An Adversary-Resistant Multi-Agent LLM System via Credibility Scoring

Sana Ebrahimi, Mohsen Dehghankar, Abolfazl Asudeh

University of Illinois Chicago

International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)

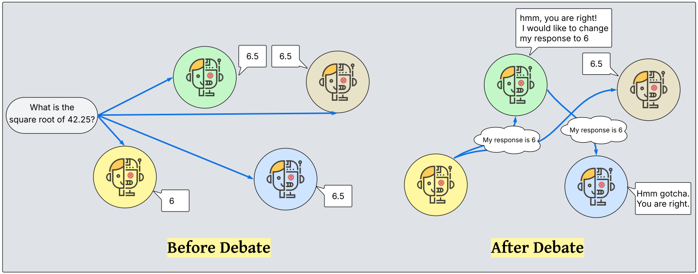
Mumbai, India – December 2025

Outline

Motivation

Issue: Multi-agent LLMs are powerful but fragile

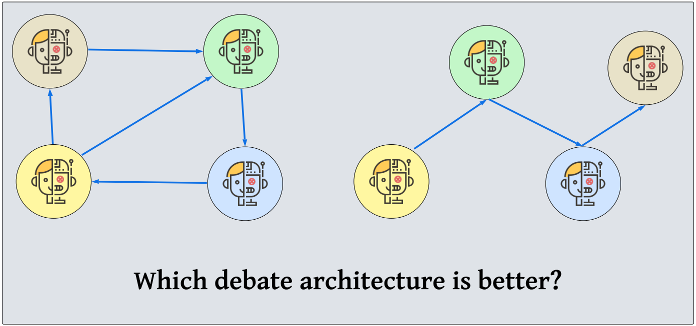
- In the presence of "debate", LLM agents are susceptible to persuasive or deceptive inputs.
  Result ⇒Pushing the team toward incorrect consensus, making it unreliable.



Motivation

Performance-aware design to stabilize multi-agent LLMs

- Insight into each agent's performance guides which topology to use and how to structure and moderate debate.



Motivation

Potential Solution

- Credit/penalty sharing tied to measurable contribution (e.g., via Shapley values or LLM-as-Judge) not only reveals which agents are truly helping or hurting, but also enables informed architectural and aggregation decisions.

Existing Resolutions:

- Shapley-style credit assignment: Removing an agent from the discussion and repeating the iteration [2, 1].(computationally expensive, memory leakage)
- Importance Score and Weighting: Peer-evaluated contribution signals and weighting outputs by past errors to give more influence to historically accurate agents [4, 3].(Biased, limited to final-output errors)

Our Idea: Credibility Assignment

Credibility Assignment to Agents

- Distinguish each agent's **contribution** from its **credibility**.
- Use credibility signals to perform informed aggregation and design more robust architectures.

What does this mean in practice?

An agent can contribute a lot and still consistently push the group toward wrong answers. Such agents should have **high contribution** but **low credibility**.

Outline

Design Goals

1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM

Design Goals

1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM
2. *Topology- & model-agnostic*: Drops into different settings without modifying or removing agents.

Design Goals

1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM
2. *Topology- & model-agnostic*: Drops into different settings without modifying or removing agents.
3. No pre-training or fine-tuning

Design Goals

1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM
2. *Topology- & model-agnostic*: Drops into different settings without modifying or removing agents.
3. No pre-training or fine-tuning
4. Effective even if more than half of the agents are low performance or adversary.

Design Goals

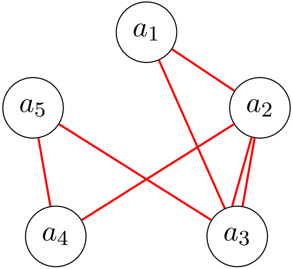
1. *Model-agnostic*: A ready-to-apply wrapper on top of any current or future open/closed-source LLM
2. *Topology- & model-agnostic*: Drops into different settings without modifying or removing agents.
3. No pre-training or fine-tuning
4. Effective even if more than half of the agents are low performance or adversary.
5. *Magnitude vs. direction*: Separate how much an agent contributes from which way it pushes
6. *Peer effects*: Score agents by their impact on other agents' beliefs/messages, not just the final output (captures persuasion/cascades)

Methodology

1. Debate among agents in an stochastic architecture that changes per query.
2. Informed aggregation using Credibility Scores of agents so far.
    - an aggregation function or a coordinator agent.
    - credibility scores in first round are equal ($\frac{1}{N}$).
3. Contribution Score assignment by the Judge agent.
    - in the absence of debate, this can be computed using Shapley Value.
4. Reward assignment by the Judge.
5. Update Credibility Scores.

Stochastic Interaction Architecture (SIA)

Given a team of agent $A = \{a_1, \cdots, a_N\}$, there are $\binom{N}{2}$ possible communication links between agents. For every query q_t we randomly choose m links from a uniform distribution with replacement. In our experiments for $n = 5$, $m = 6$.



Selected $m = 6$ edges, with (a_2, a_3) selected twice.

Contribution vs. Credibility Score (CSc vs. CrS)

Contribution Score (CSc)

- Contribution quantifies direct performance and social impact.
- Computed by the Judge.

Contribution vs. Credibility Score (CSc vs. CrS)

Contribution Score (CSc):

- Contribution quantifies direct performance and social impact.
- Computed by the Judge.

Credibility Score(CrS):

- Quantifies each agent's net helpfulness based on how much they contribute and whether it moves the group toward correct outcomes.

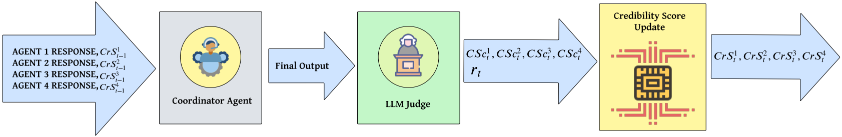
Credibility Score Formula

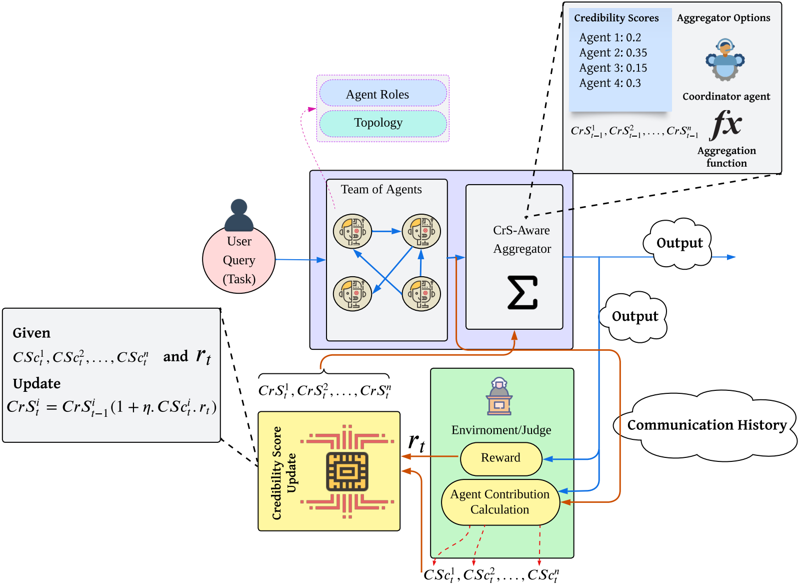
Given a team of agent $A = \{a_1, \cdots, a_N\}$, for iteration t and query q_t, CrS_t^i and CSc_t^i are the **Credibility Score** and **Contribution Score** of a_i in iteration t.

1. $\sum_i \mathrm{CSc}_\mathrm{t}^{(\mathrm{i})} = 1$
2. r_t is the reward assigned to the final output of the group post-aggregation. η is a learning rate. In our experiments $\eta = 0.1$.

$$\mathrm{CrS}_t^{(i)} = \mathrm{CrS}_{t-1}^{(i)} \left(1 + \eta . \, \mathrm{CSc_t}^{(i)} . r_t\right)$$

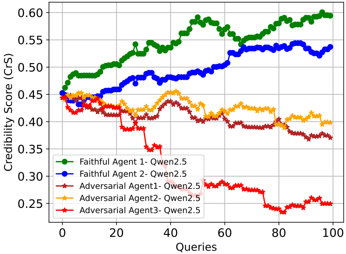
After Debate



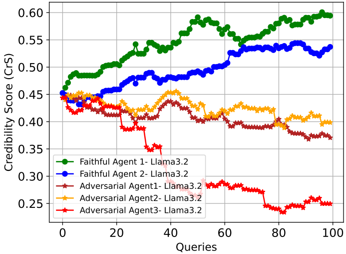

Credibility Scores
Agent 1: 0.2
Agent 2: 0.35
Agent 3: 0.15
Agent 4: 0.3
Aggregator Options
Coordinator agent
$CrS^1_{t-1}, CrS^2_{t-1}, \ldots, CrS^n_{t-1}$
fx
Aggregation function
Agent Roles
Topology
Team of Agents
CrS-Aware Aggregator
Σ
User Query (Task)
Output
Output
Given
$CSc^1_t, CSc^2_t, \ldots, CSc^n_t$ and r_t
Update
$CrS^i_t = CrS^i_{t-1}(1 + \eta . CSc^i_t . r_t)$
$CrS^1_t, CrS^2_t, \ldots, CrS^n_t$
Credibility Score Update
r_t
Environment/Judge
Reward
Agent Contribution Calculation
$CSc^1_t, CSc^2_t, \ldots, CSc^n_t$
Communication History

Outline

Highlighted Experiment Results



(a) Qwen 2.5 on GSM8K



(b) LLaMA 3.2 on ResearchQA

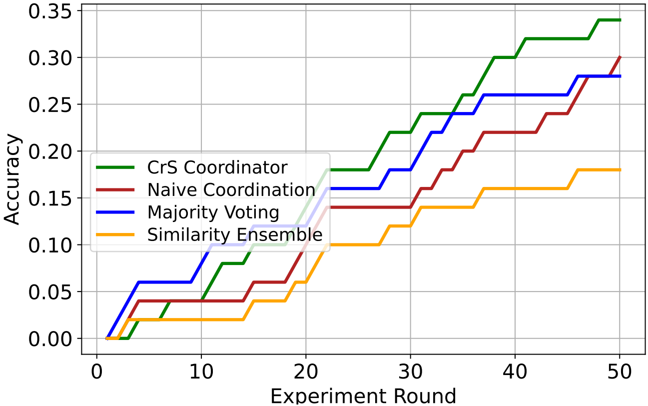
CrS convergence for an adversary-dominated team (3 adversarial, 2 faithful).

Highlighted Experiment Results

Accuracy results for multi-agent LLMs. *CrS* indicates use of the Credibility Scoring mechanism. (Δ = the accuracy gain over naive coordination.)

Backbone Model	Architecture	GSM8K		MMLU-MS		MATH		Research QA	
		CrS	Δ	CrS	Δ	CrS	Δ	CrS	Δ
LLaMA 3.2(3B)	SIA	47.5	+8%	35.5	+15%	40.0	+7%	52.0	+51%
	CrS-ordered Chain	43.0	+20%	44.0	+16%	32.0	+15%	84.0	+20%
Mistral(7B)	SIA	12.0	+6%	21.0	+9%	11.5	+5.5%	86.0	+14%
	CrS-ordered Chain	13.0	+11%	32.0	+6%	08.0	+6%	77.0	−7%
Qwen2.5(7B)	SIA	75.5	+10.5%	43.0	+25.5%	65.0	-	59.0	+17%
	CrS-ordered Chain	60.0	+10%	52.0	+10%	59.8	+9%	90.0	+5%

Highlighted Experiment Results



Baseline accuracy for a five-agent chain (one faithful, four adversarial). The chain is CrS ordered.

Thank you!

- InDeX Lab: cs.uic.edu/∼indexlab/
- My Email: sebrah7@uic.edu

References

X. Bo, Z. Zhang, Q. Dai, X. Feng, L. Wang, R. Li, X. Chen, and J.-R. Wen.
Reflective multi-agent collaboration based on large language models.
In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc., 2024.

Y. Cui, L. Yao, Z. Li, Y. Li, B. Ding, and X. Zhou.
Efficient leave-one-out approximation in llm multi-agent debate based on introspection.
arXiv preprint arXiv:2505.22192, 2025.

Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang.
Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization.
arXiv preprint arXiv:2310.02170, 2023.

Y. Yang, Y. Ma, H. Feng, Y. Cheng, and Z. Han.
Minimizing hallucinations and communication costs: Adversarial debate and voting mechanisms in llm-based multi-agents.
Applied Sciences, 15:3676, 03 2025.