Learning Collaborative Reasoning Strategies Through Trust-Weighted Multi-Agent Consensus

Association for the Advancement of Artificial Intelligence

EKbana

1. E.K. Solutions Pvt. Ltd., Nepal
2. Stevens Institute of Technology, USA
3. Institut de F´ısica d'Altes Energies, Spain
4. Indiana University East, USA

AUTHORS

Projan Shakya* {1}, Kristina Ghimire* {1}, Kashish Bataju* {1}, Ashwini Mandal {1}, Sadikshya Gyawali {1}, Manish Awale {1}, Manish Dahal {1}, Shital Adhikari {1, 2}, Sanjay Rijal {1, 3}, You Young {4}, Vaghawan Ojha {1}

Correspondence to: vpo4@msstate.edu, vaghawan.ojha@ekbana.net

* Equal Contribution

Introduction

Large Language Models show remarkable reasoning skills, but their performance is often variable. On multi-step or high-stakes problems, individual models can produce inconsistent or incorrect answers due to inherent biases or domain-specific knowledge gaps.

- **Variability:** The same model can give different answers to the same problem.
- **Brittleness:** Performance can degrade significantly on slightly different problem phrasings.
- **Limited Dependability:** This makes it difficult to trust LLMs for fully autonomous problem-solving.

We present a **Multi-Agent Reinforcement Learning (MARL)** framework for collaborative reasoning, where a network of agents learns to trust, weight, and integrate each other's contributions to reach more robust consensus than a single LLM.

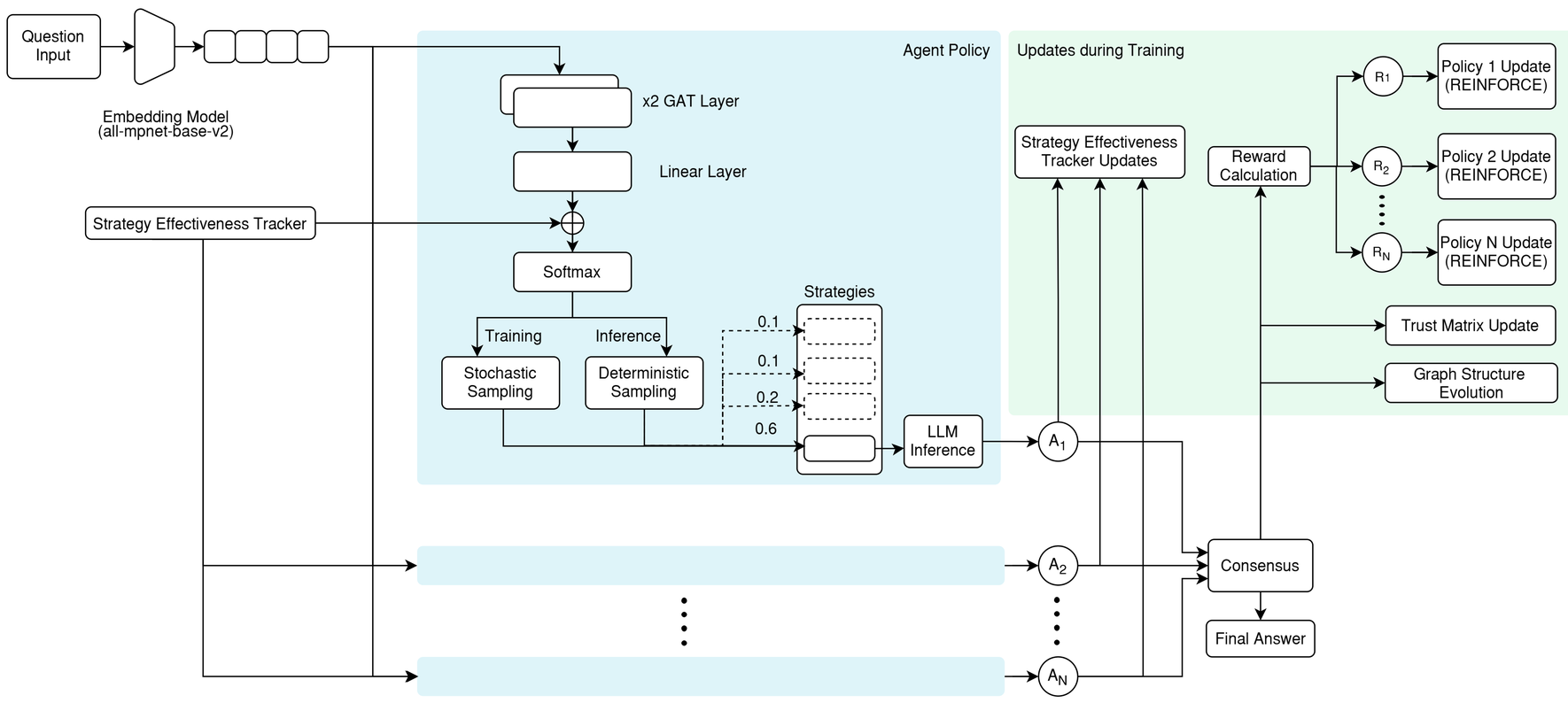
Key Contributions

- We apply a trust-aware multi-agent reinforcement learning framework that enables collaborative reasoning among multiple LLM agents, using a Graph Attention Network (GAT) to model trust and prioritize reliable contributions.
- We introduce adaptive strategy selection and reward mechanisms that balance individual performance, consensus alignment, and reasoning diversity to improve decision-making and exploration.
- Our approach strengthens collaboration, yielding consistent improvements in consensus accuracy across model sizes and benchmarks, independent of the underlying architecture.

Datasets Used

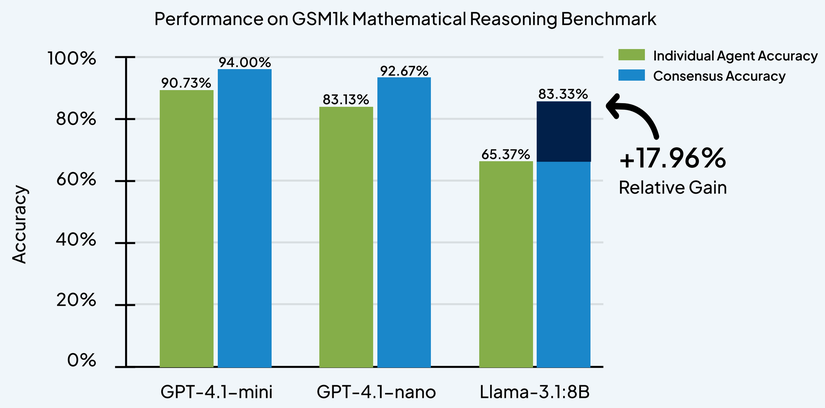
Dataset	Type	Description	Usage
ARC-Challenge	Science	Grade-School reasoning; includes causal, comparative, hypothetical questions	Training & Inference
GSM8k	Mathematics	Grade-School math problems with solutions; arithmetic, algebra, multi-step	Training
GSM1k	Mathematics	Held-out math problems similar to GSM8K for robustness testing	Inference

System Architecture

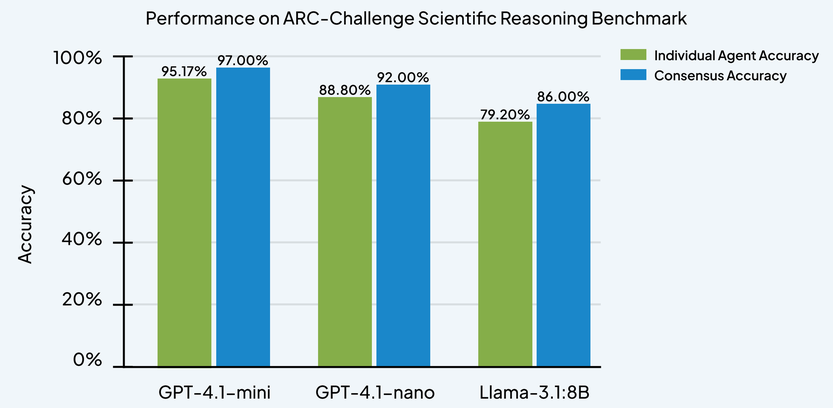


Experiments

Collaborative Reasoning Delivers Substantial Gains, Especially for Smaller Models



The Framework Also Improves Stability in Scientific and Commonsense Reasoning



Benchmark Comparison

GSM1K

Model	Accuracy
Ours MARL-GAT (GPT-4.1-mini)	**94.00%**
GPT-4.1-mini †	93.33%
GPT-4o[1]	92.90%
Ours MARL-GAT (GPT-4.1-nano)	**92.67%**
GPT-4[1]	92.30%
GPT-4.1-nano †	90.00%
Ours MARL-GAT (Llama-3.1 8B)	**83.30%**
Llama-3.1:8B †	77.33%
Llama-3.1:8B[1]	69.0%

ARC-Challenge

Model	Accuracy
Ours MARL-GAT (GPT-4.1-mini)	**97.00%**
Llama-3.1:405B[3]	96.90%
GPT-4[2]	96.40%
GPT-4.1-mini †	94.33%
Ours MARL-GAT (GPT-4.1-nano)	**92.00%**
GPT-4.1-nano †	91.67%
Ours MARL-GAT (Llama-3.1 8B)	**86.00%**
Llama-3.1:8B †	84.67%
Llama-3.1:8B[3]	83.40%

1- Zhang et al. 2024 | 2 - HyperAI 2025 | 3 - Grattafiori et al. 2024 | † - Our zero-shot experiments

Ablation Studies Confirm Our Core Design Choices

Reinforcement Learning Outperforms Supervised Learning

Finding
Reward-driven adaptation enhances coordination and leads to higher consensus accuracy.

Reinforcement Learning Consensus: 92.67%

Supervised Learning Consensus: 92.00%

Simple Individual Rewards Are More Effective

Finding
A direct correctness signal (+1.0 / -0.5) provides a more stable learning objective than complex consensus-aware rewards.

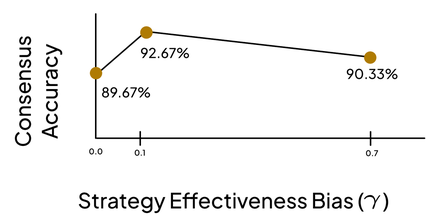
Base Reward consistently achieved higher consensus accuracy (e.g., 94.00% vs 93.67% for GPT-4.1-mini on GSM1K),

$$R_{\text{modified}} = R_{\text{individual}} + \Delta R,$$

$$\Delta R = \begin{cases} +\delta^{+}, & \text{if consensus is incorrect and agent is correct,} \\ -\delta^{-}, & \text{if consensus is incorrect and agent is incorrect,} \\ +\delta^{\text{minor}}, & \text{if consensus is correct and agent is in the correct minority,} \\ 0, & \text{otherwise.} \end{cases}$$

A Moderate Bias Towards Proven Strategies is Optimal

Finding
A small bias ($\gamma = 0.1$) balances exploiting known effective strategies with exploring new ones, maximizing consensus, while a higher bias ($\gamma = 0.7$) limits exploration and no bias ($\gamma = 0.0$) yields the lowest accuracy.



Methodology

Overall Pipeline

Questions are embedded, agents choose strategies via a GAT-based policy, generate answers with LLMs, and a trust-weighted consensus selects the final output.

GAT Policy Network

$$\pi = (\mathbf{W} \cdot \mathrm{GAT}_2(\mathrm{GAT}_1(\mathbf{X}, \mathbf{E}), \mathbf{E}))$$

$\mathbf{X}$ is the node feature matrix, $\mathbf{E}$ is the edge index tensor, $\mathbf{W}$ is the output weight matrix, and π is the logits obtained from policy network, also referred to as GAT_{output}.

Strategy Effectiveness

A global tracker adaptively biases agents toward historically successful strategies by retrieving similar past outcomes and updating effectiveness via an exponential moving average.

$$S_{s,q}^{(t+1)} = (1-\alpha) S_{s,q}^{(t)} + \alpha \cdot c_{s,q}^{(t)}$$

$S_{s,q}^{(t)}$ denotes the effectiveness score for strategy s on question embedding q at time t. $c_{s,q}^{(t)}$ is the binary correctness indicator (1 if correct, 0 if incorrect). α is EMA smoothing factor.

Strategy Selection

Agents use GAT outputs plus effectiveness bias to sample diverse reasoning strategies tailored to the question and graph context.

$$P(\text{strategy}) = \text{softmax}\left(\text{GAT}_{\text{output}} + \gamma \times S_{s,q}^{(t)}\right)$$

γ is a tunable parameter controlling the influence of effectiveness bias $S_{s,q}^{(t)}$.

Consensus

A trust-weighted consensus aggregates agent answers, giving higher influence to reliable agents while preserving diverse inputs.

For each candidate answer, the total score is computed as the sum of weights of agents selecting that answer

$$S_{\text{answer}} = \sum_{\text{agent} \in P_{\text{answer}}} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} T_{ij},$$

$\mathcal{N}_i = \{j \mid T_{ij} > 0, j \neq i\}$ denotes the set of agents trusted by agent i. The final consensus is determined by applying a softmax function over these scores to produce a probability distribution:

$$P(\text{answer}_i) = \frac{\exp(S_{\text{answer}_i})}{\sum_j \exp(S_{\text{answer}_j})}$$

P_{answer} denotes the set of agents selecting that answer. The answer with the highest probability is selected as the consensus answer.

Rewards

$$R_{\text{individual}} = \begin{cases} +1.0, & \text{if correct,} \\ -0.5, & \text{if incorrect.} \end{cases}$$

References

[1] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks," 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[2] OpenAI, "Gpt api," https://platform.openai.com/, 2023, accessed: 2025-08-18.

[3] A. Grattafiori, A. Dubey, A. Jauhri, and et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[4] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," 2021. [Online]. Available: https://arxiv.org/abs/1911.10635

[5] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, 2018, pp. 1955-1967.

[6] H. Zhang, J. Da, D. Lee, V. Robinson, C. Wu, W. Song, T. Zhao, P. Raja, C. Zhuang, D. Slack, Q. Lyu, S. Hendryx, R. Kaplan, M. Lunati, and S. Yue, "A careful examination of large language model performance on grade school arithmetic," 2024. [Online]. Available: https://arxiv.org/abs/2405.00332

Conclusion

Effective but costlier
Improves robustness across tasks, but increases compute use and inherits biases from base models.

Multi-agent collaboration boosts reasoning
Coordinated agents outperform single models through shared decision-making.

Adaptive trust networks
Agents learn who to trust, strengthening links with reliable collaborators over time.

Larger performance gains for smaller models
Collective reasoning provides large performance gains for weaker models while stabilizing stronger ones.