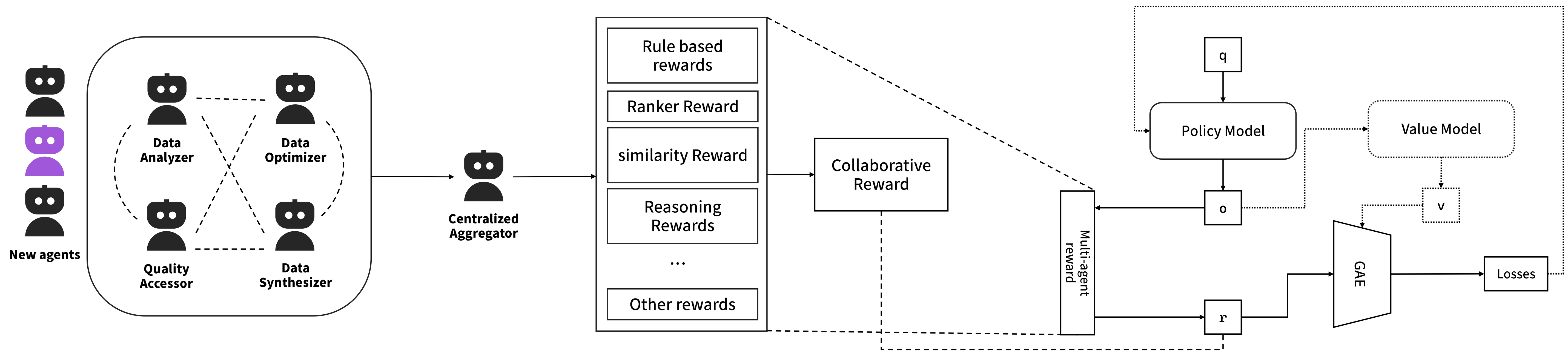
Multi-Agent Collaborative Reward Design for Enhancing Reasoning in Reinforcement Learning

Pei Yang, Ke Zhang, Ji Wang, Xiao Chen, Yuxin Tang, Eric Yang, Lynn AI, Bill

Abstract

We present Collaborative Reward Modeling (CRM), a frame-work that replaces a single black-box reward model witha coordinated team of specialist evaluators to improve ro-bustness and interpretability in RLHF. Conventional rewardmodels struggle to jointly optimize multiple, sometimes con-flicting, preference dimensions (e.g., factuality, helpfulness,safety) and offer limited transparency into why a score is as-signed. CRM addresses these issues by decomposing prefer-ence evaluation into domain-specific agents that each producepartial signals, alongside global evaluators such as ranker-based and embedding-similarity rewards. A centralized ag-gregator fuses these signals at each timestep, balancing fac-tors like step-wise correctness, multi-agent agreement, andrepetition penalties, yielding a single training reward com-patible with standard RL pipelines. The policy is optimizedwith advantage-based updates (e.g., GAE), while a valuemodel regresses to the aggregated reward, enabling multi-perspective reward shaping without requiring additional hu-man annotations beyond those used to train the evaluators. Weevaluate CRM on RewardBench, a benchmark suite alignedwith multi-dimensional preference evaluation, demonstratinga practical, modular path to more transparent reward model-ing and more stable optimization.



Methodology: Collaborative Reward Model (CRM)

We propose a Collaborative Reward Model (CRM) that enhances policy optimization by replacing traditional monolithic scalar rewards with a distributed, multi-agent evaluation framework. This ecosystem employs four specialist agents—the Data Optimizer, Quality Assessor, Data Synthesizer, and Data Analyzer—to cooperatively evaluate rollouts from complementary perspectives, ensuring robustness and diversity. The framework constructs a unified objective, R_{collab}, by aggregating multi-dimensional signals including step-level outcome verification, model-level semantic similarity (R_{sim}), and explicit constraints such as accuracy (R_{acc}), formatting(R_{fmt}), and repetition penalties. Finally, these heterogeneous signals are fused via a central aggregator into a scalar reward for standard RL updates using Generalized Advantage Estimation (GAE), guiding the policy $\pi_\theta$to balance factual correctness, reasoning clarity, and linguistic fluency in a transparent and extensible manner.

Table 1: Result of MARM in RewardBench, Math and GSM8K

Methods	Chat	Chat Hard	Safety	Reasoning	Math	GSM8K
Two Agents (Data Analyzer + Data Optimizer)						
Qwen2.5-0.5B-ins	0.193	0.561	0.561	0.598	0.139	0.08%
MARM	0.190	0.557	0.553	**0.659**	0.149	19.64%
MARM(rerank)	0.182	0.545	**0.566**	0.423	0.136	22.16%
MARM(emb)	**0.198**	**0.561**	0.536	0.567	0.131	**22.33%**
Three Agents (Data Analyzer + Data Optimizer + Quality Assessor)						
Qwen2.5-0.5B-ins	0.193	0.561	0.561	0.598	0.139	0.08%
MARM	0.190	0.557	0.553	**0.659**	0.149	19.64%
MARM(rerank)	0.190	**0.567**	0.538	0.398	0.143	22.87%
MARM(emb)	**0.199**	0.532	**0.570**	0.637	0.141	**23.15%**
Four Agents (Data Analyzer + Data Optimizer + Quality Assessor + Data Synthesizer)						
Qwen2.5-0.5B-ins	**0.193**	0.561	0.561	0.598	0.139	0.08%
MARM	0.190	0.557	0.553	**0.659**	0.149	19.64%
MARM(rerank)	0.182	**0.568**	0.527	0.610	**0.192**	**29.87%**
MARM(emb)	0.179	0.557	**0.573**	0.578	0.152	27.60%

Experiments and Results

We evaluated the proposed CRM framework using the Qwen2.5-0.5B-Instruct model optimized via Generalized Reinforcement Policy Optimization (GRPO) on RewardBench, GSM8K, and Math benchmarks. By progressively testing configurations from two to four agents, we demonstrated that the integration of specialized roles—specifically the Quality Assessor and Data Synthesizer—yields substantial gains in reasoning structure and generalization, with the full four-agent MARM variant achieving the highest accuracy (e.g., improving GSM8K performance to 29.87%). Our results confirm that the centralized reward aggregation strategy effectively balances these enhancements in mathematical precision and logical consistency without compromising general conversational fluency, thereby validating CRM as a robust, scalable, and modular approach for multi-dimensional policy optimization.