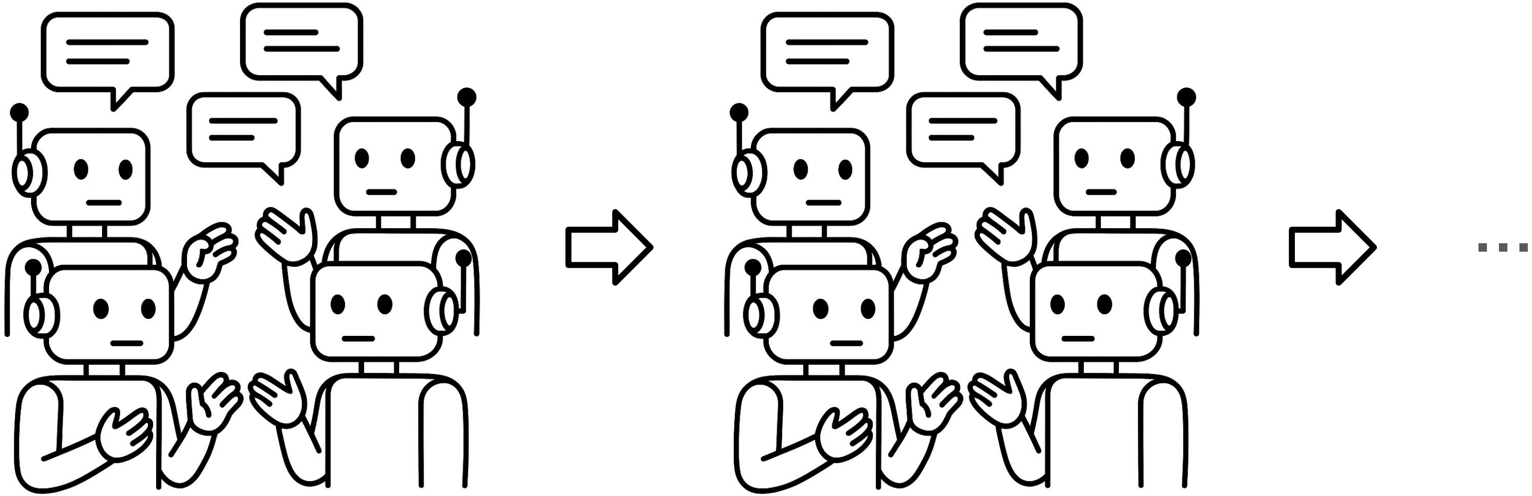
Debate or Vote:
Which Yields Better Decisions in Multi-Agent LLMs?

Hyeong Kyu Choi, Xiaojin Zhu, Sharon Li

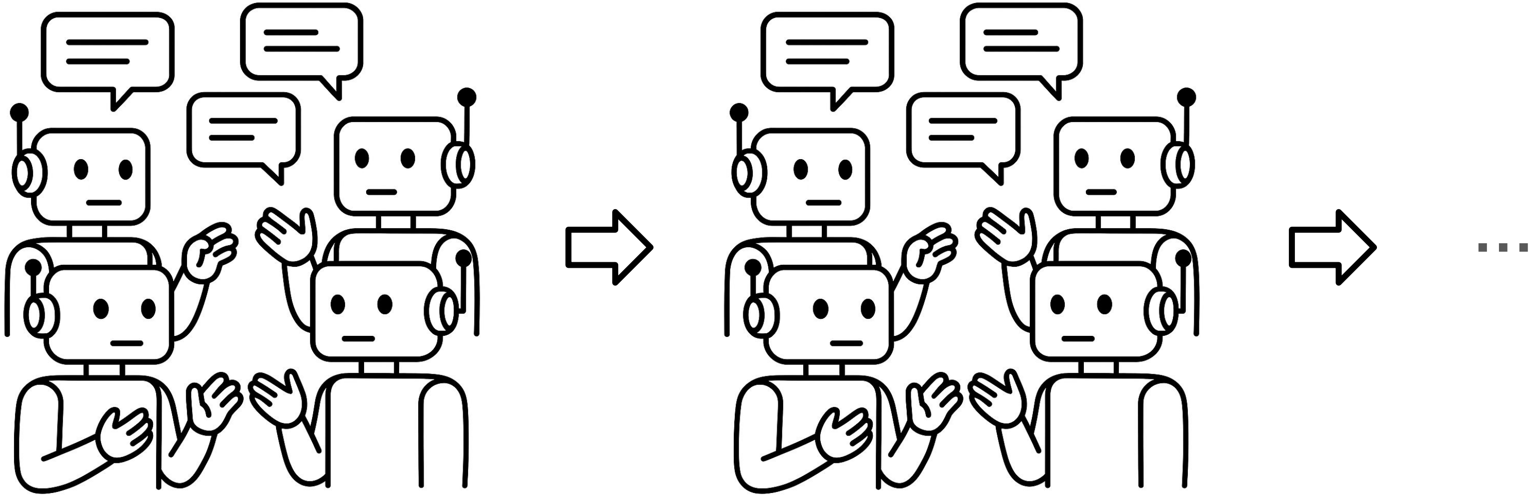
THE UNIVERSITY of WISCONSIN MADISON

NEURAL INFORMATION PROCESSING SYSTEMS

LLM Multi-Agent Debate?

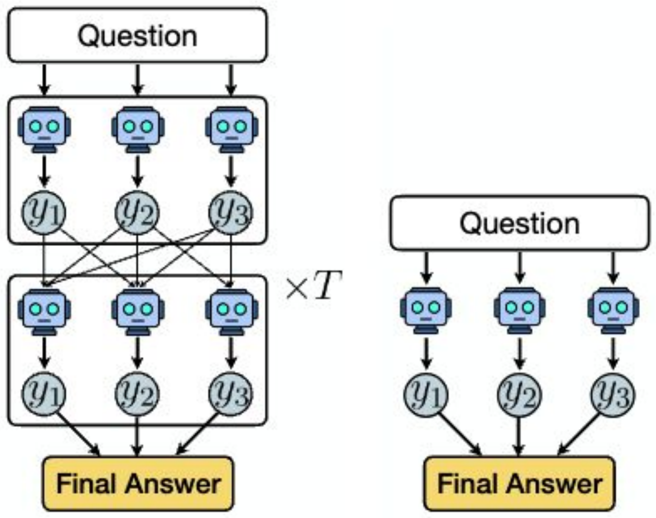


LLM Multi-Agent Debate?



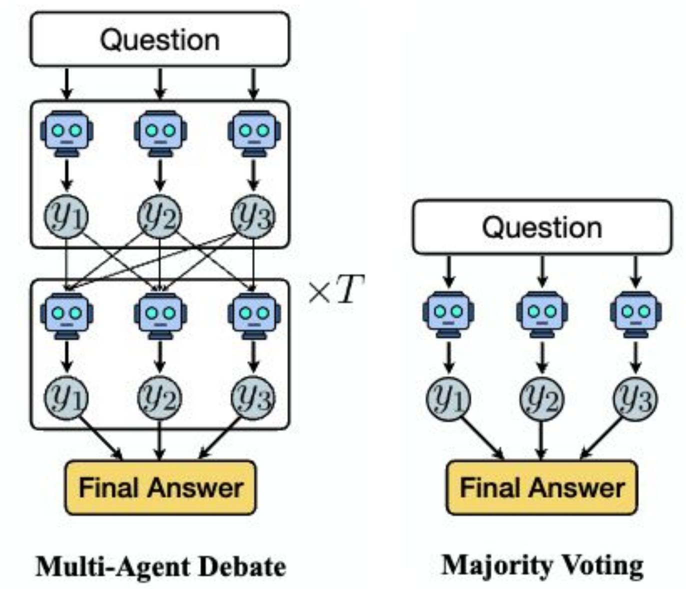
Is MAD meaningfully improving performance through *interaction*?

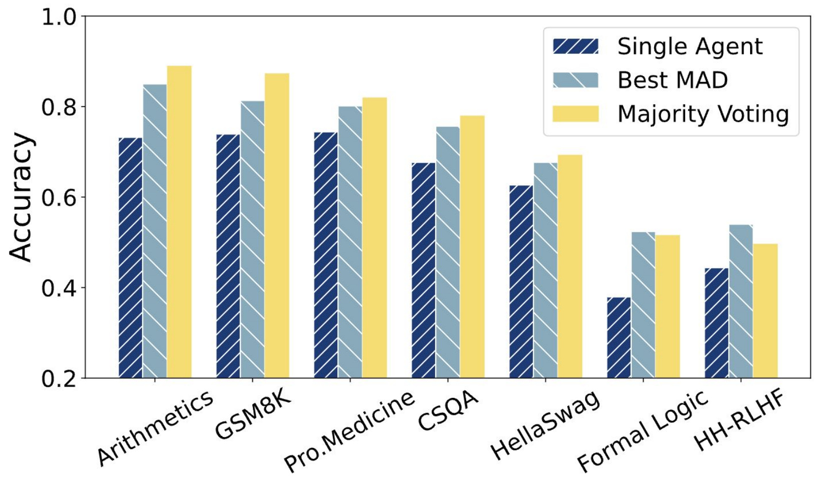
Multi-Agent Debate vs. Majority Voting



Multi-Agent Debate **Majority Voting**

Multi-Agent Debate vs. Majority Voting





Debate is no better than Voting

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\begin{aligned} \textit{\textbf{(Belief sampling)}} \quad & \boldsymbol{\theta}_{i,t} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{i,t}), \\ \textit{\textbf{(Response generation)}} \quad & y_{i,t} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{i,t}). \end{aligned}$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as: $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.*

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector* $\boxed{\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)})} \in \mathbb{R}_+^K$*, where each entry* $\alpha_{i,t}^{(k)}$ *reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\textit{(Belief sampling)} \quad \boldsymbol{\theta}_{i,t} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{i,t}),$$
$$\textit{(Response generation)} \quad y_{i,t} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{i,t}).$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector* $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$*, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as:* $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\boxed{\textbf{\textit{(Belief sampling)}} \quad \boldsymbol{\theta}_{i,t} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{i,t}),}$$
$$\textbf{\textit{(Response generation)}} \quad y_{i,t} \sim \text{Categorical}(\boldsymbol{\theta}_{i,t}).$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as: $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.*

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\textit{(Belief sampling)} \quad \boldsymbol{\theta}_{i,t} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{i,t}),$$

$$\boxed{\textit{(Response generation)} \quad y_{i,t} \sim \text{Categorical}(\boldsymbol{\theta}_{i,t}).}$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as: $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.*

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\begin{aligned} \textbf{\textit{(Belief sampling)}} \quad & \boldsymbol{\theta}_{i,t} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{i,t}), \\ \textbf{\textit{(Response generation)}} \quad & y_{i,t} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{i,t}). \end{aligned}$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as: $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.*

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

$$\textit{\textbf{(Belief sampling)}} \quad \boldsymbol{\theta}_{i,t} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_{i,t}),$$
$$\textit{\textbf{(Response generation)}} \quad y_{i,t} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{i,t}).$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\boxed{\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)})} \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as: $\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}$.*

Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial

Definition 1. (Agent Response Generation via DCM) *At round t, each agent i is associated with a belief vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t}^{(1)}, \ldots, \alpha_{i,t}^{(K)}) \in \mathbb{R}_+^K$, where each entry $\alpha_{i,t}^{(k)}$ reflects the agent's belief in response option $k \in \mathcal{A}$. To generate a response $y_{i,t}$, the agent follows a two-step process:*

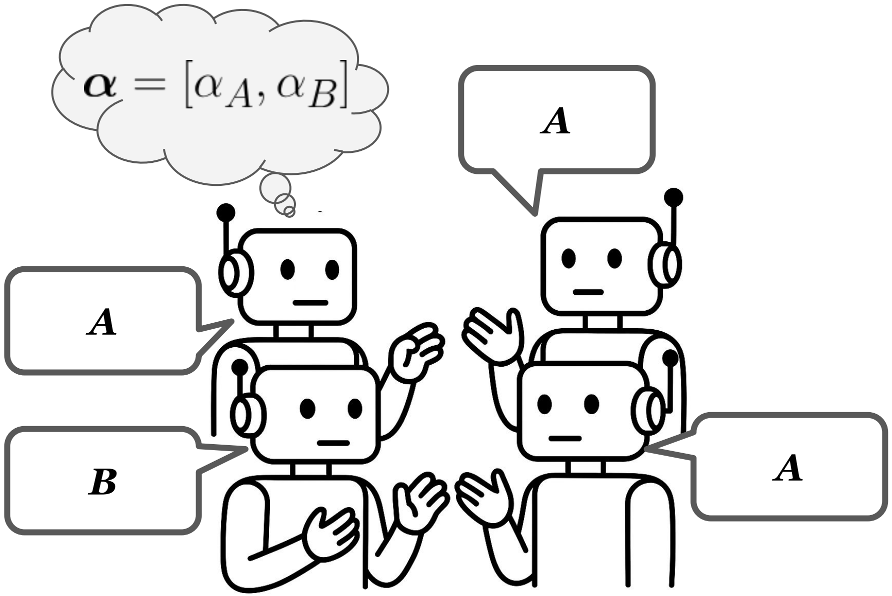
$$\begin{aligned} &\textbf{\textit{(Belief sampling)}} && \boldsymbol{\theta}_{i,t} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{i,t}), \\ &\textbf{\textit{(Response generation)}} && y_{i,t} \sim \text{Categorical}(\boldsymbol{\theta}_{i,t}). \end{aligned}$$

The marginal probability of generating any particular response $y_{i,t} \in \mathcal{A}$—after integrating out the randomness in $\boldsymbol{\theta}_{i,t}$—is given by $P(y_{i,t} = k \mid \boldsymbol{\alpha}_{i,t}) = \alpha_{i,t}^{(k)} / \sum_{j \in \mathcal{A}} \alpha_{i,t}^{(j)}$.

Definition 2. (Bayesian Belief Update from Neighbor Responses) *Let $\{y_{j,t-1} \mid j \in \mathcal{N}(i)\}$ be the set of responses observed by agent i from its neighbors $\mathcal{N}(i)$ at round t. These responses induce a count vector $\mathbf{c}_{i,t} = (c_{i,t}^{(1)}, \ldots, c_{i,t}^{(K)}) \in \mathbb{N}^K$, where $c_{i,t}^{(k)}$ denotes the number of neighbors who selected response k. Then, the agent updates its Dirichlet parameter as* $\boxed{\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{i,t-1} + \mathbf{c}_{i,t}.}$

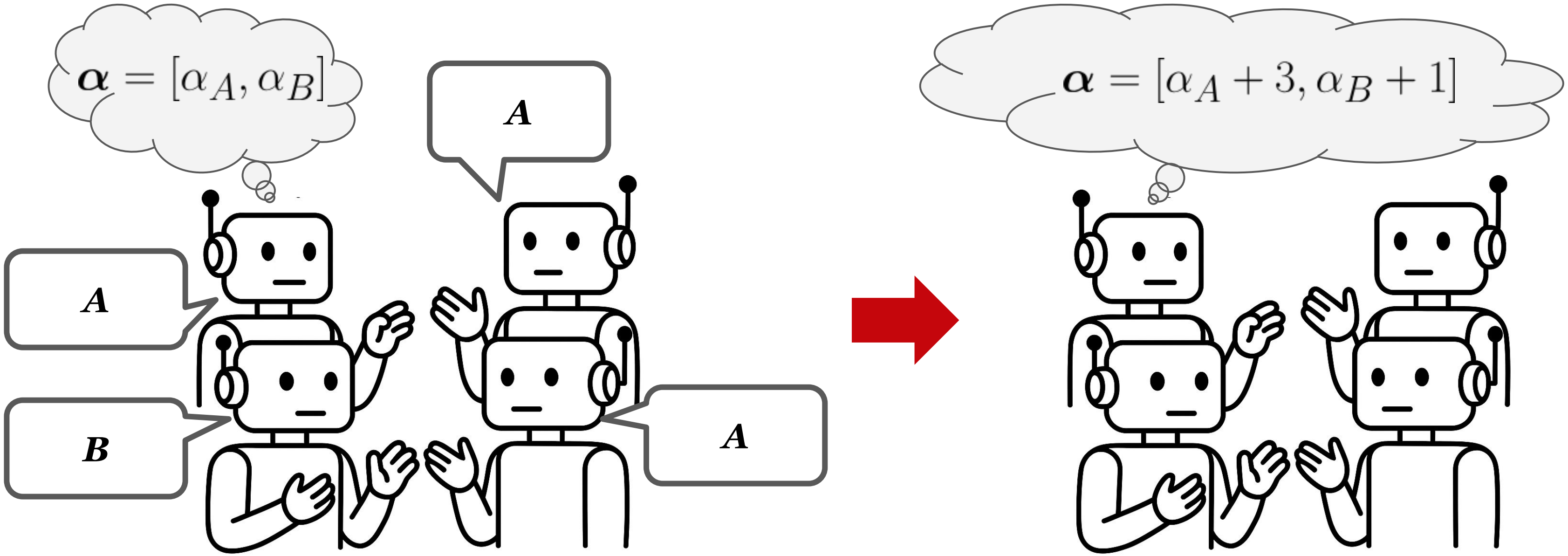
Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial



Conceptual Framework

MAD as a Dirichlet-Compound-Multinomial



MAD is a Martingale

Theorem 2. (Martingale Behavior of Multi-Agent Debate) *For any agent i at round $t > 0$, if*

$$\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} p_{j,t-1} = p_{i,t-1},$$

then sequence $\{p_{i,t}\}_{t \geq 0}$ forms a martingale. That is, the expected belief at the next round equals the current belief:

$$\mathbb{E}[p_{i,t} \mid \boldsymbol{\alpha}_{t-1}] = p_{i,t-1}.$$

MAD is a Martingale

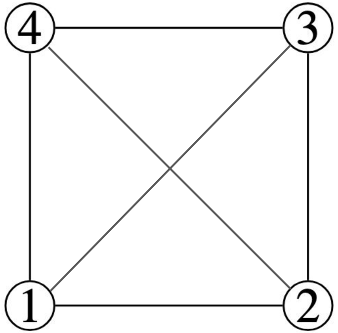
Theorem 2. (Martingale Behavior of Multi-Agent Debate) *For any agent i at round $t > 0$, if*

$$\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} p_{j,t-1} = p_{i,t-1},$$

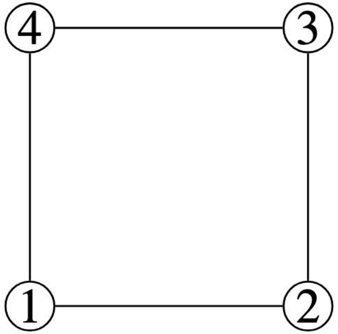
then sequence $\{p_{i,t}\}_{t \geq 0}$ forms a martingale. That is, the expected belief at the next round equals the current belief:

$$\mathbb{E}[p_{i,t} \mid \boldsymbol{\alpha}_{t-1}] = p_{i,t-1}.$$

Proof in Appendix D.3

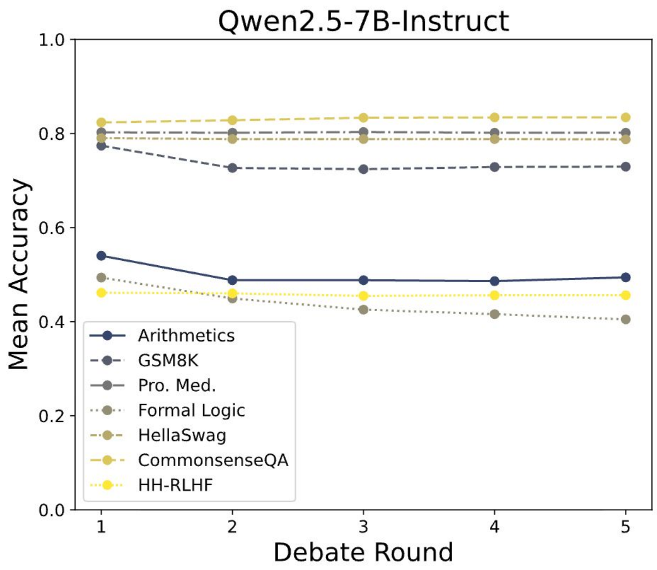


Decentralized MAD
(Martingale Guaranteed)



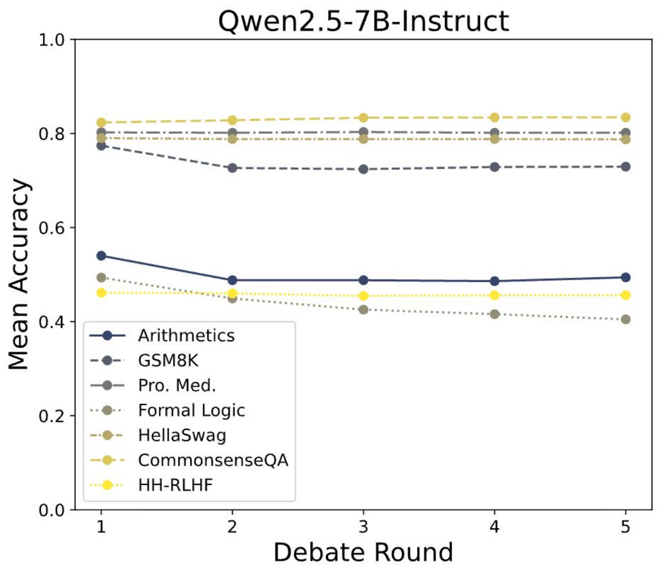
Sparse MAD (li et al.**)**
(Martingale Not Guaranteed)

Empirical Justifications



★ **Direct evaluation of agent-wise accuracy**

Empirical Justifications



★ **Direct evaluation of agent-wise accuracy**

Methods	Qwen2.5-32B-Instruct GSM8K	HellaSwag
Single-Agent		
Single-agent baseline	0.7566 ± .02	0.8700 ± .01
Multi-Agent		
Decentralized MAD ($T = 2$)	0.9367	**0.8767**
Decentralized MAD ($T = 3$)	0.9200	0.8733
Decentralized MAD ($T = 5$)	0.9300	0.8667
Sparse MAD ($T = 2$)	0.9400	0.8700
Sparse MAD ($T = 3$)	**0.9433**	0.8667
Sparse MAD ($T = 5$)	0.9400	0.8700
Centralized MAD ($T = 2$)	0.8000	0.8633
Centralized MAD ($T = 3$)	0.8333	0.8467
Centralized MAD ($T = 5$)	0.8333	0.8233
Majority Voting	**0.9433**	**0.8767**

★ **Larger models**

Methods	Qwen2.5-7B-Instruct GSM8K	MMLU (Pro.Med.)
Single-Agent		
Single-agent baseline	0.6813± .04	0.8257 ± .01
Multi-Agent		
Decentralized MAD ($T = 2$)	0.7867	0.8493
Decentralized MAD ($T = 3$)	0.7467	0.8493
Decentralized MAD ($T = 5$)	0.6567	**0.8529**
Sparse MAD ($T = 2$)	0.8667	0.8272
Sparse MAD ($T = 3$)	0.8300	0.8346
Sparse MAD ($T = 5$)	0.7533	0.8309
Centralized MAD ($T = 2$)	0.6567	0.8088
Centralized MAD ($T = 3$)	0.6367	0.8051
Centralized MAD ($T = 5$)	0.5700	0.8125
Majority Voting	**0.9300**	0.8309

★ **Heterogeneous agents**

hyeongkyu.choi@wisc.edu

www.froilanchoi.com