. Ziegler et al., "Fine-tuning language models from human preferences,", 2019.

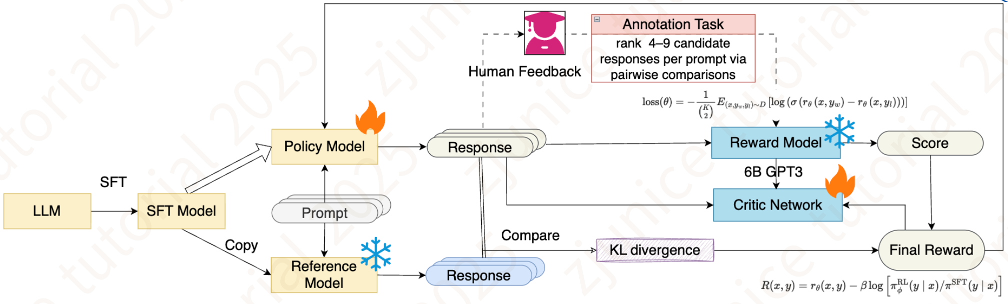
System Framework 3: RLHF_v2[3]



- **Core Contribution:**
  - Expert Annotators & **Pairwise Preference Labeling**;
  - Improved Alignment through **SFT + RLHF** Pipeline;
  - Reward Model Initialization from SFT Policy.

[3] N. Stiennon et al., "Learning to summarize with human feedback,", 2020.

System Framework 4: RLHF_v3 (InstructGPT)[4]



- **Core Contribution:**
  - Higher-Quality Expert Labelers;
  - Multi-Response Generation with **Full Pairwise Ranking** ;
  - Scaling Reward Model with GPT-3.

[4] L. Ouyang et al., "Training language models to follow instructions with human feedback,", 2022.

System Framework of RL in LLM

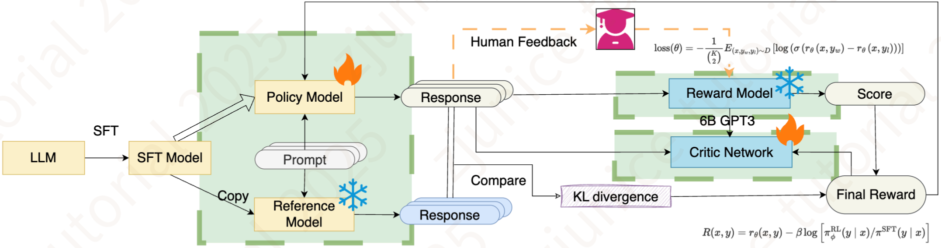


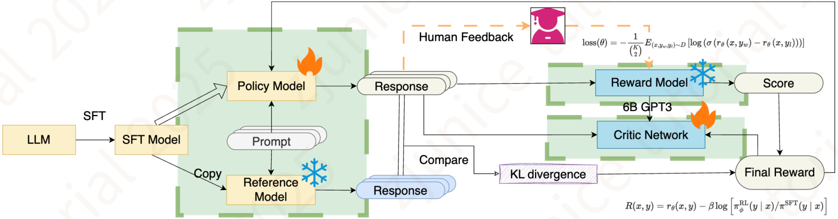
Figure: Complete workflow of RLHF.

Components

- Actor
- Critic (×)
- Reward (×)

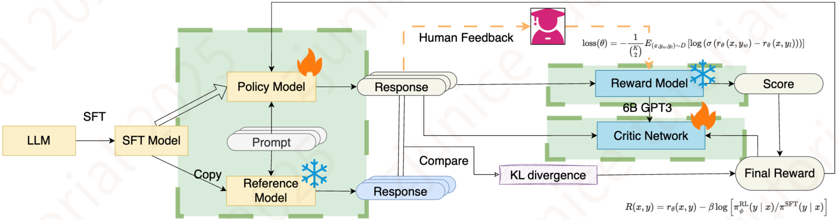
Technical Classifications

Technical Classifications 1: Reward



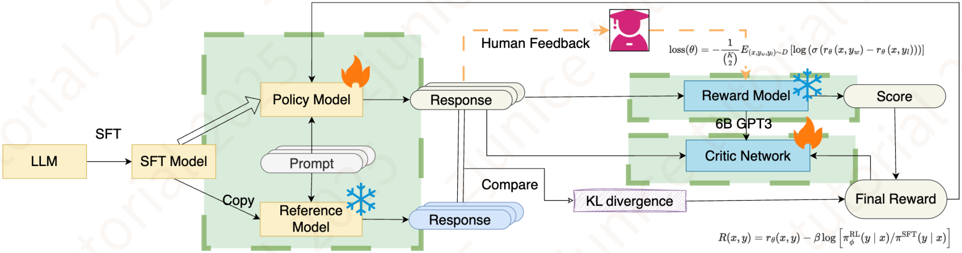
- **Implicit Rewards:** Learn from comparisons. *e.g.,* *DPO*
- **Explicit Rewards:**
  - **By source:** *source illustration*
    1. **External**: RM is a separate small LLM.
    2. **Internal**: shares parameters with actor, with a reward head.
  - **By temporal granularity:** *PRM vs. ORM illustration*
    1. **Outcome Reward Model (ORM)**: Outcome-level. *e.g.,* *GRPO*
    2. **Process Reward Model (PRM)**: Sequence- or token-level. *e.g.,* *PRIME*

Technical Classifications2: Critic Design



- **With Critic:**
  Estimates expected value for current policy.
  1. Initialized from RM. *e.g., InstructGPT*
  2. Initialized from policy and perform Value-Pretraining. *e.g.,* *VAPO*
- **Without Critic:**
  Relative ranking used as signal (no value function).
  Key idea: Avoids bias from value estimation. *e.g.,* *RLOO*, *DAPO*, GRPO

Challenges in Reward Modeling



- **Distribution Shift/Overoptimization:** Reward model becomes misaligned.
- **Reward Hacking:** Model exploits reward shortcuts.
- **Reward Sparsity:** Good samples are rare.
- **Cost:** Human preference labels are expensive.
- **Non-additivity:** Mixed (PRM and ORM) reward scales cause instability.

Thank you

Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
https://nice.rongpeng.info

Appendix

Source of Reward Model

LLM + Value Head (Linear Layer)

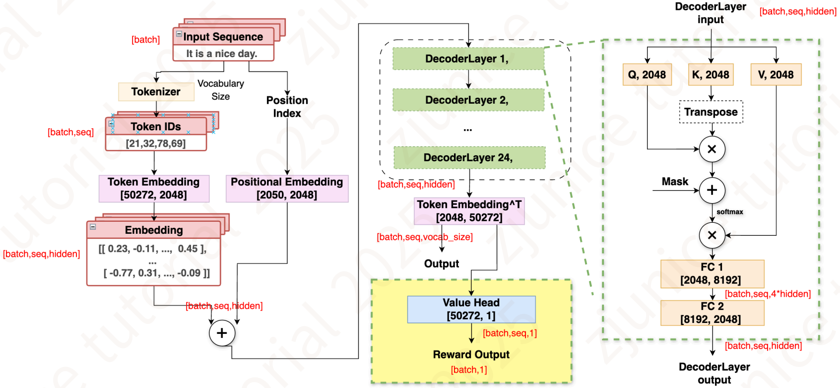


Figure: external vs. internal source of RM.

Back Link to reward

Outcome Reward Model (ORM) vs. Process Reward Model (PRM)

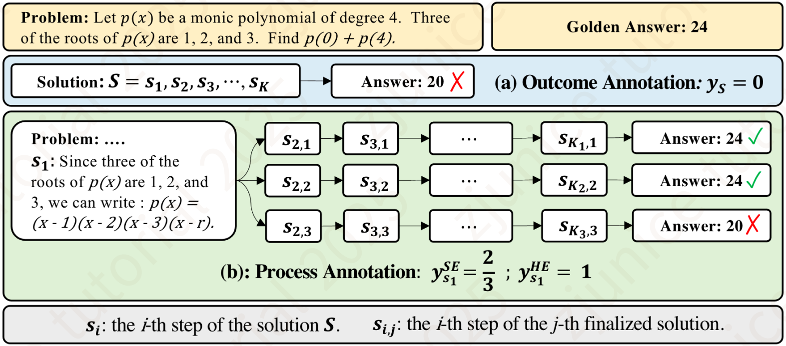
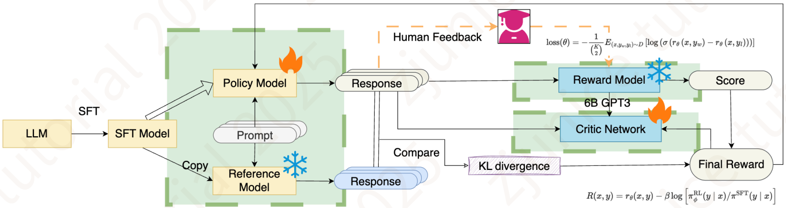


Figure: Example of different RM in Math-Shepherd[5].

Back Link to reward

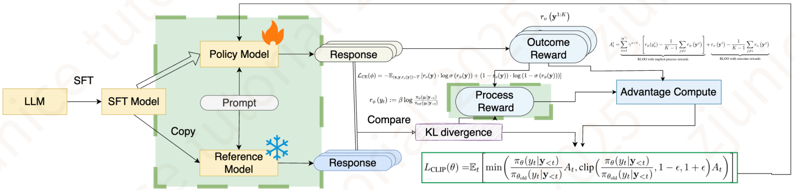
[5] P. Wang et al., “Math-shepherd: Verify and reinforce llms step-by-step without human annotations,” in *Proc. ACL*, 2024.

PRIME:Process Reinforcement through IMplicit rEwards[6]



Component

- Actor
- ~~Critic~~
- Reward



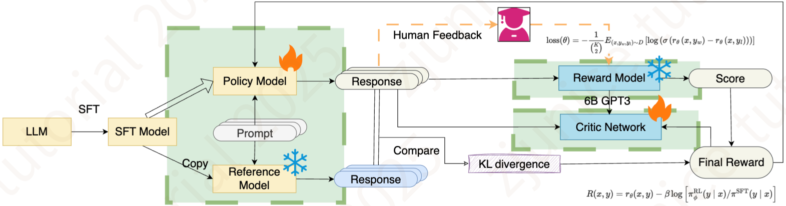
PRIME

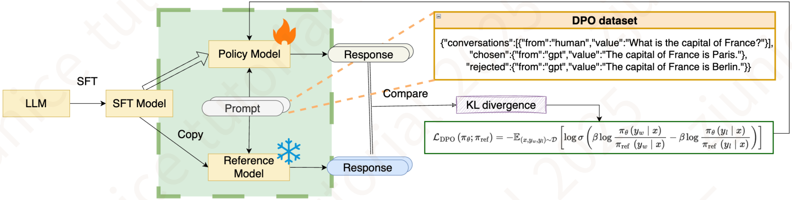
- PRM Supervision from ORM.

Back Link to reward

[6] G. Cui et al., "Process reinforcement through implicit rewards," *arXiv preprint arXiv:2502.01456, 2025* .

DPO: Direct Preference Optimization[7]





Component

- Actor
- ~~Critic~~
- ~~Reward~~

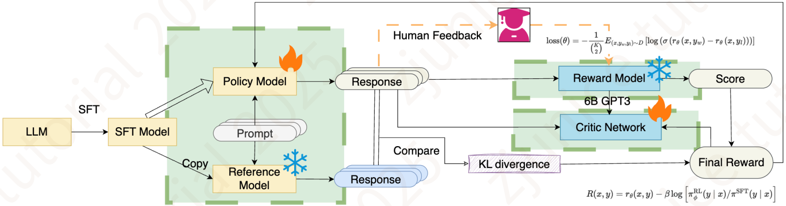
DPO

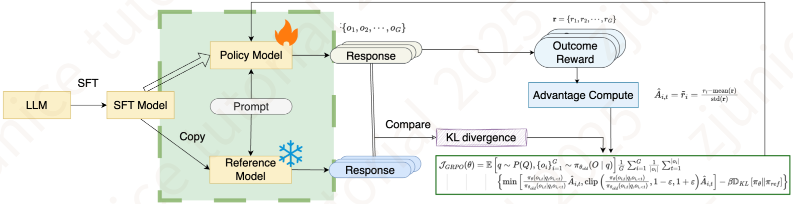
- pairwise comparison

Back Link to Reward

[7] R. Rafailov et al., "Direct preference optimization: Your language model is secretly a reward model,", 2023.

GRPO:Group Relative Policy Optimization[8,9]





Component

- Actor
- ~~Critic~~
- Reward

GRPO

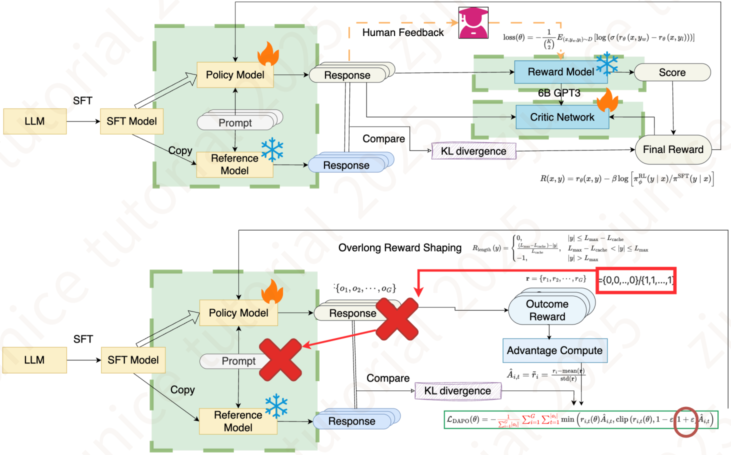
- Average group advantage.

Back Link to Reward   Back Link to Critic

[8] Z. Shao et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300, 2024* .

[9] D. Guo et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948, 2025* .

DAPO:Decoupled Clip and Dynamic sAmpling Policy Optimization[10]



- ~~Critic~~
- ~~Reward~~

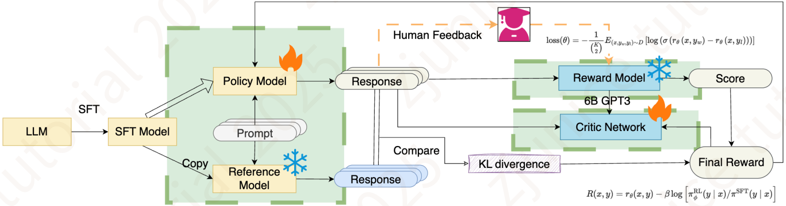
DAPO

- Remove KL.
- Clip-Higher.
- Dynamic Sampling.
- Token-Level Policy Gradient Loss.
- Overlong Reward Shaping

Back Link to Critic

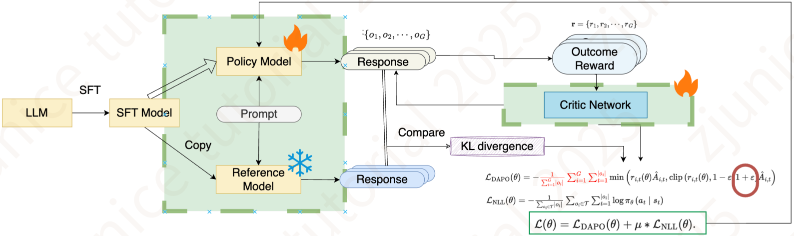
[10] Q. Yu et al., “Dapo: An open-source llm reinforcement learning system at scale,” *arXiv preprint arXiv:2503.14476, 2025* .

VAPO:Value Augmented proximal Policy Optimization[11]



Component

- Actor
- Critic
- ~~Reward~~



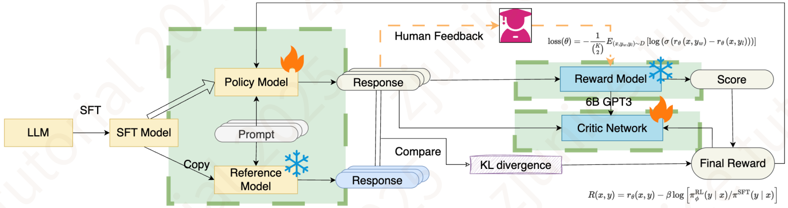
VAPO

- Value-Pretraining.
- Positive Example LM Loss.

Back Link to Critic

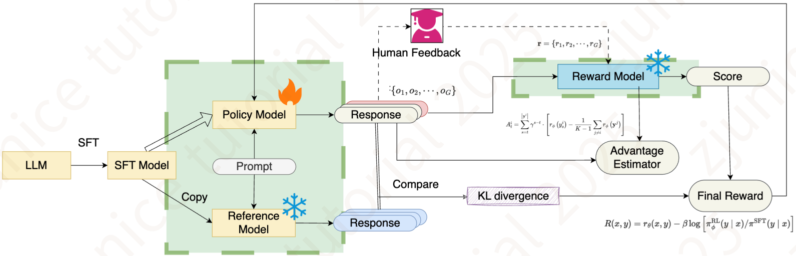
[11] Y. Yue et al., "Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks," *arXiv preprint arXiv:2504.05118, 2025* .

RLOO:REINFORCE Leave-One-Out[12]



Component

- Actor
- ~~Critic~~
- Reward



RLOO

- The average of the other samples in the group.

Back Link to Critic

[12] A. Ahmadian et al., "Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms," *arXiv preprint arXiv:2402.14740, 2024* .