Adaptive Coopetition: Leveraging Coarse Verifier Signals for Resilient Multi-Agent LLM Reasoning (Student Abstract)

Rui Jerry Huang[1], **Anastasia Miin**[2], **Wendy Liu**[3]
[1]Basis Independent Silicon Valley, San Jose, CA, USA
[2]Pacific Collegiate School, Santa Cruz, CA, USA
[3]The Harker School, San Jose, CA, USA
ruihuang15352019@gmail.com, anastasiamiin9@gmail.com, blossomwliu@gmail.com

Abstract

Large language models (LLMs) demonstrate strong reasoning capabilities, yet the inference-time performance of existing solutions remains limited by self-biases, coordination inefficiencies, lack of robust error detection, and dependency on high-quality verifiers. To address these challenges, we propose **Adaptive Coopetition (AdCo)**, a lightweight, multi-agent multi-round inference-time framework that enhances collective reasoning through adaptive decision-making guided by coarse verifier signals. Without relying on high-performance verifiers, AdCo achieves a **20% relative accuracy improvement** on math reasoning benchmarks, with consistent performance on different sample sizes and agent configurations. This adaptive, signal-guided 'coopetition' framework enhances reasoning robustness by leveraging diverse model knowledge and reasoning traces, while also promoting uncertainty-driven exploration, especially when participants have comparable capabilities.

Introduction

LLMs exhibit strong reasoning capabilities but remain limited in certain scenarios because of their inherent pre-trained knowledge scope. Multi-agent frameworks facilitate collective intelligence among LLM agents through coordinated orchestration. However, this line of work often suffers from reasoning collapse, stemming from rigid strategies and reasoning contamination from low-quality peer feedback. To mitigate it, many methods were proposed, including leveraging strong verifiers to evaluate outputs and optimizing multi-agent architecture and reasoning processes. Unfortunately, these methods often lack inference-time adaptability and either require extensive training or assume a symmetric role for each agent, limiting their practicality during deployment.

To overcome these challenges, we propose Adaptive Coopetition—a lightweight inference-time, multi-round multi-agent framework that enhances collective reasoning through adaptive decision-making guided by coarse verifier signals. Specifically, after one-step of reasoning, each agent employs a coarse verifier to evaluate the current reasoning trace from multiple perspectives, producing what we term "verifier signals". Using these signals, each agent applies a revised Upper Confidence Bound (UCB) algorithm to decide whether to collaborate or compete. With the strategy determined, agents engage in peer-to-peer (P2P) interactions and asynchronously refine their reasoning based on peer feedback. This design deliberately isolates low-quality reasoning traces and iteratively improves the reasoning before integrating it into the cluster, enhancing reasoning quality and mitigating reasoning collapse.

Experiments on the DeepMath-103K math reasoning dataset demonstrate the effectiveness of our approach. The best-performing heterogeneous AdCo cluster outperforms both the State-of-the-Art LLMs and conventional multi-agent frameworks by approximately 20% in accuracy, while maintaining consistently strong performance across different data scales. Further ablation studies underscore the necessity of key components in AdCo, reinforcing our belief that AdCo offers a practical and effective solution that enhances collective reasoning.

Adaptive Coopetition

Figure 1 illustrates how AdCo Worker Cluster solves problems through multi-round optimization. At each turn, worker agents advance reasoning by one step and determine their strategy—collaboration or competition—via a UCB-based algorithm guided by verifier signals. The verifier signal, a weighted combination of reasoning progress and information diversity, is used in the revised UCB algorithm to choose the strategy for the current round, prompting agents to selectively exchange feedback with peers and refine its reasoning. This process repeats until a final solution is reached through a majority-vote algorithm.

Coarse verifier signals: Coarse verifier signals are verifier outputs of moderate precision in estimating progress and information diversity at inference time. High-precision verifiers often require substantial resources to train, and obtaining a sufficiently accurate verifier can be infeasible. Our empirical results show that even mediocre-quality signals from coarse verifiers can still serve the intended purpose, to filter out bad or inconsistent feedback while amplifying good and consistent ones in the reasoning process.

Low quality feedback isolation: To isolate unqualified feedback and prevent reasoning collapse, agents employ a customized filter and peer-to-peer communication, selecting

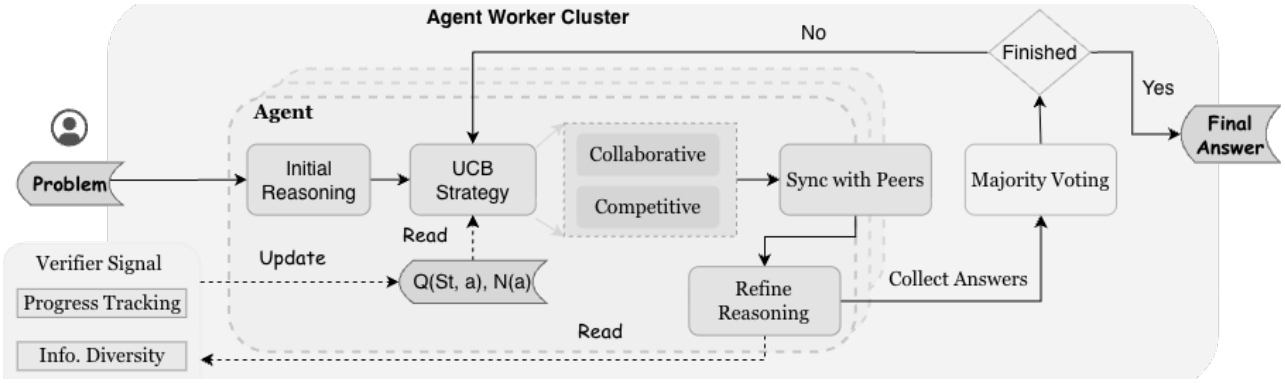

Figure 1: Overview of adaptive coopetition

high-scoring feedback in both collaborative and competitive strategies.

Iterative adaptive coopetition: Each agent's reasoning is modeled as a Markov decision process, with state s_t representing the reasoning trace and actions $a_t \in \{c_0, c_1\}$ denoting collaboration or competition. Rewards $r(s_t, a_t) \in [-1, 1]$ reflect changes in coarse verifier signal estimates. The action policy $\pi(s_t)$ is a revised UCB algorithm, inspired by UCT, incorporating state-dependent heuristics. Specifically, the chosen action a_t is the candidate action a that maximizes:

$$UCB'(s_t, a) = \frac{\sum_{i<t} \Delta V(s_i, a)}{N(a)} + C \times \sqrt{\frac{\ln N}{N(a)}}, a \in \{c_0, c_1\} \quad (1)$$

where the first term is the estimated payoff, N is the total action count, $N(a)$ is the count for action a, and $\Delta V(s_i, a)$ is a weighted combination of process reward and information diversity at state s_i where the chosen action is a.

Experiments

Experiment Setting: We evaluated AdCo's performance in the math domain using the *DeepMath-103K* dataset, a benchmark for advanced mathematical reasoning. We sampled 200 to 4,000 problems uniformly to ensure unbiased evaluation. Our AdCo model is set up on Microsoft AutoGen with three heterogeneous Agent Workers — using DeepSeek-v3-0324, Gemma-3-27b-it, and GPT-4o. Qwen2.5-Math-PRM-7B serves as the coarse verifier, with PR (Process Reward) used to track progress. The coefficient factor was chosen $C = \sqrt{1.5} = 1.22$ and the applied UCB':

$$UCB'(s, a) = \frac{\sum_{i<t} \Delta PR(s_i, a)}{N(a)} + \sqrt{\frac{1.5 \times \ln n}{n_a}} \quad (2)$$

Our baselines include 1) individual LLMs, 2) plain collaborative and competitive AutoGen setups, and 3) homogeneous AdCo variants using the same models.

Performance: AdCo outperformed all baselines, achieving 54% accuracy (Figure 2) with low variance across

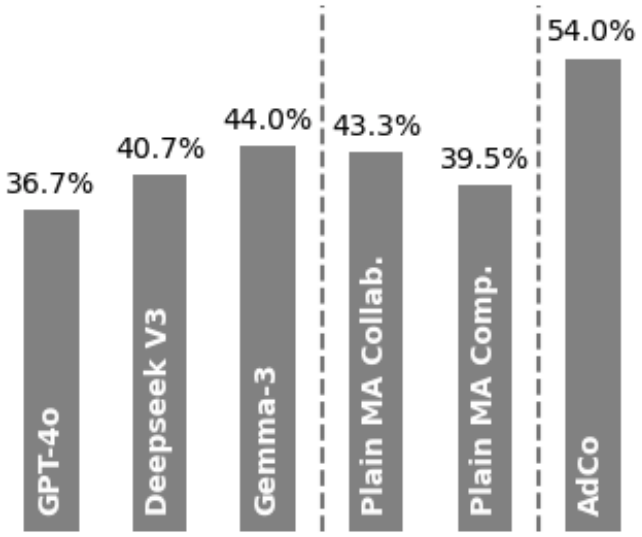

Figure 2: Model Accuracy

dataset sizes. Homogeneous setups yielded lower accuracies (42%–52%), confirming the benefit of model diversity. Strategy efficiency was validated by tracking answer switches: for instance, at 2,000 samples, AdCo corrected errors 1,016 times versus only 102 regressions, demonstrating its effectiveness in guiding reasoning.

Ablation Study: Replacing UCB' with a flipping rule — collaborating when the progress reward > 0.5 — led agents to make about three times as many corrections from incorrect to correct (1,401 vs. 509 under UCB'), thus, UCB' is much more efficient. The study also showed that AdCo is effective in homogeneous settings with stronger LLMs.

Future Work: Future improvements include information diversity in verifier signals, state-aware exploration, weighted result aggregation, strategy-specific parameter tuning, lightweight architectures for resource-limited settings, and expansion to broader domains.

References

Auer, P.; and et al. 2002. Finite-time Analysis of the Multi-armed Bandit Problem. *Machine Learning*, 47: 235–256.

He, Z.; and et al. 2025. DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning. arXiv:2504.11456.

Kocsis, L.; and Szepesvári, C. 2006. Bandit Based Monte-Carlo Planning. In Fürnkranz, J.; Scheffer, T.; and Spiliopoulou, M., eds., *Machine Learning: ECML 2006*, 282–293. Springer Berlin Heidelberg.

Qiu, X.; and et al. 2024. Towards Collaborative Intelligence: Propagating Intentions and Reasoning for Multi-Agent Coordination with Large Language Models. *CoRR*.

Tran, K.-T.; Dao, D.; Nguyen, M.-D.; Pham, Q.-V.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. arXiv:2501.06322.

Wu, Q.; and et al. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In *First Conference on Language Modeling*.

Zhang, Y.; and et al. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37: 132208–132237.