ICML
International Conference
On Machine Learning

TMLR
TRUSTWORTHY MACHINE LEARNING AND REASONING

From *Debate* to *Equilibrium*:

Belief-Driven Multi-Agent LLM Reasoning via Bayesian Nash Equilibrium

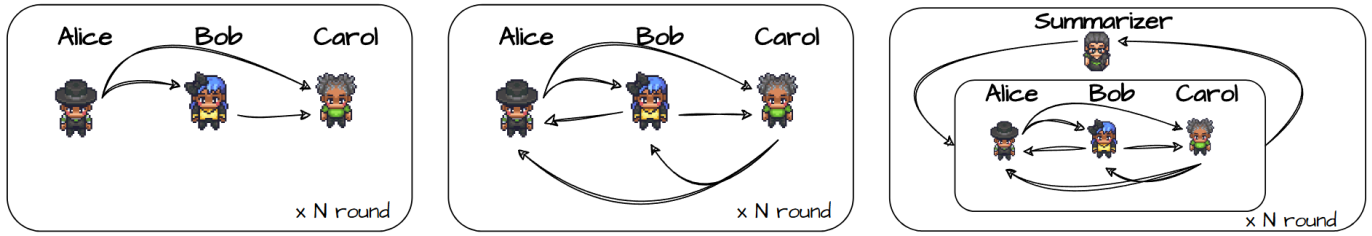
Yi Xie
Fudan University

with Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu and Bo Han

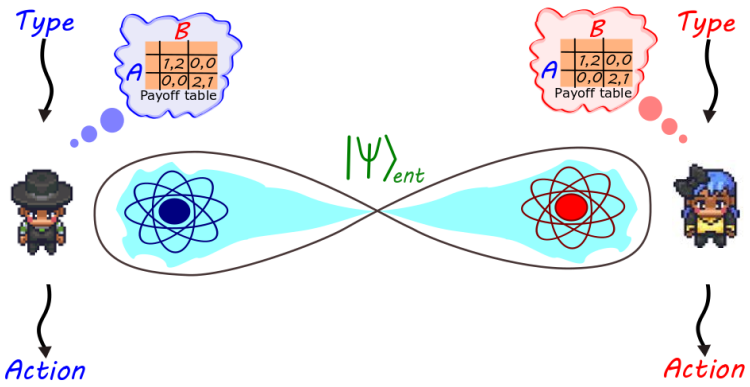
Email: 22210860116@m.fudan.edu.cn
Paper: https://arxiv.org/abs/2506.08292
Code: https://github.com/tmlr-group/ECON.

Main Contributions

Multi Agent Debate → **Formalize** → **Bayesian Game**



* (a) One-by-One (b) Simultaneous-Talk (c) Simultaneous-Talk-with-Summarizer



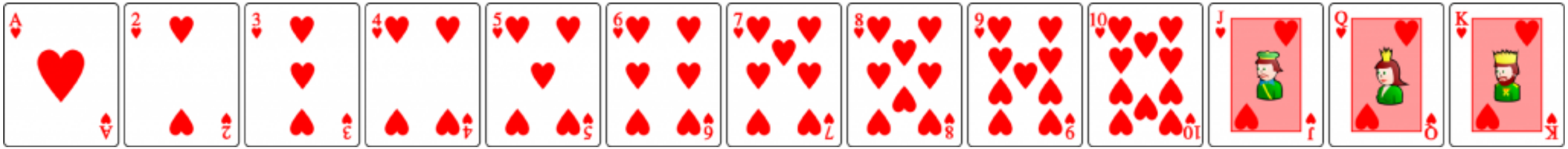
Main Contribution:

1. We formalize Bayesian Nash Equilibrium for multi-agent LLM systems
2. We introduce ECON to implement BNE via belief-based coordination
3. ECON outperforms both existing single-agent and multi-agent approaches, and validate its efficiency to scale to larger ensembles

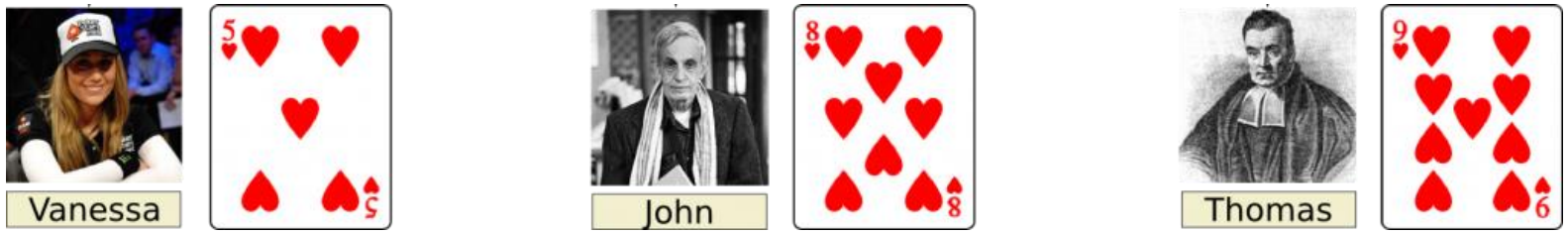
* ChatEval: https://arxiv.org/abs/2308.07201

Background | Highest Card Game

*



Rule: **If you think you get the highest value card, you should say yes.**



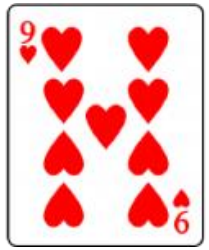
Type (private information)

* https://www.science4all.org/article/bayesian-games-how-to-model-poker/

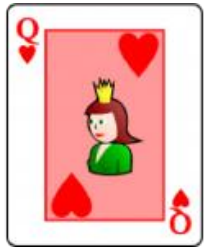
Background | Highest Card Game

Type:

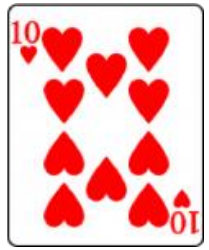

Vanessa




John




Thomas



Belief:

(8/12) x (7/11) ≈ 0.42

(11/12) x (10/11) ≈ 0.83

(9/12) x (8/11) ≈ 0.55

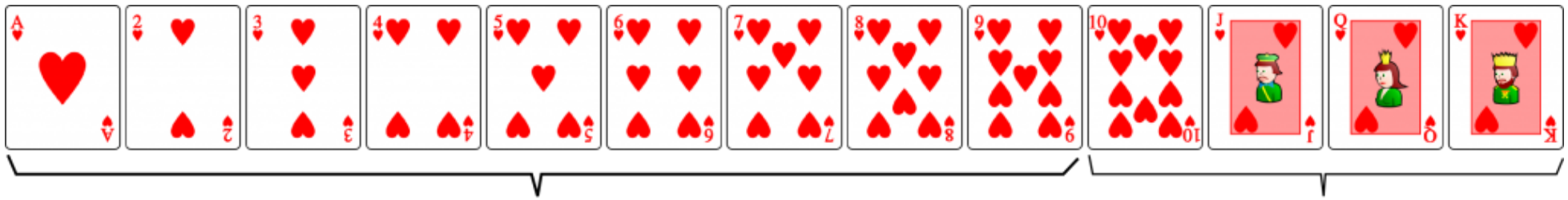
Action:

NO

YES

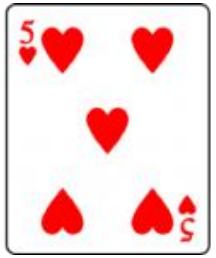
YES

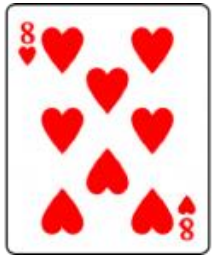
Strategy:

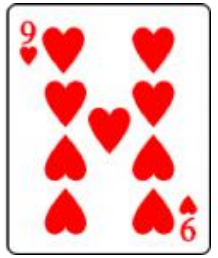


Background | Sequential Highest Card Game

Type:







Sequence:

Action: **NO**

Belief update:
Vanessa <=9

$(7/8)x(6/11) \approx 0.48$

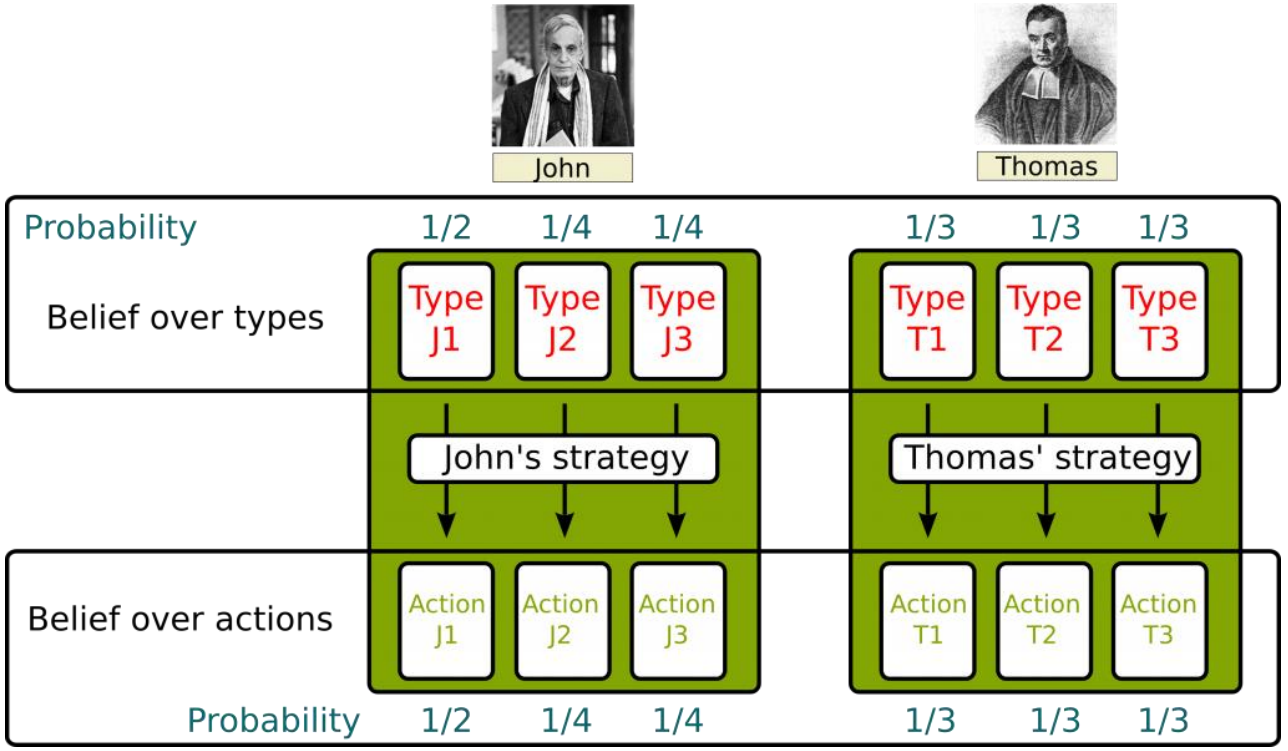
Action: **NO**

Belief update:

Vanessa <=9;
John <=9

Action: **YES**

Background | Sequential Highest Card Game

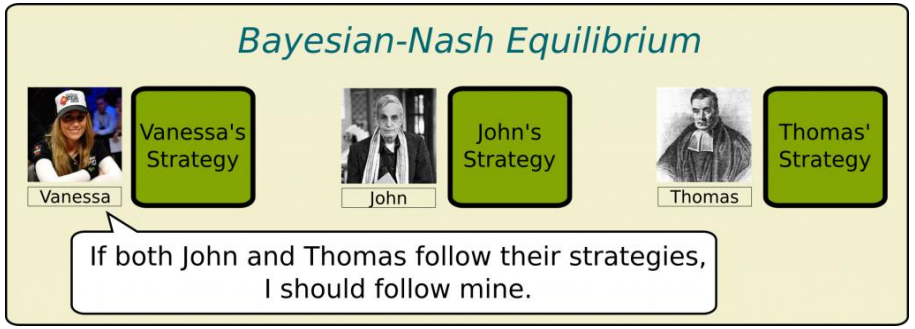


BNE:

1. strategy profile

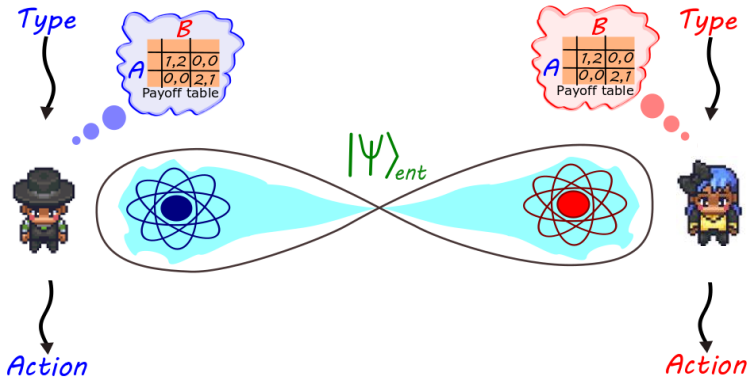
2. a best-response strategy to best-response strategies

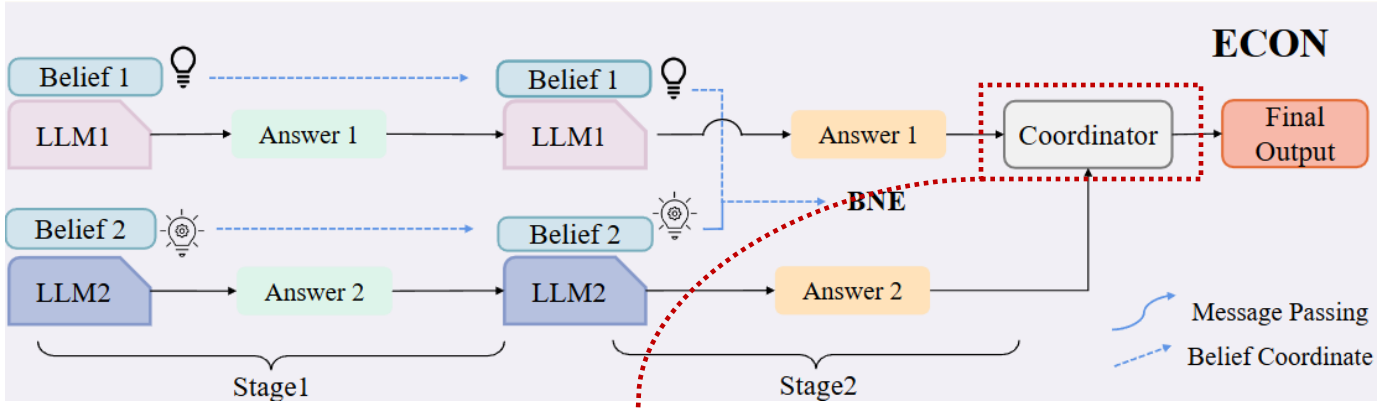
Back to Vanessa:



Background | Bayesian Nash Equilibrium

Incomplete Information (what we need)

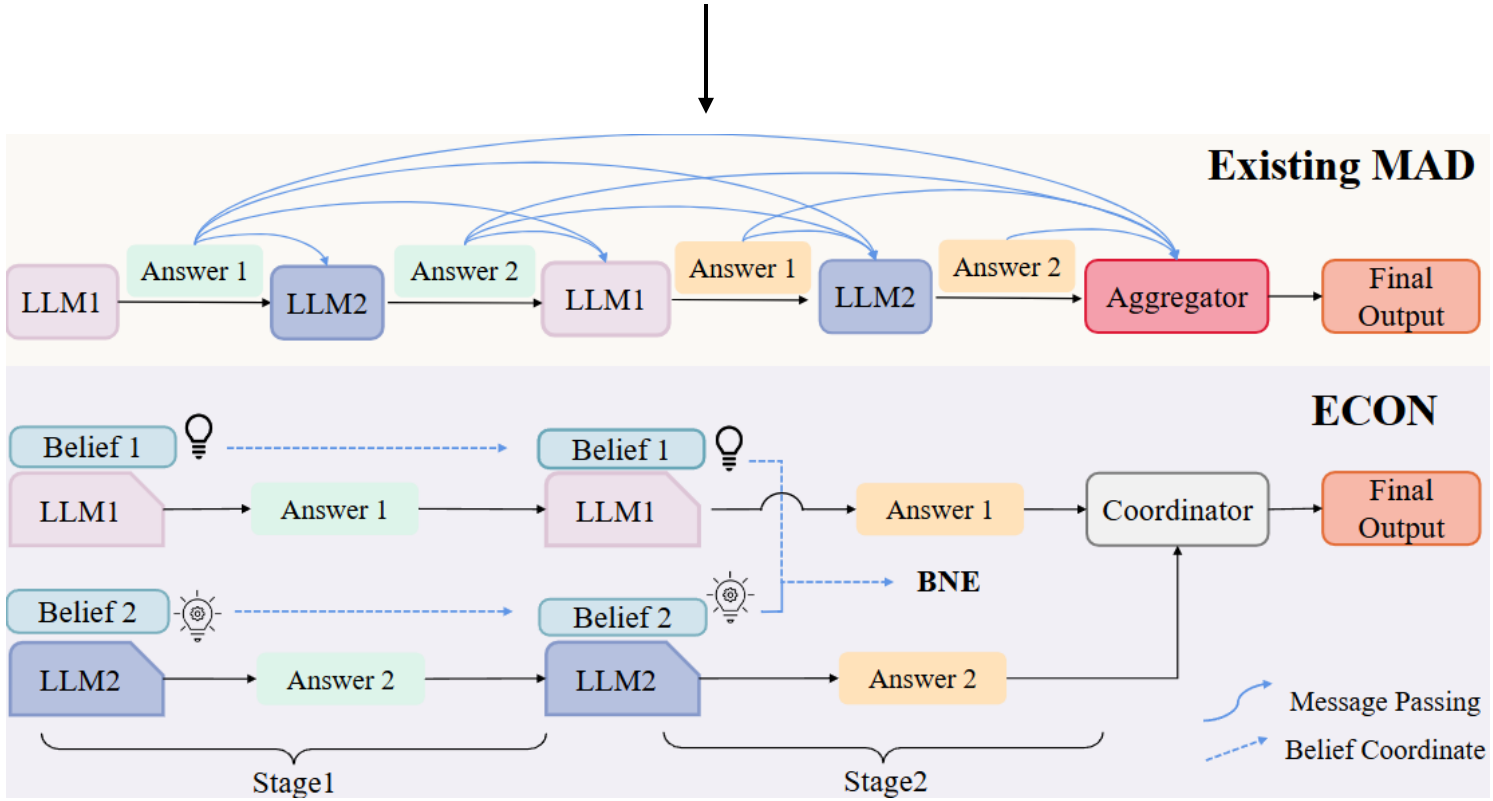




No rule for LLM output, so we need a coordinator

Method | Overview

Find a BNE in MA LLM reasoning process :



- **Existence:**

Does a BNE even exist in this complex system? (optimization target)

- **Convergence:**

How to evaluate? Can our learning algorithm reliably guide towards this BNE?

Method | Existence of BNE

Satisfy the conditions of Glicksberg's Fixed-Point Theorem for BNE exists

↓

Glicksberg's Conditions

- Strategy space is compact and convex.
- Payoff function is continuous.
- Payoff function is quasi-concave

→

ECON's Implementation

- Action space
- Reward Function Design
- Centralized Mixing Network

Method | Bayesian Regret

How do we measure convergence to this equilibrium?

Regret measures the cumulative performance loss of a learning agent compared to the optimal BNE strategy over time. It quantifies "how much better" the agent could have performed.

Formal Definition (Total Regret over T steps):

$$R(T) = \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=1}^{T} (V_i^*(s_t) - V_i^{\pi_t}(s_t))\right]$$

Sublinear Regret : The average regret approaches zero. This implies the agent's policy is **converging** to the optimal BNE. **(Our Goal)**

Method | Convergence

To achieve sublinear regret we need:

Theoretical Assumption

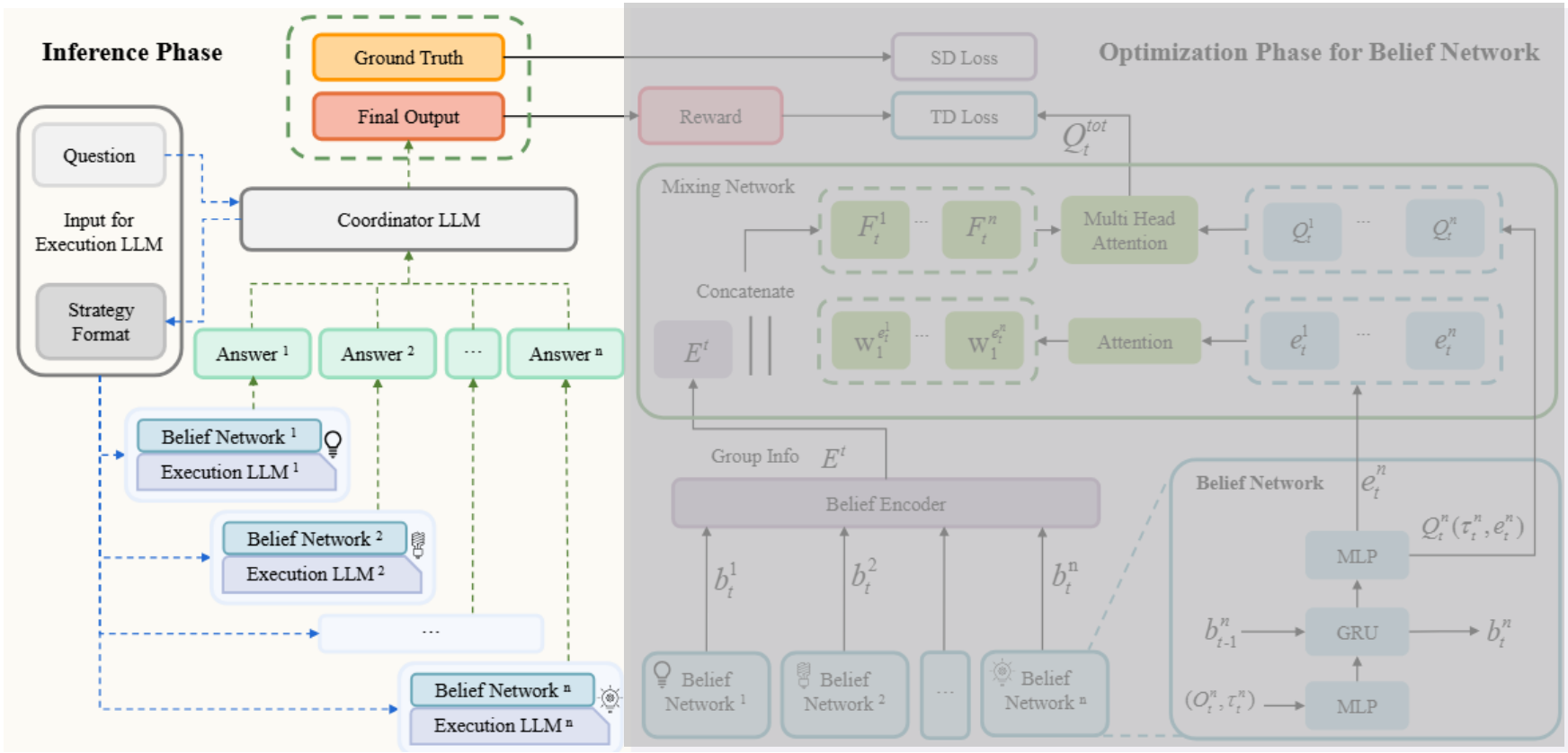
- Bounded Rewards
- Approximate Posterior Alignment
- Game Regularity
- Concentrability

ECON's Implementation

- Reward Function Design
- Belief Encoder
- Belief Networks & Soft Update
- Experience Replay Buffer

Method | Inference Stage

Inference with no direct communication.



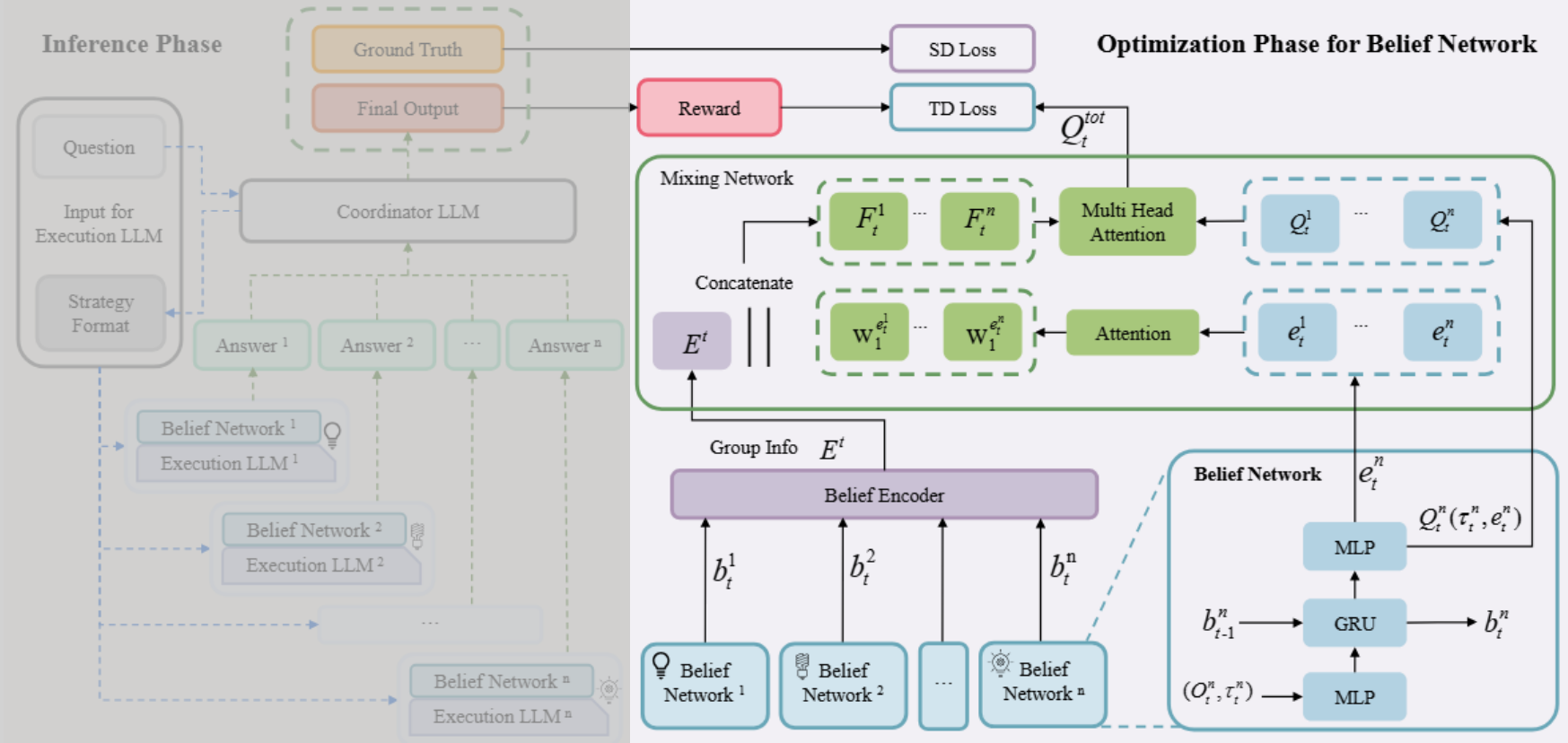
Guidance:
The **Coordinator LLM** receives the *Question* and generates a high-level *Strategy and Format*.

Independent Reasoning:
Each **Execution LLM**, guided by its *Belief Network*, takes this guidance and independently produces an Answer.

Aggregation:
The *Coordinator LLM* collects all Answers and synthesizes the Final Output.

Method | Optimization Stage

Optimization Phase is a top-down execution flow to approach BNE.



Local Beliefs to Global Representation:
Local Belief are fed into a shared **Belief Encoder**, creating a global group representation

Value Decomposition:
A central **Mixing Network** takes the local Q-values with the global information to compute a total, global Q-value

Loss-driven Updates: Based on the final Reward and the global Q-value, losses are calculated to update all components.

Experiments | Major results

	Validation (#180)						Test (#1,000)					
	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro			Micro	Macro	Micro	Macro	
Greedy Search	100	74.4	0	60.8	37.8	0	100	72.0	0	52.4	31.8	0
Two-stage												
Mixtral-8x7B-MoE	49.4	30.0	0	1.2	0.6	0	51.2	32.2	0.2	0.7	0.4	0
Gemini Pro	28.9	18.9	0	0.5	0.6	0	39.1	24.9	0	0.6	0.1	0
GPT-3.5-Turbo	86.7	54.0	0	0	0	0	91.8	57.9	0	0.5	0.6	0
GPT-4-Turbo	89.4	61.1	2.8	15.2	10.6	0.6	93.1	63.3	2.0	10.5	5.5	0.6
Debate (GPT-4) @3round	95.2	67.3	6.7	22.7	13.1	2.3	97.8	72.4	11.3	17.4	12.1	3.7
ECON (GPT-4)	**100**	**71.4**	**15.6**	**32.1**	**25.7**	**7.2**	**100**	**82.1**	**26.6**	**32.4**	**17.6**	**9.3**
Sole-planning												
$\text{Direct}_{\text{GPT-3.5-Turbo}}$	100	60.2	4.4	11.0	2.8	0	100	59.5	2.7	9.5	4.4	0.6
$\text{CoT}_{\text{GPT-3.5-Turbo}}$	100	66.3	3.3	11.9	5.0	0	100	64.4	2.3	9.8	3.8	0.4
$\text{ReAct}_{\text{GPT-3.5-Turbo}}$	82.2	47.6	3.9	11.4	6.7	0.6	81.6	45.9	2.5	10.7	3.1	0.7
$\text{Reflexion}_{\text{GPT-3.5-Turbo}}$	93.9	53.8	2.8	11.0	2.8	0	92.1	52.1	2.2	9.9	3.8	0.6
$\text{Direct}_{\text{Mixtral-8x7B-MoE}}$	100	68.1	5.0	3.3	1.1	0	99.3	67.0	3.7	3.9	1.6	0.7
$\text{Direct}_{\text{Gemini Pro}}$	93.9	65.0	8.3	9.3	4.4	0.6	93.7	64.7	7.9	10.6	4.7	2.1
$\text{Direct}_{\text{GPT-4-Turbo}}$	**100**	80.4	17.2	47.1	22.2	4.4	**100**	80.6	15.2	44.3	23.1	4.4
Debate (GPT-4)	97.7	78.9	15.6	43.3	20.6	6.7	98.2	79.5	18.8	41.7	22.9	7.1
ECON (GPT-4)	**100**	**83.3**	**22.2**	**51.7**	**27.8**	**12.9**	**100**	**84.2**	**23.5**	**49.8**	**28.7**	**15.2**

Observation:

SOTA Performance, ECON outperforms strong baselines across 6 diverse reasoning and planning benchmarks. Especially in **Complex Planning,** more than **doubles** the final pass rate on TravelPlanner vs. 3-round debate (9.3% vs 3.7%).

Experiments | Heterogeneous results

Method	GSM-Hard	MATH
Baselines		
ECON	51.43	81.47
LLaMA 3.1 7B (Few-shot CoT)	42.23	62.71
ECON Configurations		
Homo. (3× LLaMA3.1 8B)	48.71	67.70
Homo. (3× LLaMA3.1 405B)	61.29	89.24
Hetero. (LLaMA3.1 8B, LLaMA3 8B, Mixtral 7B)	45.24	74.24
Hetero. (Mixtral 8×22B, Qwen1.5 110B, LLaMA3.1 405B)	55.73	85.46

Observation: when using a **homogeneous** set of agents, stronger models deliver better results.
A mix of different models remains effective and robustly outperforms baselines.

Experiments | Consumption

Dataset	Inference Strategy	LLaMA3.1 70B		Mixtral 8x7b		Mixtral 8x22b	
		Token Usage	Performance	Token Usage	Performance	Token Usage	Performance
MATH	Multi-Agent Debate (3 rounds)	2154.87	71.58	1462.12	31.28	5345.56	67.41
	RAP	2653.27	68.71	1737.73	33.99	6668.55	62.53
	ECON (with detailed strategy)	3270.06	72.38	2150.23	26.18	8054.03	68.23
	Self Consistency (64 rounds)	11917.00	67.39	8066.21	31.58	29616.13	62.21
	ECON	1629.79	81.47	1128.23	35.02	4270.86	72.29
GSM8K	Multi-Agent Debate (3 rounds)	1391.57	86.32	1463.40	70.19	5714.05	81.95
	RAP	1907.86	81.33	1248.66	72.03	6517.77	76.97
	ECON (with detailed strategy)	2772.24	85.17	1188.13	65.37	9341.60	81.46
	Self Consistency (64 rounds)	9574.25	89.56	6601.34	71.08	24671.91	86.24
	ECON	1131.65	92.70	1284.98	76.97	4715.31	88.20
GSM-Hard	Multi-Agent Debate (3 rounds)	3030.73	41.98	1478.14	20.04	9250.78	45.21
	RAP	1768.72	38.97	1036.11	22.47	6464.52	42.79
	ECON (with detailed strategy)	3662.64	44.12	2239.07	18.52	11464.98	41.04
	Self Consistency (64 rounds)	16724.69	39.76	11668.19	22.47	74544.25	44.19
	ECON	1518.76	51.43	1271.53	25.76	7101.62	47.58

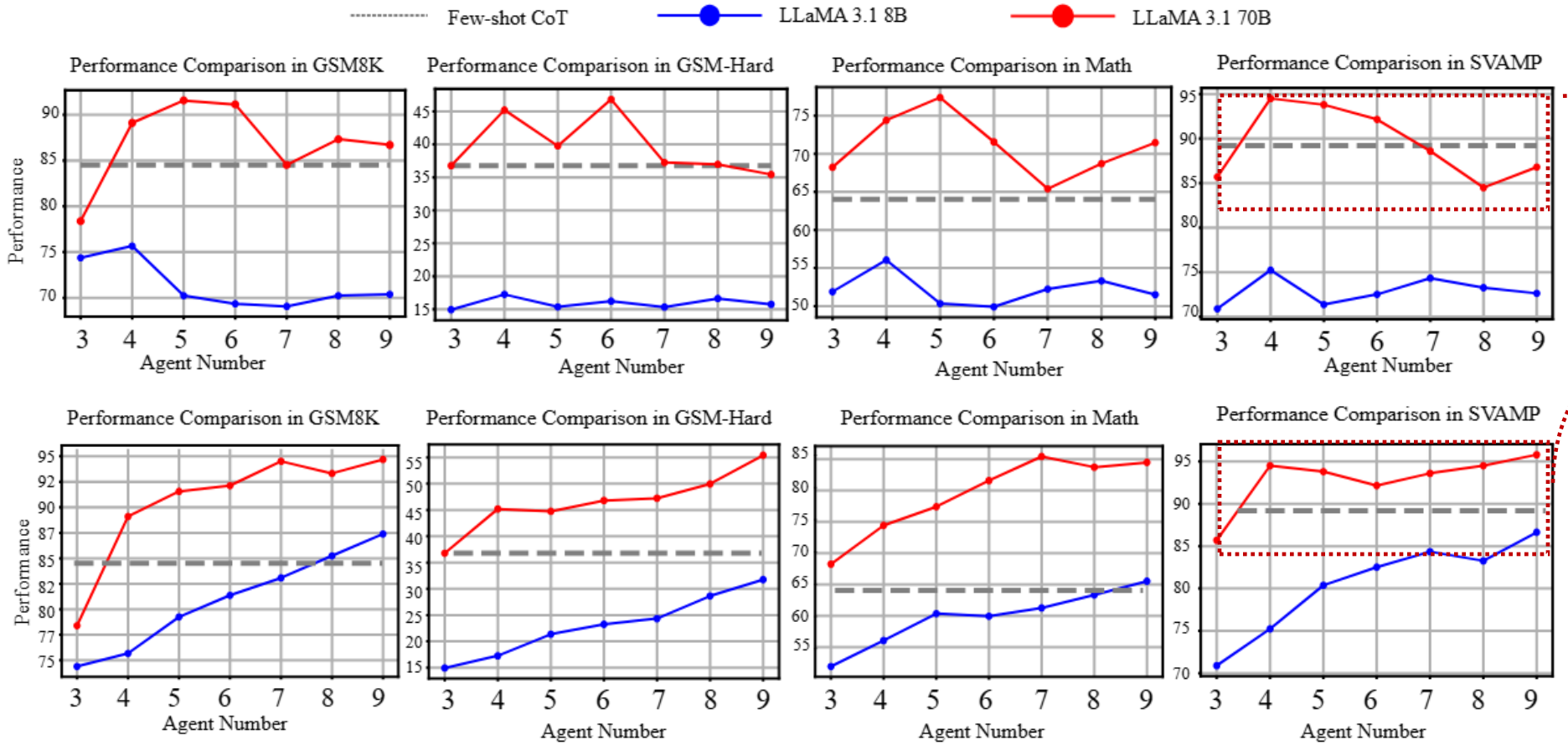
Dataset & Model	Complete Info	Consumption
GSM8K - LLaMA 3.1 8B	81.4	80.3 (+35.6%)
GSM8K - LLaMA 3.1 70B	96.1	96.7 (+42.7%)
GSM-Hard - LLaMA 3.1 8B	30.2	29.9 (+62.3%)
GSM-Hard - LLaMA 3.1 70B	53.6	51.4 (+40.9%)
MATH - LLaMA 3.1 8B	59.6	60.4 (+33.8%)
MATH - LLaMA 3.1 70B	83.1	81.5 (+39.4%)

Observation:

High Efficiency:
Reduces token usage by **21.4%** compared to 3 round debate while achieving better performance.

Incomplete Information is Key:
Full communication increases token cost by **~42%** for minimal gain, validating our core design principle.

Experiments | Scalability



Observation:

Naive Scaling Fails: A single coordinator becomes a bottleneck (top row).

Local-Global Nash Scaling Succeeds: Our hierarchical approach unlocks significant performance gains, achieving an **+18.1%** boost when scaling from 3 to 9 agents (bottom row).

Take-home messages

Summary

1. We successfully **formalized** the multi-agent LLM reasoning problem as a **Bayesian Game**, moving beyond heuristic debate.
2. We introduced **ECON**, a novel framework that practically and efficiently guides agents toward a **Bayesian Nash Equilibrium (BNE)**.
3. Our method is supported by rigorous theory, including proofs for **BNE existence** and a **sublinear regret bound** that guarantees convergence.

Thanks you!

Yi Xie

22210860116@m.fudan.edu.cn