Group Perspective Matters: Regulating Debate Relationships Can Mitigate Blind Conformity in Multi-Agent Debate

Hao Wu[1,2], Shoucheng Song[1,2], Chang Yao[1,2], Huaiyu Wan[1,2], Youfang Lin[1,2], Kai Lv[1,2]*
[1]School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing, China
{insis_wh,insis_songsc,yaochang,hywan,yflin,lvkai}@bjtu.edu.cn

Abstract

Multi-Agent Debate (MAD) improves the reasoning performance of Large Language Models (LLMs) through multi-round interaction. However, LLMs in MAD are highly susceptible to blind conformity. Existing individual evaluation methods, typically based on confidence or perplexity, fail to reflect the correctness of reasoning and may even exacerbate blind conformity. To address this, we shift the perspective from individual evaluation to group interaction. We define mutual referencing among LLMs as **Debate Relationships** and recognize that regulating these relationships is the key to mitigating blind conformity. In this paper, we propose a novel framework for **D**ynamically r**E**gulating deb**A**te **R**elationships (DEAR) from the group perspective. At first, DEAR quantifies consensus and divergence as *group evidence* to capture the debate state. Then, DEAR operates through three stages: 1) What: perceiving group consultation tendency and uncertainty; 2) Who: introducing a Selection RL-Agent to dynamically select reference peers; and 3) How: adopting a Behavior RL-Agent to adaptively adjust generation behaviors. Notably, we formulate the execution of the two RL-Agents as a sequential decision-making process, jointly optimizing via multi-agent reinforcement learning. Extensive experiments demonstrate that DEAR achieves superior performance while significantly reducing token consumption.

1 Introduction

In recent years, Multi-Agent Debate (MAD) has gained widespread attention as an effective paradigm for enhancing the reasoning capabilities of Large Language Models (LLMs) [1, 2, 3]. Compared to the independent reasoning of a single model, MAD effectively mitigates hallucinations [4, 5, 6] by organizing multiple LLMs to debate around a common problem. However, in MAD, LLMs are highly susceptible to **blind conformity**—the minority often abandons their independent reasoning to align with the majority [7, 8, 9].

Current methods evaluate individual LLM reasoning' reliability using confidence [10, 11, 12, 13] or perplexity [14]. By restricting the propagation of uncertain reasoning, these approaches moderately improve debate quality [11]. However, both confidence and perplexity suffer from inherent flaws. For confidence-based methods, they prompt LLMs to output confidence scores alongside their responses. Yet LLMs exhibit "overconfidence" [15], frequently assigning high confidence to incorrect reasoning even with confidence calibration [10, 12]. For perplexity-based methods, they calculate the token-level negative log-probability of the reasoning to quantify LLMs' uncertainty. Yet perplexity measures textual fluency rather than reasoning correctness, meaning LLMs may easily generate

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

incorrect responses with low perplexity [16, 17]. Moreover, the above methods apply fixed thresholds to propagate only high-confidence or low-perplexity responses. However, such unreliable individual evaluation may amplify the influence of incorrect reasoning and exacerbate blind conformity [18].

The limitations of evaluation methods motivate us to rethink the nature of blind conformity: rather than a single reasoning error, blind conformity is a group phenomenon driven by majority dominance. As this phenomenon emerges from group interactions, static individual evaluations are insufficient to capture the evolving group dynamics. Therefore, to mitigate blind conformity, we need to shift our perspective from individual evaluation to group interactions.

In group interactions, we argue that debate is a dynamic process of viewpoint collision and integration. During this process, LLMs continuously update their reasoning by referencing peer opinions. We define the above mutual referencing as **Debate Relationship**. This relationship spans the entire group interaction and serves as the medium for viewpoint exchange. A beneficial debate relationship facilitates the reasonable adoption of peer responses, enabling LLMs to enhance their reasoning. Therefore, effectively regulating the debate relationships at the group level offers a viable pathway to mitigate the phenomenon of blind conformity.

Building on the above insight, we propose a novel framework for **D**ynamically r**E**gulating deb**A**te **R**elationships (DEAR) from the group perspective. To capture the debate state after each round, DEAR initially quantifies the group consensus and divergence as *group evidence*. Leveraging this evidence, DEAR regulates the debate relationships through three stages: perceiving consultation tendency and uncertainty (**What**), selecting reference peers (**Who**), and guiding generation behaviors (**How**). Specifically, in the first stage, we employ Subjective Logic (SL) [19] to decouple the group evidence into consultation tendency and uncertainty among LLMs. In the second stage, we introduce a Selection RL-Agent to learn an adaptive peer selection policy to determine which peers each LLM should reference. Finally, based on the established debate relationships, a Behavior RL-Agent is adopted to adjust the generation behaviors of individual LLMs. Furthermore, we formulate the ordered execution of the above RL-Agents as a sequential decision-making process. To guarantee the coordinated optimization of these policies, we innovatively employ the Heterogeneous-Agent Proximal Policy Optimization (HAPPO) algorithm [20] for end-to-end joint training. Extensive experiments on four mathematical reasoning tasks and four question-answering (QA) tasks demonstrate the effectiveness of our proposed method.

Our main contributions are summarized as follows:

- We shift the perspective from individual evaluation to group interactions centered on debate relationships, effectively mitigating blind conformity in Multi-Agent Debate.
- We propose DEAR, a novel framework that employs group evidence to dynamically regulate debate relationships through: What–Who–How stages.
- We conduct extensive experiments to demonstrate that DEAR achieves superior performance while significantly reducing token consumption.

2 Related Works

2.1 Single-Agent Reasoning.

Several works propose various single-agent reasoning policies to augment the reasoning capabilities of LLMs [21, 22, 23, 24, 25, 26]. However, constrained by the intrinsic knowledge boundaries of an individual model, these methods struggle to correct hallucinations and reasoning errors.

2.2 Multi-Agent Debate.

Multi-Agent Debate (MAD) leverages multi-round LLM interactions to improve reasoning performance [1, 2, 27, 28, 29], yet LLMs frequently exhibit blind conformity during debate [7, 8]. Existing methods filter responses using confidence or perplexity to improve debate quality [10, 11, 14, 13], but the inherent unreliability of these metrics may instead exacerbate blind conformity [18]. FreeMAD [30] attempts to address the error propagation caused by incorrect majority responses through heuristic scoring of LLMs' reasoning trajectories. However, evaluating the correctness of individual reasoning is fundamentally infeasible during the debate process without ground truth.

More detailed related works and discussion can be seen in Appendix A.

3 Problem Definition

We aim to mitigate blind conformity in the Multi-Agent Debate (MAD) system by regulating debate relationships from the group perspective, without fine-tuning any internal parameters of Large Language Models (LLMs). In this setting, we formulate this regulation process as a Partially Observable Markov Decision Process (POMDP) defined by the tuple $\langle \mathcal{N}, \mathcal{V}, \mathcal{S}, \Omega, \mathcal{O}, \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{N} = \{1, 2\}$ represents the set of the Selection RL-Agent (with policy ρ) and the Behavior RL-Agent (with policy π); $\mathcal{V} = \{1, \ldots, V\}$ denotes the set of debating LLMs, with V representing their number; $\mathcal{S}$ denotes the state space; Ω represents the observation space; $\mathcal{O}$ is the observation function; $\mathcal{I}$ and $\mathcal{A}$ denote the respective action spaces for the Selection and Behavior RL-Agents; $\mathcal{T}$ is the state transition function; $\mathcal{R}$ is the reward function that assigns an environmental reward R_{debate} based on the correctness of the final answer at the terminal round T; and γ is the discount factor.

Within this formulation, the Selection RL-Agent receives an observation $o^\rho \in \Omega$ derived from the group evidence, and outputs selection actions $\boldsymbol{ids} \in \mathcal{I}$ to regulate which peer opinions each LLM should reference. The Behavior RL-Agent takes the observation and action of the Selection RL-Agent as conditional inputs, i.e., $o^\pi = (o^\rho, \boldsymbol{ids}) \in \Omega \times \mathcal{I}$, and outputs generation behaviors $\boldsymbol{a} \in \mathcal{A}$, where $a_i = [T_i, Top_p_i]$, to adjust the temperature and nucleus sampling parameters for each LLM. According to the probability chain rule, the joint policy of the two RL-Agents can be defined as:

$$\boldsymbol{\pi}^{\boldsymbol{all}}(\boldsymbol{ids}, \boldsymbol{a} \mid s) = \rho(\boldsymbol{ids} \mid o^\rho) \cdot \pi(\boldsymbol{a} \mid o^\pi). \tag{1}$$

4 Method

To mitigate blind conformity in MAD, our core idea is to dynamically regulate debate relationships from the group perspective. As illustrated in Figure 1, after each generation round, DEAR first quantifies the group consensus and divergence as group evidence to capture the debate state (▷ Section 4.1). Driven by this, DEAR operates through three stages: (1) **What** (▷ Section 4.2): We apply Subjective Logic (SL) [19] to decouple the group evidence into consultation tendency and uncertainty, providing the foundation for regulating debate relationships. (2) **Who** (▷ Section 4.3): A Selection RL-Agent is introduced to select which peer responses each LLM should reference, dynamically regulating the debate relationships. (3) **How** (▷ Section 4.4): A Behavior RL-Agent is adopted to adjust the generation behaviors of LLMs within the established debate relationships. Finally, we employ the Heterogeneous-Agent Proximal Policy Optimization (HAPPO) algorithm [20] for joint training of the above RL-Agents (▷ Section 4.5).

4.1 Group Evidence Extraction

To dynamically regulate debate relationships, we should first realize the current debate state. During the debate, the consensus and divergence among LLMs directly represent the collision and integration of viewpoints, objectively reflecting the current group dynamics. Therefore, we quantify the above two critical indicators as group evidence to capture the debate state.

Specifically, we extract group evidence from paired LLM responses (Ans_i, Ans_j). Inherently, the final answers serve as the most direct indication of LLMs' viewpoint. The alignment between two answers directly signifies whether they are in consensus or in divergence. However, standard outputs (e.g., a number or option) yield excessively sparse semantics for meaningful comparison. To address this, we prompt each LLM to summarize its reasoning process into a conclusion. Subsequently, we quantify the consensus and divergence across two dimensions.

Conclusion Similarity. Conclusions represent the LLMs' most direct judgments. To quantify conclusion similarity, we project the conclusions into a latent space $h_i \in \mathbb{R}^d$ using a pre-trained text encoder $\xi(\cdot)$ and calculate their scaled cosine similarity:

$$e_{ij}^{(con)} = \frac{1}{2}\left(1 + \frac{h_i^\top h_j}{\|h_i\|_2 \|h_j\|_2}\right). \tag{2}$$

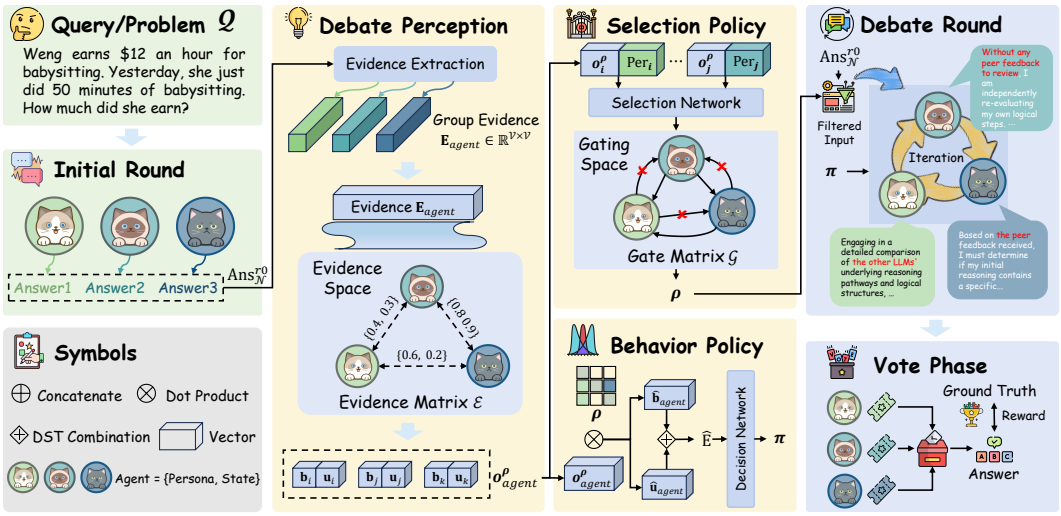

Figure 1: The overall framework of DEAR. We employ a two-round debate as an example to detail our method. (1) *Debate Perception*, consisting of extracting group evidence and **What**: perceiving group consultation tendency and uncertainty. (2) *Selection Policy*, representing **Who**: adaptively selecting reference peers. (3) *Behavior Policy*, denoting **How**: adjusting generation behavior.

Reasoning Similarity. Conclusion similarity only captures the alignment of final judgments, but fails to reveal the underlying logical reasoning process. To address this, we further quantify reasoning similarity. Following the same procedure as conclusion similarity, we map the reasoning processes of LLMs into latent representations $c_i \in \mathbb{R}^d$. We then quantify the reasoning similarity with scaled cosine similarity:

$$e_{ij}^{(rea)} = \frac{1}{2}\left(1 + \frac{c_i^\top c_j}{\|c_i|_2 \|c_j|_2}\right). \tag{3}$$

The above two similarities satisfy $e_{ij}^{(con)}, e_{ij}^{(rea)} \in [0, 1]$. We adopt the arithmetic mean to compute the scalar evidence score e_{ij} (where we set $e_{ii} = 0$). To capture the debate state, we aggregate all pairwise scores to formulate the overall group evidence matrix $\mathbf{E}_{agent}$. By measuring inter-LLM similarity independent of task-specific semantics, this evidence matrix is inherently task-agnostic and generalizes well across diverse reasoning tasks.

4.2 Group Evidence Decouple (What)

While group evidence objectively captures consensus and divergence, it fails to reflect an LLM's subjective consultation tendency toward its peers. Moreover, when facing conflicting opinions, LLMs inevitably experience uncertainty regarding peer credibility. This uncertainty is the key to perceiving whether they will abandon independent reasoning and blindly conform. Therefore, perceiving both the consultation tendency and the uncertainty is essential to effectively regulate debate relationships.

To realize this, we employ Subjective Logic (SL) [19]. SL provides a theoretical framework that translates evidence into a distribution of belief masses across multiple targets, accompanied by an overall uncertainty mass. Accordingly, we decouple LLM i's evidence into two components: the consultation tendency (the belief mass) b_{ij} toward peer j, and the corresponding uncertainty u_i. These variables satisfy the following zero-sum constraint:

$$u_i + \sum_{j\in\mathcal{V}} b_{ij} = 1, \quad b_{ij} \geq 0, \quad u_i \geq 0. \tag{4}$$

In SL, transforming evidence into the belief mass requires a base weight to represent the initial total uncertainty (i.e., when no evidence has been collected). We set this base weight to the number of LLMs V. Defining the total accumulated evidence as $E_i = \sum_{j\in\mathcal{V}} e_{ij}$, the overall evidence strength

of LLM i is $S_i = E_i + V$. The belief mass b_{ij} and the uncertainty u_i can be derived as:

$$b_{ij} = \frac{e_{ij}}{S_i}, \quad u_i = \frac{V}{S_i}. \tag{5}$$

Based on this formulation, we perceive the LLMs' subjective consultation tendencies and uncertainties to serve the subsequent regulation.

4.3 Adaptive Debater Selection (Who)

With the consultation tendencies and uncertainties, we adopt them to regulate the debate relationships by determining the specific peers each LLM should reference. However, relying on static heuristic rules to make decisions is inadequate for the evolving dynamics of debate. Thus, we introduce a Selection RL-Agent to learn an adaptive peer selection policy ρ for group-level regulation.

To achieve this, we construct a group observation o^ρ by concatenating the consultation tendency b_i, uncertainty u_i, and a one-hot identity embedding $ID_i \in \{0, 1\}$ of all LLMs:

$$o^\rho = [b_1 \parallel u_1 \parallel \mathrm{ID}_1 \parallel \cdots \parallel b_V \parallel u_V \parallel \mathrm{ID}_\mathrm{V}]. \tag{6}$$

Then, we learn a mapping network $f_\rho : \mathbb{R}^{V(2V+1)} \to [0, 1]^{V\times(V-1)}$ that maps o^ρ to a preference matrix $\mathcal{P} = [p_{i\leftarrow j}]_{i\neq j}$. Fundamentally, regulating debate relationships is a binary decision-making process: an LLM either references a peer's response or does not. To realize this, we formulate the discrete peer selection action $\boldsymbol{ids} \in \mathcal{I}$ as a set of independent Bernoulli variables parameterized by $\mathcal{P}$. Accordingly, each binary action $ids_{i\leftarrow j} \in \{0, 1\}$, indicating whether LLM i references peer j, is sampled from the following joint policy distribution:

$$\boldsymbol{ids} \sim \rho(\boldsymbol{ids} \mid o^\rho) = \prod_{i=1}^{V} \prod_{j=1,\ j\neq i}^{V} (p_{i\leftarrow j})^{ids_{i\leftarrow j}} (1 - p_{i\leftarrow j})^{1-ids_{i\leftarrow j}}. \tag{7}$$

The action $\boldsymbol{ids}$ serves as the interaction mask for the subsequent debate round. By achieving group-level regulation of debate relationships, the RL-Agent selectively severs the transmission of harmful responses while preserving beneficial reasoning exchange, thereby mitigating blind conformity.

4.4 Generation Behavior Adjustment (How)

After selecting a distinct subset of reference peers, each LLM should integrate the selected peers' evidence into its own decisions. Traditional mean aggregation fails to distinguish between consensus and divergence, which are essential for determining whether an LLM should converge on existing opinions or explore breakthroughs. Instead, we employ the Dempster-Shafer Evidence Theory (DST) combination rule to fuse evidence.

Definition 4.4.1 (*Dempster's Combination Rule*). *Given two independent belief sets $\mathcal{M}_1 = (\{b_{1k}\}_{k=1}^V, u_1)$ and $\mathcal{M}_2 = (\{b_{2k}\}_{k=1}^V, u_2)$, their fused joint belief $\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2 = (\{b_k\}_{k=1}^V, u)$ is computed as:*

$$b_k = \frac{1}{1-C}(b_{1k}b_{2k} + b_{1k}u_2 + b_{2k}u_1), \quad u = \frac{1}{1-C}u_1u_2,$$

where $C = \sum_{p\neq q} b_{1p}b_{2q}$ quantifies the degree of conflict between the two beliefs, and $\frac{1}{1-C}$ is the normalization factor.

Based on this, LLM i fuses its own belief set $\mathcal{M}_i$ with peers from the selected subset $\mathcal{X}_i = \{j \in \mathcal{V} \mid j \neq i, ids_{i\leftarrow j} = 1\}$ by the subsequent rule:

$$\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2 \oplus \ldots \mathcal{M}_V. \tag{8}$$

Once we have determined the fused belief $\hat{\mathcal{M}}_i = \left(\hat{b}_i, \hat{u}_i\right)$, the evidence $\hat{E}_i$ can be derived as: $\hat{S}_i = \frac{V}{\hat{u}_i}, \hat{E}_i = \hat{b}_i \times \hat{S}_i$. Taking $\hat{E}_i$ as input, we sample the continuous variable $z_i = [z_{i,T}, z_{i,P}]^\top$ from a diagonal Gaussian distribution parameterized by the Behavior RL-Agent policy π:

$$z_i \sim \pi(\cdot \mid \hat{E}_i) = \mathcal{N}\left(\boldsymbol{\mu}(\hat{E}_i), \mathrm{diag}(\boldsymbol{\nu}(\hat{E}_i)^2)\right). \tag{9}$$

To ensure the generation behaviors fall within the valid ranges $[T_{min}, T_{max}]$ and $[P_{min}, P_{max}]$, we apply a scaling mapping to the sampled variable z_i:

$$T_i = T_{min} + (T_{max} - T_{min}) \cdot \sigma(z_{i,T}), \tag{10}$$

$$Top_p_i = P_{min} + (P_{max} - P_{min}) \cdot \sigma(z_{i,P}), \tag{11}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. Finally, the Behavior RL-Agent outputs the continuous generation behavior $a_i = [T_i, Top_p_i]$, effectively adjusting the LLM's generation behavior in the subsequent debate round.

4.5 Optimization Scheme

Within DEAR, the Selection RL-Agent first generates the peer selection action $\boldsymbol{ids}$. Subsequently, the Behavior RL-Agent dynamically adjusts the LLM's generation behavior based on $\boldsymbol{ids}$. This ordered execution constitutes an explicit sequential decision-making process. To ensure coordinated optimization, we employ the Heterogeneous-Agent Proximal Policy Optimization (HAPPO) algorithm [20] for end-to-end joint training.

The Selection RL-Agent Optimization. Given that the Selection RL-Agent acts first in each iteration, we prioritize its optimization. Taking o^ρ as its input, this RL-Agent outputs a binary action $\boldsymbol{ids}$ for all LLMs. Its policy update ratio is defined as $r_\phi^{\rho,t} = \frac{\rho_\phi(\boldsymbol{ids}^t|o^\rho)}{\rho_{\phi_{old}}(\boldsymbol{ids}^t|o^\rho)}$. The Selection RL-Agent is optimized based on the advantage function $\hat{A}^t$, with the objective formulated as:

$$\mathcal{L}_\rho(\phi) = \mathbb{E}\left[-\min\left(r_\phi^{\rho,t}\hat{A}^t, \text{clip}(r_\phi^{\rho,t}, 1\pm\epsilon)\hat{A}^t\right)\right]. \tag{12}$$

Here, $\hat{A}^t$ represents the Generalized Advantage Estimation (GAE), which is computed via the Temporal Difference (TD) error δ^t:

$$\hat{A}^t = \sum_{l=0}^{T-t}(\gamma\lambda)^l\delta^{t+l}, \quad \delta^t = \mathcal{R}^t + \gamma V_\psi(s^{t+1}) - V_\psi(s^t), \quad \mathcal{R}^t = \begin{cases} R_{debate}, & t = T \\ 0, & t < T \end{cases}, \tag{13}$$

where R_{debate} denotes the environmental reward, T denotes the number of debate rounds.

The Behavior RL-Agent Optimization. Following the update of the Selection RL-Agent, we proceed to optimize the Behavior RL-Agent. Specifically, the sequential surrogate advantage for the Behavior RL-Agent is defined as: $M^{\pi,t} = r_\phi^{\rho,t}\hat{A}^t$, where $r_\phi^{\rho,t}$ is the importance sampling ratio previously computed by the Selection RL-Agent. This term accounts for the distributional change induced by the updated Selection policy when optimizing the conditional Behavior policy. Since the Sigmoid mapping is a deterministic transformation, the essence of updating the Behavior RL-Agent is to optimize the parameterized distribution of the variable $\boldsymbol{z}^t$. Therefore, the corresponding policy update ratio is computed as $r_\theta^{\pi,t} = \frac{\pi_\theta(\boldsymbol{z}^t|\hat{E}^t)}{\pi_{\theta_{old}}(\boldsymbol{z}^t|\hat{E}^t)}$. The optimization objective is defined as:

$$\mathcal{L}_\pi(\theta) = \mathbb{E}\left[-\min\left(r_\theta^{\pi,t}M^{\pi,t}, \text{clip}(r_\theta^{\pi,t}, 1\pm\epsilon)M^{\pi,t}\right)\right]. \tag{14}$$

The Global Critic Optimization. To support the advantage estimation for the above RL-Agents, the centralized critic $V_\psi(s)$ precisely fits the expected return $\hat{r}^t = \hat{A}^t + V_{\psi_{old}}(s^t)$. We additionally apply value clipping to stabilize critic updates. The objective is expressed as:

$$\mathcal{L}_{critic}(\psi) = \mathbb{E}\left[\max\left((V_\psi(s^t) - \hat{r}^t)^2, \left(\text{clip}(V_\psi(s^t), V_{\psi_{old}}(s^t)\pm\epsilon) - \hat{r}^t\right)^2\right)\right]. \tag{15}$$

The pseudo-code for DEAR is given in Appendix H. Through the end-to-end training, we obtain the coordinated pair of Selection RL-Agent and Behavior RL-Agent for subsequent evaluation.

5 Experiments

In this section, we design extensive experiments to answer: (**RQ1**) How does DEAR perform against existing MAD methods in terms of both accuracy and token consumption? (**RQ2**) How scalable is DEAR across varying foundation models and debate configurations? (**RQ3**) What is the specific contribution of key components to the performance of DEAR?

Table 1: Performance comparison on `Qwen3-8B`. We highlight the **optimal** and suboptimal results. **Acc. (%)** denotes accuracy, and **T.** (K) represents the average tokens per task.

Math Reasoning Tasks											
Architecture	**Method**	**ID**		**OOD**						**Avg. (Math)**	
		GSM8K		**AIME24**		**GSM-Hard**		**MATH-500**			
		Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
Single-Agent	CoT	91.5	0.4	30.0	4.1	57.0	0.4	70.8	1.2	62.3	1.5
	CoT-SC	93.5	5.4	**36.7**	64.2	60.3	8.4	75.8	16.9	66.6	23.7
Multi-Agent	Majority Voting	93.0	1.1	33.3	11.7	60.5	1.7	74.0	3.4	65.2	4.5
	MAD	93.5	11.1	30.0	75.7	61.5	16.3	76.3	30.8	65.3	33.5
	DMAD	91.5	17.5	33.3	98.2	61.8	23.4	78.8	41.0	66.4	45.0
	CortexDebate	**94.0**	15.9	30.0	100.0	60.5	21.3	69.3	35.6	63.5	43.2
	MAD-M^2(O)	92.0	8.7	30.0	47.9	61.5	10.0	77.0	20.8	65.1	21.9
	ECON	91.0	8.0	–	–	61.0	9.3	72.0	13.3	–	–
	DEAR (Ours)	**94.0**	7.6	**36.7**	21.2	**73.5**	6.2	**85.0**	8.4	**72.3**	10.9

QA Tasks											
Architecture	**Method**	**ID**		**OOD**						**Avg. (QA)**	
		ARC-C		**MMLU (Pro.H.)**		**TruthfulQA**		**GPQA Dia.**			
		Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
Single-Agent	CoT	90.0	0.4	56.8	0.4	63.3	0.5	36.7	2.0	61.7	0.8
	CoT-SC	**95.0**	8.1	61.5	6.2	70.5	7.5	40.7	25.6	66.9	11.9
Multi-Agent	Majority Voting	91.0	1.5	66.5	2.1	74.0	1.9	41.3	6.5	68.2	3.0
	MAD	90.0	15.7	62.5	30.8	68.8	20.2	40.7	30.4	65.5	24.3
	DMAD	91.5	22.6	64.0	39.2	72.0	27.7	38.0	64.1	66.4	38.4
	CortexDebate	91.8	11.8	60.0	19.8	70.8	12.6	46.0	61.8	67.2	26.5
	MAD-M^2(O)	90.0	4.7	65.0	14.5	73.0	6.8	42.7	29.8	67.7	14.0
	ECON	90.5	13.3	64.8	10.9	64.5	9.2	31.3	16.4	62.8	12.5
	DEAR (Ours)	93.3	4.6	**68.5**	6.9	**80.8**	4.8	**52.0**	13.9	**73.7**	7.6

5.1 Experimental Settings

Datasets and Benchmarks. To evaluate the effectiveness of DEAR, we employ four math reasoning datasets (i.e., GSM8K [31], AIME24, GSM-Hard [32], and MATH-500 [33]) and four question answering (QA) datasets (i.e., ARC-C [34], MMLU_Pro (Health) [35], TruthfulQA [36], and GPQA Diamond [37]). Among these, AIME24 and GPQA Diamond are challenging tasks.

Baselines. For single-agent baselines, we adopt Chain-of-Thought (CoT) [21] and Self-Consistency Chain-of-Thoughts (CoT-SC) [22] with 16 independent reasoning paths. For multi-agent baselines, we employ Majority Voting, Multi-Agent Debate (MAD) [1], Diverse Multi-Agent Debate (DMAD) [38], and CortexDebate [10]. We employ MAD-M²(O) [14] for open-source LLMs, and regarding the MAD-M^2(O) for perplexity calculation, we adopt its variant, MAD-M^2(S) [14] for closed-source LLMs. Additionally, for training-based reinforcement learning methods, we select ECON [28] as our baseline. Notably, we refrain from fine-tuning any internal parameters of LLMs and exclusively train the Selection RL-Agent and the Behavior RL-Agent.

Detailed descriptions of the benchmarks and baselines are provided in Appendix B.

Implementation Details. Only divergent viewpoints are worthy of debate. The diverse responses of LLMs can introduce multidimensional perspectives into MAD [39, 40, 12, 41]. Therefore, for the homogeneous debate system, we assign a distinct persona to each LLM following [39]. To ensure fair comparison, all baselines are initialized with the same personas as DEAR. For single-agent baselines, we compute the average accuracy across different personas. The pre-trained text encoder $\xi(\cdot)$ uses BGE-M3 [42]. Comprehensive LLM persona configurations and additional implementation details are provided in Appendix B.3.

5.2 Main Results (RQ1)

We evaluate the performance of DEAR using both an open-source model (`Qwen3-8B`) [43] and a closed-source model (`GPT-4o-mini`). For math reasoning tasks, DEAR is trained on the GSM8K training set and evaluated on all math test sets, with GSM8K as the in-distribution (ID) benchmark

Table 2: Performance and token consumption comparison on advanced closed-source models.

LLM	Method	Math Reasoning Tasks								QA Tasks							
		ID		OOD				Avg. (Math)		ID		OOD				Avg. (QA)	
		GSM-Hard		AIME24		IMO AnswerBench				TruthfulQA		MedXpertQA (Text)		GPQA Dia.			
		Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
`Claude-4.5-Sonnet`	MAD	65.0	8.74	60.0	40.7	23.0	39.9	49.3	29.8	74.0	5.95	35.0	29.1	76.0	32.2	61.7	22.4
	CortexDebate	69.0	12.64	53.3	57.6	21.0	57.3	47.8	42.5	79.0	7.35	29.0	48.4	74.0	50.4	60.7	35.4
	MAD-M^2(S)	70.0	14.39	63.3	44.8	25.0	40.0	52.8	33.1	77.0	6.49	31.0	30.7	78.0	36.5	62.0	24.6
	DEAR (Ours)	**76.0**	7.26	**76.7**	26.6	**35.0**	31.7	**62.6**	21.9	**83.0**	4.27	**40.0**	23.2	**82.0**	24.0	**68.3**	17.2
`Gemini-3.1-Flash-Lite`	MAD	68.0	15.93	63.3	35.4	22.0	34.1	51.1	28.5	89.0	13.75	54.0	25.3	73.0	28.0	72.0	22.4
	CortexDebate	71.0	17.25	66.7	36.0	21.0	50.4	52.9	34.6	91.0	16.81	44.0	31.2	80.0	38.8	71.7	28.9
	MAD-M^2(S)	72.0	19.63	66.7	33.1	24.0	33.3	54.2	28.7	90.0	18.45	55.0	23.2	81.0	27.9	75.3	23.2
	DEAR (Ours)	**75.0**	12.74	**73.3**	19.6	**27.0**	19.1	**58.4**	17.1	**93.0**	10.28	**59.0**	10.6	**84.0**	13.6	**78.7**	11.5

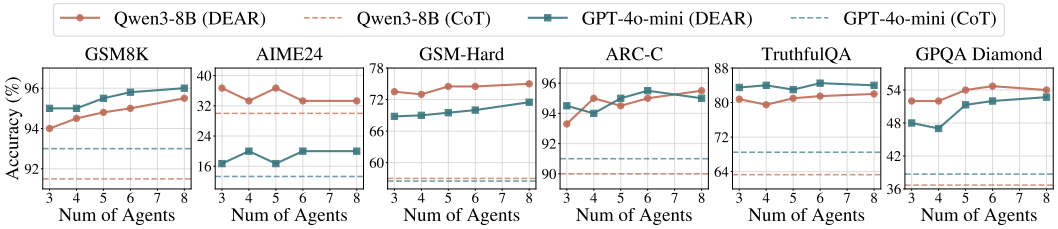

Figure 2: Effect of scaling the number of agents in the case of `Qwen3-8B` and `GPT-4o-mini`.

and the others as out-of-distribution (OOD) benchmarks. For QA tasks, DEAR is trained on the ARC-C training set and evaluated on all QA test sets, with ARC-C as ID and the others as OOD. Table 1 presents the evaluation results, with additional results provided in Appendix Table 3. Our key observations are summarized as follows:

***Observation 1*: DEAR achieves the optimal or suboptimal results across diverse task domains**. As shown in Table 1 and 3, DEAR significantly outperforms baseline methods in average accuracy for both math reasoning and QA tasks. Even in challenging datasets, AIME24 and GPQA Diamond, DEAR yields substantial performance gains. These results show that MAD's fully-connected topology tends to amplify erroneous reasoning. Meanwhile, individual evaluation methods such as CortexDebate and MAD-M^2(O) rely on unreliable confidence or perplexity, which exacerbate blind conformity and compromise overall performance. In contrast, DEAR regulates debate relationships based on objective group evidence rather than flawed metrics. Using this evidence, the Selection RL-Agent dynamically selects which peers to reference to shield reasonable minority reasoning. Meanwhile, the Behavior RL-Agent utilizes DST-fused evidence to adaptively balance exploration and convergence, guiding the LLMs toward a valid consensus.

***Observation 2*: DEAR significantly reduces token consumption during the debate**. Current methods typically rely on massive token expenditure to improve accuracy. This reliance is especially obvious in debate methods such as DMAD and individual LLM policies like CoT-SC. In contrast, DEAR's efficiency stems from the dynamic peer selection driven by the Selection RL-Agent. By perceiving the consultation tendency and uncertainty, the RL-Agent adaptively severs valueless peer references, preserving the beneficial reasoning exchanges, thereby reducing token consumption.

5.3 Scalability Analysis (RQ2)

This section evaluates DEAR across two dimensions: (1) adaptability across open- and closed-source models, and (2) scalability with respect to the number of debate rounds and LLM agents.

Different Model Adaptability. To verify compatibility, we test DEAR on advanced closed-source models `Claude-4.5-Sonnet` and `Gemini-3.1-Flash-Lite`. We also incorporate the exceptionally challenging datasets IMOAnswerBench [44] and MedXpertQA (Text) [45], while randomly sampling 200 examples from GSM-Hard and TruthfulQA to construct the math and QA training sets, respectively. As shown in Table 2, DEAR seamlessly adapts to diverse models while achieving superior performance on complex tasks. Several existing MAD methods rely on extracting token-level probabilities to calculate perplexity [46, 14]. However, closed-source models and APIs restrict access to these probability distributions, severely limiting their applicability. DEAR eliminates this reliance

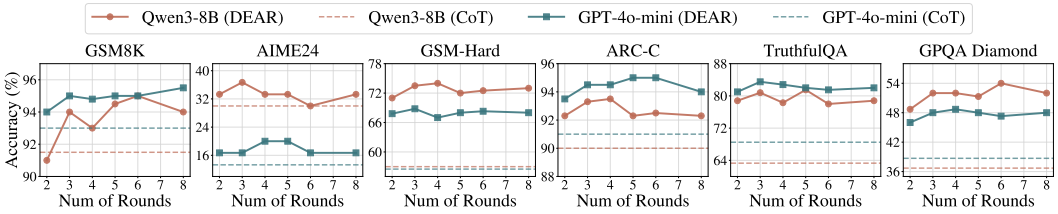

Figure 3: Effect of scaling the number of debate rounds in the case of `Qwen3-8B` and `GPT-4o-mini`.

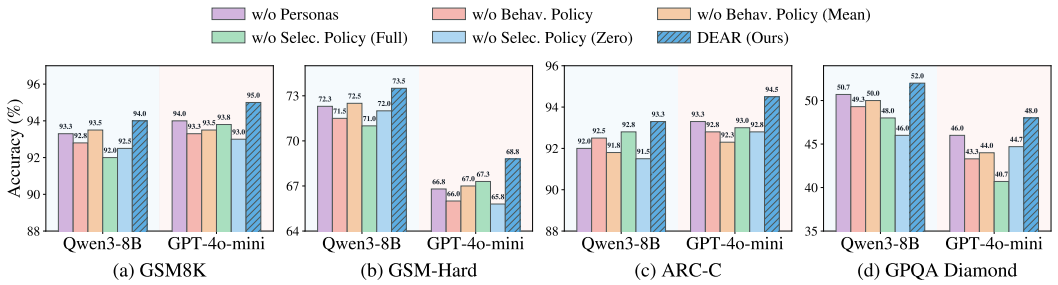

Figure 4: Ablation study of the proposed DEAR on four datasets.

by utilizing group evidence, which quantifies inter-LLM similarity based solely on semantic outputs, thereby decoupling debate regulation from internal token distributions and specific tasks.

Scaling of Debate Rounds & Number of LLMs. In this section, we evaluate the scalability of our framework. When scaling the number of LLMs, we fix the debate at three rounds. We also assign specific personas to each LLM following the configuration in [39]. Conversely, when scaling the debate rounds, we fix the LLM count at three. As illustrated in Figures 2 and 3, the performance of DEAR consistently improves as the number of LLMs increases. This demonstrates that the Selection RL-Agent effectively identifies reliable peers. In terms of scaling the debate rounds, DEAR converges to near-optimal performance within fewer rounds. This is attributed to the Behavior RL-Agent's adjustment of generation behaviors, which facilitates valid consensus formation in the early rounds.

5.4 Ablation Study (RQ3)

To validate the effectiveness of each components within DEAR, we design several ablated variants. Specifically, we remove diverse personas (w/o Personas). For the Behavior RL-Agent, we either completely remove it (w/o Behav. Policy) or substitute the DST fusion rule with simple mean-pooling (w/o Behav. Policy (Mean)). As for the Selection RL-Agent, we introduce two variants: a fully connected setting where all debaters are involved (w/o Selec. Policy (Full)), and a fully disconnected setting where selection is disabled (w/o Selec. Policy (Zero)).

Figure 4 presents the results of these variants across four datasets using `Qwen3-8B` and `GPT-4o-mini`. The results confirm that removing any individual module degrades overall performance. Among all components, the Selection RL-Agent demonstrates the most substantial impact, particularly on complex reasoning tasks, indicating that regulating which peer should be referenced is the primary driver of debate quality. Disabling the Behavior RL-Agent also degrades, and replacing DST fusion with mean-pooling leads to performance decay, demonstrating that mean aggregation fails to distinguish between consensus and divergence among selected peers. Finally, while diverse personas contribute positively, their impact is relatively smaller compared to the RL-Agents. We investigate the performance of various diversity configurations and present the results in Appendix E.

6 Conclusion and Future Works

In this paper, we propose DEAR, a novel framework that mitigates blind conformity by dynamically regulating debate relationships from the group perspective. Operating through the What-Who-How

stages, DEAR utilizes group evidence to drive two RL-Agents for adaptive peer selection and behavior adjustment. To guarantee coordinated optimization, DEAR employs a multi-agent reinforcement learning algorithm for end-to-end joint training. Extensive experiments demonstrate that DEAR achieves superior performance while significantly reducing token consumption.

In future works, we plan to extend DEAR to more complex scenarios, such as generative tasks and multimodal domains. This exploration will include applications in collaborative coding and open-ended planning, as well as situations involving visual reasoning and cross-modal analysis.

7 Acknowledgments

This work was supported by the the National Natural Science Foundation of China (Grant No. 62576029).

References

[1] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

[2] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[3] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, 2025.

[4] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904, 2024.

[5] Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*, 2023.

[6] Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[7] Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5123–5139, 2025.

[8] Mengdie Flora Wang, Haochen Xie, Guanghui Wang, Aijing Gao, Guang Yang, Ziyuan Li, Qucy Wei Qiu, Fangwei Han, Hengzhi Qiu, Yajing Huang, et al. From debate to decision: Conformal social choice for safe multi-agent deliberation. *arXiv preprint arXiv:2604.07667*, 2026.

[9] Ruiwen Zhou, Maojia Song, Xiaobao Wu, Sitao Cheng, Xunjian Yin, Yuxi Xie, Zhuoqun Hao, Wenyue Hua, Liangming Pan, Soujanya Poria, et al. Epistemic context learning: Building trust the right way in llm-based multi-agent systems. *arXiv preprint arXiv:2601.21742*, 2026.

[10] Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9503–9523, 2025.

[11] Yilin Bai. Confidencecal: Enhancing llms reliability through confidence calibration in multi-agent debate. In *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*, pages 221–226. IEEE, 2024.

[12] Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. *arXiv preprint arXiv:2601.19921*, 2026.

[13] Haocheng Yang, Fengxiang Cheng, Tianjun Yao, Mengyue Yang, Jiajun Chai, Xiaohan Wang, Guojun Yin, Wei Lin, Soummya Kar, Fenrong Liu, et al. Mad-logic: Multi-agent debate enhances symbolic translation and reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. *arXiv preprint arXiv:2603.20215*, 2026.

[15] Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*, 2024.

[16] Petar Veličković, Federico Barbero, Christos Perivolaropoulos, Simon Osindero, and Razvan Pascanu. Perplexity cannot always tell right from wrong. *arXiv preprint arXiv:2601.22950*, 2026.

[17] Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. Can perplexity reflect large language model's ability in long text understanding? *arXiv preprint arXiv:2405.06105*, 2024.

[18] Pradyumna Shyama Prasad and Minh Nhat Nguyen. When two llms debate, both think they'll win. *arXiv preprint arXiv:2505.19184*, 2025.

[19] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[20] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[23] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[24] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

[25] Jin Ziqi and Wei Lu. Tab-cot: Zero-shot tabular chain of thought. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10259–10277, 2023.

[26] KaShun Shum, Shizhe Diao, and Tong Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, 2023.

[27] Haoran Li, Ziyi Su, Yun Xue, Zhiliang Tian, Yiping Song, and Minlie Huang. Advancing collaborative debates with role differentiation through multi-agent reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22655–22666, 2025.

[28] Xie Yi, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. From debate to equilibrium: Belief-driven multi-agent llm reasoning via bayesian nash equilibrium. *arXiv preprint arXiv:2506.08292*, 2025.

[29] Yexiong Lin, Shanshan Ye, Yu Yao, Zhen Fang, Bo Han, and Tongliang Liu. Squidagent: Parallelize wisely, coordinate efficiently. *arXiv preprint arXiv:2610.08647*, 2026.

[30] Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025.

[31] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[32] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.

[33] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[34] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

[35] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[36] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[37] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*, 2024.

[38] Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*, 2025.

[39] Yingxuan Yang, Chengrui Qu, Muning Wen, Laixi Shi, Ying Wen, Weinan Zhang, Adam Wierman, and Shangding Gu. Understanding agent scaling in llm-based multi-agent systems via diversity. *arXiv preprint arXiv:2602.03794*, 2026.

[40] Manh Nguyen, Anh Nguyen, Dung Nguyen, Svetha Venkatesh, and Hung Le. Hear both sides: Efficient multi-agent debate via diversity-aware message retention. *arXiv preprint arXiv:2603.20640*, 2026.

[41] Zhenghao Li, Zhi Zheng, Wei Chen, Jielun Zhao, Yong Chen, Tong Xu, and Enhong Chen. Dynadebate: Breaking homogeneity in multi-agent debate with dynamic path generation. *arXiv preprint arXiv:2601.05746*, 2026.

[42] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[43] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[44] Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL `https://aclanthology.org/2025.emnlp-main.1794/`.

[45] Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pages 80961–80990. PMLR, 2025.

[46] Dan Qiao, Binbin Chen, Fengyu Cai, Jianlong Chen, Wenhao Li, Fuxin Jiang, Zuzhi Chen, Hongyuan Zha, Tieying Zhang, and Baoxiang Wang. Epistemic gain, aleatoric cost: Uncertainty decomposition in multi-agent debate for math reasoning. *arXiv preprint arXiv:2603.01221*, 2026.

[47] Shuyuan Zhao, Wei Chen, Boyan Shi, Liyong Zhou, Shuohao Lin, and Huaiyu Wan. Spatial-temporal knowledge distillation for takeaway recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13365–13373, 2025.

[48] Shuyuan Zhao, Wei Chen, Weijie Zhang, Xinrui Hou, Junfeng Shen, Boyan Shi, Shengnan Guo, Youfang Lin, and Huaiyu Wan. Stk-adapter: Incorporating evolving graph and event chain for temporal knowledge graph extrapolation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19763–19778, 2026.

[49] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, 2024.

[50] Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, Xitai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S2-mad: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9393–9408, 2025.

[51] Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*, 2024.

[52] Chao Wang, Han Lin, Huaze Tang, Huijing Lin, and Wenbo Ding. Rumad: Reinforcement-unifying multi-agent debate. *arXiv preprint arXiv:2602.23864*, 2026.

[53] Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

[54] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008.

[55] Glenn Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1(330-331):3, 1992.

[56] Audun Jøsang and Robin Hankin. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion*, pages 1225–1232. IEEE, 2012.

[57] Dale Fixsen and Ronald PS Mahler. The modified dempster-shafer approach to classification. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(1):96–104, 2002.

[58] Thierry Denoeux. A neural network classifier based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

[59] Yongchuan Tang, Xu Zhang, Ying Zhou, Yubo Huang, and Deyun Zhou. A new correlation belief function in dempster-shafer evidence theory and its application in classification. *Scientific Reports*, 13(1):7609, 2023.

[60] Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. Detecting misbehaviors of large vision-language models by evidential uncertainty quantification. *arXiv preprint arXiv:2602.05535*, 2026.

[61] Shenshen Li, Wenxin Meng, Lei Wang, Hao Yang, Chong Peng, Peng Yan, Fumin Shen, Jingkuan Song, Heng Tao Shen, and Xing Xu. From observation to understanding: Front-door adjustments with uncertainty calibration for enhancing egocentric reasoning in lvlms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19152–19169, 2025.

[62] Jie Shi, Xiaodong Yue, Wei Liu, Yufei Chen, and Feifan Dong. Not all inconsistency is equal: Decomposing lvlm uncertainty into belief divergence and belief conflict. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25339–25347, 2026.

[63] Tao Huang, Zhekun Liu, Rui Wang, Yang Zhang, and Liping Jing. Visual hallucination detection in large vision-language models via evidential conflict. *International Journal of Approximate Reasoning*, 186: 109507, 2025.

[64] Siyuan Wang, James R Foulds, Md Osman Gani, and Shimei Pan. Llm-based corroborating and refuting evidence retrieval for scientific claim verification. *arXiv preprint arXiv:2503.07937*, 2025.

[65] Fan Liu, Yue Feng, Zhao Xu, Lixin Su, Xinyu Ma, Dawei Yin, and Hao Liu. Jailjudge: A comprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework. *arXiv preprint arXiv:2410.12855*, 2024.

[66] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[67] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

[68] Hao Wu, Shoucheng Song, Chang Yao, Sheng Han, Huaiyu Wan, Youfang Lin, and Kai Lv. Think how your teammates think: Active inference can benefit decentralized execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29749–29757, 2026.

[69] Shoucheng Song, Youfang Lin, Sheng Han, Chang Yao, Hao Wu, Shuo Wang, and Kai Lv. Code: Communication delay-tolerant multi-agent collaboration via dual alignment of intent and timeliness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23304–23312, 2025.

[70] Chang Yao, Youfang Lin, Shoucheng Song, Hao Wu, Yuqing Ma, Shang Han, and Kai Lv. From general relation patterns to task-specific decision-making in continual multi-agent coordination. *arXiv preprint arXiv:2507.06004*, 2025.

[71] Chang Yao, Youfang Lin, Shoucheng Song, Hao Wu, Shengkun Yang, Yuqing Ma, and Kai Lv. Role-level inductive bias for cross-task generalization in multi-agent reinforcement learning. In *Forty-third International Conference on Machine Learning*.

[72] Shoucheng Song, Chang Yao, Hao Wu, Youfang Lin, Sheng Han, and Kai Lv. Towards high semantic fidelity: Hyperdimensional symbolic messages in multi-agent communication, 2026. URL `https://arxiv.org/abs/2610.23092`.

[73] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[74] Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[75] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[76] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[77] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

A Detailed Related Works

A.1 Single-Agent Reasoning.

Large Language Models (LLMs) have shown formidable reasoning capabilities by independently executing complete reasoning processes. However, direct answer generation often fails on complex tasks because it forces models to jump to conclusions without intermediate logical deduction. To address this, some methods [21, 22, 26, 47, 48] have proposed various prompt-based internal reasoning policies. The most representative paradigm is Chain-of-Thought (CoT) [21], which guides models to decompose complex problems into sequential steps through few-shot exemplars or zero-shot instructions such as "Let's think step by step" [21]. To further enhance robustness and leverage the diversity of generation paths, Self-Consistency Chain-of-Thought (CoT-SC) [22] samples multiple reasoning trajectories for the same problem and extracts the final answer via majority voting. As research progressed, single-agent reasoning evolved beyond linear paradigms to explore more complex logical topologies. For instance, Tab-CoT [25] introduces highly structured reasoning. Tree-of-Thoughts (ToT) [23] constructs a tree topology by generating multiple candidate answers at each step, facilitating multi-path exploration and self-evaluation. Despite these advances, single-agent reasoning remains constrained by the knowledge and reasoning boundaries of an individual model, making it difficult to reliably correct hallucinations and reasoning errors.

A.2 Multi-Agent Debate.

Multi-Agent Debate (MAD) [1] scales conventional reasoning paradigms by allowing multiple LLMs to engage in multi-round interactions, ultimately reaching a consensus through majority voting. However, fully connected debate incurs substantial token overhead. To improve efficiency, subsequent studies reduce interaction costs by sparsifying debate topologies or eliminating unnecessary debate rounds. S-MAD [49] and S^2-MAD [50] prune redundant information flows, while GD [51] restricts

Table 3: Performance comparison on `GPT-4o-mini`. We highlight the **optimal** and suboptimal results. **Acc. (%)** denotes accuracy, and **T.** (K) represents the average tokens per task.

Math Reasoning Tasks											
Architecture	**Method**	**ID**		**OOD**						**Avg. (Math)**	
		GSM8K		**AIME24**		**GSM-Hard**		**MATH-500**			
		Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
Single-Agent	CoT	93.0	0.5	13.3	1.1	56.5	0.5	65.8	0.8	57.2	0.7
	CoT-SC	**96.3**	6.0	10.0	17.3	59.3	8.5	73.3	12.5	59.7	11.1
Multi-Agent	Majority Voting	94.5	1.2	6.7	3.3	57.3	1.4	63.3	2.2	55.5	2.0
	MAD	94.5	13.3	13.3	36.1	54.8	15.3	64.8	24.2	56.9	22.2
	DMAD	93.8	17.5	10.0	37.0	55.3	19.3	72.3	27.6	57.9	25.4
	CortexDebate	93.5	22.6	6.7	34.4	56.3	25.6	69.3	36.6	56.5	29.8
	MAD-M^2(S)	94.0	9.0	10.0	32.3	53.5	9.7	71.0	21.1	57.1	18.0
	ECON	91.5	8.1	–	–	52.0	9.0	65.0	12.9	–	–
	DEAR (Ours)	95.0	6.3	**16.7**	18.7	**68.8**	7.4	**79.8**	12.0	**65.1**	11.1

QA Tasks											
Architecture	**Method**	**ID**		**OOD**						**Avg. (QA)**	
		ARC-C		**MMLU (Pro.H.)**		**TruthfulQA**		**GPQA Dia.**			
		Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
Single-Agent	CoT	91.0	0.3	71.0	0.3	68.5	0.3	38.7	1.0	67.3	0.5
	CoT-SC	93.0	6.6	73.5	7.2	70.5	7.3	44.0	16.2	70.3	9.3
Multi-Agent	Majority Voting	93.8	1.2	73.0	1.3	68.0	1.4	40.7	3.0	68.9	1.7
	MAD	93.0	12.5	73.5	19.4	72.0	14.0	40.0	32.1	69.6	19.5
	DMAD	92.0	17.7	71.3	25.5	77.8	19.2	35.3	31.8	69.1	23.6
	CortexDebate	92.0	12.0	70.3	15.5	74.5	11.3	42.0	42.7	69.7	20.4
	MAD-M^2(S)	93.8	6.1	73.0	10.4	62.5	7.3	42.7	20.2	68.0	11.0
	ECON	92.5	9.5	71.0	10.0	73.0	8.8	42.0	14.8	69.6	10.8
	DEAR (Ours)	**94.5**	5.6	**74.0**	8.5	**83.5**	5.2	**48.0**	14.3	**75.0**	8.4

global debate to local groups. RUMAD [52] introduces a reinforcement learning-based controller to generate dynamic debate topologies. However, it does not adapt the generation behavior of LLMs.

Beyond interaction sparsification, several methods further assess response quality. CortexDebate [10] and ConfidenceCal [11] use explicit confidence scores to select or weight opinions. MAD-M^2 [14] filters historical information through agent-generated subjective labels or perplexity. However, confidence, perplexity, and subjective assessments may not reliably reflect reasoning quality, while reliance on internal token probabilities limits applicability to closed-source models. More importantly, LLMs may prematurely abandon independent judgment under peer influence, resulting in blind conformity and error propagation [53]. Free-MAD [30] mitigates this issue by evaluating individual reasoning trajectories using heuristic scores. Nevertheless, reliably determining reasoning correctness during inference remains difficult without ground-truth supervision. In contrast, our DEAR framework bypasses flawed individual evaluations by dynamically regulating debate relationships at the group interaction level, effectively protecting the valid reasoning of the minority.

A.3 The Dempster-Shafer Evidence Theory (DST).

DST is a theory of belief functions established by Dempster [54]. It combines beliefs from multiple sources through fusion operators, producing a unified belief distribution over the available evidence [55, 56]. Early studies primarily applied DST to model epistemic uncertainty in classification [57, 58, 59]. DST has also inspired evidential learning methods that represent predictive beliefs and epistemic uncertainty through Subjective Logic (SL) [19]. Extending evidential learning to multi-source settings, With the development of LLMs, DST has been applied to identify feature-level multimodal conflicts for hallucination detection in large vision-language models (LVLMs) [60, 61, 62, 63]. In retrieval-augmented generation (RAG), DST is used to resolve conflicts among evidence retrieved from multiple sources [64], while in LLM-as-a-Judge, it supports the fusion of evaluations from multiple agents [65]. In this work, we use DST combination rules to fuse the evidence of the LLMs and their selected peers, thereby guiding the adaptation of generation behaviors.

B Experimental Details

B.1 Datasets Descriptions.

Since multi-agent, multi-round debate incurs substantially higher inference costs than standard single-agent evaluation, we randomly sample 400 examples from each benchmark whenever the dataset size permits, using the same sampled subset across all methods for consistent comparison. For open-source models, we use the full AIME24 dataset and randomly sample 150 questions from GPQA Diamond. For closed-source models in the extension experiments, due to API cost constraints, we use 200 examples from GSM-Hard and TruthfulQA for training and 100 non-overlapping examples for evaluation, while using 100 evaluation examples for GPQA Diamond, IMOAnswerBench, and MedXpertQA. The datasets are described below:

- **GSM8K** [31] comprises grade-school mathematics word problems designed to evaluate multi-step arithmetic and basic logical reasoning.
- **AIME24** consists of official problems from the 2024 American Invitational Mathematics Examination and requires rigorous mathematical reasoning.
- **GSM-Hard** [32] is a more challenging variant of GSM8K featuring larger and more complex numerical values, placing greater demands on arithmetic reasoning.
- **MATH-500** [33] is curated from the MATH benchmark and covers advanced mathematical domains such as algebra, geometry, and number theory across multiple difficulty levels.
- **ARC-C** [34] contains challenging multiple-choice questions from grade-school science examinations and requires scientific knowledge and reasoning beyond superficial text retrieval.
- **MMLU_Pro (Health)** [35] is a challenging subset of MMLU-Pro focusing on health and medicine, designed to evaluate professional knowledge and domain-specific reasoning.
- **TruthfulQA** [36] evaluates whether models can provide truthful responses when confronted with common human misconceptions.
- **GPQA Diamond** [37] is an expert-level benchmark comprising graduate-level questions in physics, biology, and chemistry, requiring specialized scientific knowledge and reasoning.
- **IMOAnswerBench** [44] contains challenging International Mathematical Olympiad problems and is used to evaluate complex and rigorous mathematical reasoning.
- **MedXpertQA** [45] focuses on complex clinical medical scenarios. We use its text-modal subset to evaluate domain-specific medical reasoning.

B.2 Baselines Descriptions

In this section, we provide detailed descriptions of each baseline used in our comparison:

- **CoT (Chain-of-Thought)** [21]: Guides large language models to generate intermediate step-by-step reasoning processes before producing the final output, serving as the foundational single-agent prompting method in this study.
- **CoT-SC (Self-Consistency Chain-of-Thought)** [22]: Builds upon CoT by performing independent sampling to generate multiple distinct reasoning paths (set to 16 in our experiments) and using Majority Voting to determine the final answer.
- **Majority Voting**: Aggregates the initial outputs from multiple independent agents, directly selecting the most frequently occurring answer as the final group decision.
- **MAD (Multi-Agent Debate)** [1]: Allows multiple agents to read and evaluate each other's reasoning processes across multiple interaction rounds, attempting to reach a group consensus through an iterative feedback mechanism.
- **DMAD (Diverse Multi-Agent Debate)** [38]: Assigns diverse perspectives to agents within the traditional debate framework to mitigate the formation of mental sets during multi-agent interactions.
- **CortexDebate** [10]: Utilizes the internal confidence of the models as a filtering metric to dynamically construct a sparse communication topology, thereby organizing information exchange pathways among agents.

- **MAD-M²** [14]: Incorporates subjective (S) and objective (O) methods to filter historical memory. The subjective approach (S) requires agents to autonomously evaluate and label records (e.g., "YES," "NO," or "NOT SURE"), whereas the objective approach (O) utilizes model perplexity to screen responses.
- **ECON** [28]: A multi-agent reinforcement learning-based debate method that strictly freezes the internal parameters of LLMs, training only external networks to adjust and optimize the generation behavior used during LLM API calls.

B.3 Implementation Details

Our pre-trained text encoder, $\xi(\cdot)$, employs BGE-M3, which outputs feature embeddings with a dimensionality of 1024. To instantiate the diverse homogeneous debate system introduced in the main text, we assign a distinct persona to each LLM. Specifically, for math reasoning tasks, the designated personas are `Rigorous_Formalist`, `Creative_Explorer`, and `Systematic_Decomposer`; for question-answering (QA) tasks, they are `Consensus_Fact_Checker`, `Eliminative_Reasoner`, and `Careful_Reader`. To guarantee a fair comparison, all baseline methods are initialized with identical persona configurations. The explicit prompt templates corresponding to each persona are detailed in Appendix G. The core reinforcement learning hyperparameters and the debate environment parameters adopted during the training process are summarized in Table 4.

Table 4: Key hyperparameters in the training process of DEAR.

Parameter	Explanation	Value
HAPPO Optimization Parameters		
γ	Discount factor	1.0
λ	GAE parameter	0.95
ϵ	PPO clip range	0.2
actor lr	Actor learning rate	5×10^{-4}
critic lr	Critic learning rate	5×10^{-4}
λ_v	Value loss coefficient	0.5
–	Optimizer type	Adam
num mini-batch	Number of mini-batches	1
ppo epoch	Number of PPO epochs	5
Debate Environment and Behavior Parameters		
V	Number of LLMs	3
T	Maximum debate rounds	3
R_{debate}	Reward for correct consensus	$+1.0$
$T_{\min}/T_{\max}$	Temperature action boundaries	$0/1.0$
$P_{\min}/P_{\max}$	Top-p action boundaries	$0/1.0$

C Analysis for HAPPO with Reward Sharing

C.1 Selection of HAPPO

We consider a cooperative Multi-Agent Debate (MAD) system, where all LLMs collaboratively strive to maximize the accuracy of the group answer. In our proposed DEAR, the Selection RL-Agent and the Behavior RL-Agent form an explicit sequential decision: the Behavior RL-Agent's input is strictly conditioned on the action output of the Selection RL-Agent. Moreover, the two RL-Agents operate over heterogeneous action spaces—the former outputs discrete binary masks, while the latter produces continuous generation behavior.

Standard algorithms such as Independent PPO (IPPO) [66] optimize each RL-Agent independently and do not explicitly account for inter-agent coordination. To coordinate the heterogeneous selection and behavior policies, DEAR adopts the sequential policy-update and importance-weighted advantage propagation mechanism of Heterogeneous-Agent Proximal Policy Optimization (HAPPO) [20]. Specifically, the Selection RL-Agent is updated first, after which its policy importance ratio scales the

Behavior RL-Agent's surrogate advantage. This sequential correction accounts for the change in the selection policy when optimizing the conditional behavior policy, providing a principled mechanism for their coordinated optimization.

C.2 Reward Sharing

Reward sharing is a standard paradigm in cooperative Multi-Agent Reinforcement Learning (MARL) [67, 68, 69, 70, 71, 72], premised on the assumption that all agents pursue a unified optimization objective. Our framework DEAR intrinsically satisfies this assumption. The Selection RL-Agent and the Behavior RL-Agent jointly regulate the debate relationships through peer selection and behavior generation to maximize the final accuracy. This collective outcome is inherently inseparable, as it arises from the coupled interplay of peer selection and behavior generation, and cannot be meaningfully attributed to either RL-Agent individually.

Furthermore, reward sharing is not merely a heuristic choice, but a necessity for HAPPO's sequential advantage decomposition to function correctly. The generalized advantage $\hat{A}^t$, derived from a shared reward and a global critic, provides the base signal that is scaled by the Selection RL-Agent's importance sampling ratio to guide the Behavior RL-Agent's update. Without a shared reward, this sequential propagation of policy dependencies would be ill-defined. Therefore, employing reward sharing is both logically sound and theoretically indispensable for the coordinated optimization of our heterogeneous RL-Agents.

D Analysis of Computational Efficiency and Training Overhead

All experiments are conducted on a workstation equipped with an AMD Ryzen 9 9950X 16-Core Processor and a single NVIDIA GeForce RTX 5090 GPU. To accelerate the empirical data collection process, we deploy 16 parallel interaction environments. For open-source models, training DEAR takes approximately 4.6 hours on GSM8K and 4.0 hours on ARC-Challenge. For closed-source models, training takes approximately 3.6 hours on GSM-Hard and 3.2 hours on TruthfulQA. These results demonstrate the training efficiency of DEAR across both open- and closed-source models.

E Additional Experiments

E.1 Minority Correction and Conformity Analysis

This section examines whether DEAR mitigates blind conformity while retaining beneficial correction in multi-agent debate. We focus on two complementary situations: incorrect minority agents adopting a correct majority answer, and correct minority agents conforming to an incorrect majority. We quantify these behaviors using Beneficial Correction Rate (BCR) and Harmful Conformity Rate (HCR), respectively. Both metrics are formally defined below.

E.1.1 Metric Definitions

Let the evaluation set contain N problems, each involving V LLMs. We index debate rounds by $t = 1, \ldots, T$, where $t = 1$ denotes the initial round and $t = T$ denotes the terminal round. T is the total number of debate rounds, and the observed transitions are $t \to t + 1$ for $t = 1, \ldots, T - 1$. For problem q_n, let $y_n^\star$ denote the ground-truth answer and $y_{n,i}^t$ denote the answer produced by the i-th LLM at round t.

When the modal answer is unique, the group majority answer at round t is defined as:

$$m_n^t = \mathrm{Mode}\left(y_{n,1}^t, \ldots, y_{n,V}^t\right), \tag{16}$$

where $\mathrm{Mode}(\cdot)$ returns the most frequently occurring answer among the V LLMs.

If two or more answers tie for the highest frequency, no unique majority answer is defined for that round. We define:

$$\mathcal{U} = \left\{(n,t) \in \{1,\ldots,N\} \times \{1,\ldots,T-1\} \;\middle|\; m_n^t \text{ is uniquely defined}\right\} \tag{17}$$

as the set of problem–round pairs with a unique majority answer and an observed subsequent round. Since BCR and HCR characterize minority updates relative to the current majority, they are computed

only over $(n,t) \in \mathcal{U}$. Terminal-round responses may serve as the destination of the last transition, but the terminal round itself is excluded from $\mathcal{U}$ as no subsequent response is generated.

Beneficial Correction Rate (BCR). BCR measures the fraction of incorrect minority answers at round t that change to the correct round-t majority answer at round $t+1$. For $(n,t) \in \mathcal{U}$, the indicator is defined as:

$$\mathcal{B}_{n,i}^{t} = \mathbb{I}\big[y_{n,i}^{t} \neq y_n^{\star} \wedge y_{n,i}^{t} \neq m_n^{t} \wedge m_n^{t} = y_n^{\star}\big]. \tag{18}$$

BCR is then calculated as:

$$\mathrm{BCR} = \frac{\sum_{(n,t)\in\mathcal{U}} \sum_{i=1}^{V} \mathcal{B}_{n,i}^{t} \mathbb{I}\big[y_{n,i}^{t+1} = m_n^{t}\big]}{\sum_{(n,t)\in\mathcal{U}} \sum_{i=1}^{V} \mathcal{B}_{n,i}^{t}}. \tag{19}$$

A higher BCR indicates more frequent beneficial shifts of minority answers toward the correct majority.

Harmful Conformity Rate (HCR). HCR measures the fraction of correct minority answers at round t that change to the incorrect round-t majority answer at round $t+1$. For $(n,t) \in \mathcal{U}$, the indicator is defined as:

$$\mathcal{H}_{n,i}^{t} = \mathbb{I}\big[y_{n,i}^{t} = y_n^{\star} \wedge y_{n,i}^{t} \neq m_n^{t} \wedge m_n^{t} \neq y_n^{\star}\big]. \tag{20}$$

HCR is then calculated as:

$$\mathrm{HCR} = \frac{\sum_{(n,t)\in\mathcal{U}} \sum_{i=1}^{V} \mathcal{H}_{n,i}^{t} \mathbb{I}\big[y_{n,i}^{t+1} = m_n^{t}\big]}{\sum_{(n,t)\in\mathcal{U}} \sum_{i=1}^{V} \mathcal{H}_{n,i}^{t}}. \tag{21}$$

A lower HCR indicates that correct minority answers are less likely to change to the incorrect majority answer, consistent with stronger resistance to harmful conformity.

Here, $\mathbb{I}[\cdot]$ denotes the indicator function, which equals 1 when the specified condition holds and 0 otherwise. BCR reflects beneficial correction and is better when higher, whereas HCR captures harmful conformity and is better when lower.

E.1.2 Results and Analysis.

Table 5: Comparison of debate methods in terms of blind conformity mitigation. We highlight the **optimal** and suboptimal results.

Method	GSM8K (ID)			MATH-500 (OOD)			ARC-C (ID)			TruthfulQA (OOD)		
	Acc. ↑	HCR ↓	BCR ↑	Acc. ↑	HCR ↓	BCR ↑	Acc. ↑	HCR ↓	BCR ↑	Acc. ↑	HCR ↓	BCR ↑
MAD	<u>93.5</u>	43.3	65.2	76.3	51.4	<u>73.9</u>	90.0	22.5	74.5	68.8	41.7	<u>79.3</u>
DMAD	91.5	23.8	60.6	<u>78.8</u>	70.8	66.4	91.5	55.0	71.1	72.0	26.7	45.5
CortexDebate	**94.0**	<u>19.2</u>	69.7	69.3	<u>23.3</u>	59.6	<u>91.8</u>	<u>16.7</u>	<u>83.3</u>	70.8	<u>20.8</u>	68.3
MAD-M^2(O)	92.0	32.7	<u>79.0</u>	77.0	70.9	69.7	90.0	62.5	79.5	<u>73.0</u>	31.0	78.9
DEAR (Ours)	**94.0**	**12.5**	**84.2**	**85.0**	**18.3**	**82.5**	**93.3**	**14.2**	**94.5**	**80.8**	**18.3**	**86.7**

We conduct these experiments using `Qwen3-8B` as the backbone LLM. As shown in Table 5, DEAR achieves the highest BCR and lowest HCR across all four datasets, indicating that incorrect minority agents are more likely to adopt a correct majority answer, while correct minority agents are less likely to conform to an incorrect majority. These results suggest that DEAR's regulation is sensitive to the evolving group state rather than uniformly encouraging agreement or independence. By jointly learning whom to reference and how to generate subsequent responses, DEAR supports selective revision instead of indiscriminate alignment with the majority.

E.2 Analysis of Different Diversity Configurations

To evaluate the impact of diversity and decouple its benefits from our proposed DEAR, we structure LLM heterogeneity into four progressive levels based on [39]. This approach systematically isolates each diversity source:

- L1 (No Diversity): Identical base models and identical default prompts.
- L2 (Persona Diversity Only): Identical base models equipped with the persona prompts specified in Section B.3.
- L3 (Model Diversity Only): Distinct base models with identical default prompts.
- L4 (Full Diversity): Distinct base models equipped with the persona prompts specified in Section B.3.

In our specific setup, the heterogeneous small-scale LLM group utilizes `LLaMA-3.1-8B-Instruct` [73], `GLM-4-9B-0414` [74], and `Qwen3-8B` [43]. The heterogeneous large-scale LLM group employs `DeepSeek-V3.2` [75], `GPT-5.4`, and `Claude-Sonnet-4.5`. For the homogeneous setting, we use `Qwen3-8B` and `Claude-Sonnet-4.5`.

Table 6: Performance comparison under different diversity configurations for small-scale LLMs.

Config.	Method	GSM8K Acc. (%)	GSM-Hard Acc. (%)	ARC-C Acc. (%)	GPQA Dia. Acc. (%)
L1	MAD	92.3	60.0	88.5	40.0
	DEAR	**93.3**	**72.3**	**92.0**	**50.7**
L2	MAD	93.5	61.5	90.0	40.7
	DEAR	**94.0**	**73.5**	**93.3**	**52.0**
L3	MAD	94.0	66.0	92.3	57.3
	DEAR	**95.0**	**77.3**	**94.5**	**67.3**
L4	MAD	94.5	66.3	93.3	62.0
	DEAR	**95.8**	**78.0**	**95.5**	**68.0**

Table 7: Performance comparison under different diversity configurations for large-scale LLMs.

Config.	Method	AIME24 Acc. (%)	IMO AnswerBench Acc. (%)	MedXpertQA (Text) Acc. (%)	GPQA Dia. Acc. (%)
L1	MAD	53.3	21.0	30.0	74.0
	DEAR	**70.0**	**32.0**	**37.0**	**80.0**
L2	MAD	60.0	23.0	35.0	76.0
	DEAR	**76.7**	**35.0**	**40.0**	**82.0**
L3	MAD	86.7	37.0	38.0	82.0
	DEAR	**93.3**	**42.0**	**42.0**	**85.0**
L4	MAD	90.0	40.0	39.0	83.0
	DEAR	**96.7**	**44.0**	**45.0**	**87.0**

The results are presented in Tables 6 and 7. Based on this, we draw the following observations:

***Observation 1*: Introducing diversity yields progressive performance gains.** From L1 to L4, debate performance improves steadily. Model-level differentiation (L3) serves as the primary driver, while persona prompts (L2/L4) provide only marginal additional gains.

***Observation 2*: DEAR consistently outperforms the MAD baseline across all configurations.** In both homogeneous and heterogeneous settings, DEAR surpasses MAD. This indicates that the performance of DEAR stems from the regulation of debate relationships by the two heterogeneous RL-Agents, rather than from a specific diversity configuration.

In summary, diversity configuration and debate relationship regulation are two orthogonal dimensions for improving debate performance. The core innovation of the DEAR lies in the dynamic regulation of debate relationships rather than the introduction of diversity. Moreover, DEAR is compatible with various diversity settings and delivers consistent performance gains under each configuration.

E.3 Comparison with Sparse Multi-Agent Debate

To assess DEAR's performance and token efficiency, we compare it with static sparse debate (S-MAD [49], S^2-MAD [50]) and an RL-based dynamic topology generation method (RUMAD) [52]. Ensuring a fair comparison, we strictly adopt RUMAD's configuration: a 6-round debate involving 6 LLMs. These LLMs are uniformly instantiated from three base models (LLaMA-3.1-8B-Instruct [73], ChatGLM-4 9B [74], and Deepseek-Math-7B-Instruct [76]), utilizing two agents per model. Performance is assessed on GSM8K [31], MMLU [77], and GPQA [37].

- MMLU: A comprehensive benchmark evaluating multi-domain reasoning capabilities in LLMs, spanning diverse fields such as STEM, humanities, and social sciences.
- GPQA: A highly challenging, graduate-level QA dataset designed to test advanced scientific reasoning across disciplines like physics, biology, and chemistry.

Table 8: Comparison of accuracy and token consumption with sparse debate methods. **Acc. (%)** denotes accuracy, and **T. (K)** represents the average tokens per task. S-MAD$_*$ and S-MAD$_o$ represent S-MAD configured with star and ring topologies, respectively. Best results are **bolded** and the second-best are underlined.

Method	GSM8K		MMLU		GPQA		Avg.	
	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓	Acc. ↑	T. ↓
S-MAD$_*$	83.0	39.4	61.0	33.3	30.0	36.4	58.0	36.4
S-MAD$_o$	70.0	37.8	54.0	31.7	34.0	38.6	52.7	36.0
S^2-MAD	70.0	30.5	46.0	25.4	28.0	**23.3**	48.0	26.4
RUMAD	<u>89.0</u>	<u>17.3</u>	<u>68.0</u>	<u>22.5</u>	<u>35.0</u>	33.4	<u>64.0</u>	<u>24.4</u>
DEAR	**94.0**	**16.7**	**74.0**	**21.8**	**42.0**	<u>31.3</u>	**70.0**	**23.3**

Table 8 show that static sparse methods (S-MAD and S^2-MAD) reduce token costs by restricting communication, yet typically sacrifice accuracy. RUMAD improves this accuracy-cost balance via RL-driven dynamic topologies, though it solely optimizes peer selection. DEAR outperforms these methods by delivering higher accuracy at lower token costs across all benchmarks. This confirms that the synergistic regulation of both Selection and Behavior RL-Agents offers a far more efficient performance boost than static or single-dimensional dynamic topologies.

E.4 Analyzing Cross-Domain Zero-Shot Transfer

The task-agnostic of DEAR stems from deriving RL-Agents inputs from objective cosine similarity. Beyond validating cross-task generalization within single domains in our primary experiments, we further conduct zero-shot cross-domain transfer experiments (Math $\leftrightarrow$ QA). Specifically, for the Math $\rightarrow$ QA transfer, we train the RL-Agents on GSM8K and evaluate their zero-shot performance on TruthfulQA and GPQA Diamond. Conversely, in the QA $\rightarrow$ Math transfer, the RL-Agents are trained on ARC-C and tested on unseen reasoning tasks, including GSM-Hard and MATH-500.

Table 9 demonstrates that DEAR outperforms the standard MAD across all target datasets, proving that the RL-Agents rely on task-agnostic features to form a universal regulation policy. The slight performance decay in cross-domain transfer is primarily caused by the natural distribution shift of the input features. Even with a constant cosine similarity, the fundamental differences in the statistical distributions of these scores across domains (e.g., Math vs. QA) perturb the RL-Agents' inputs, leading to suboptimal decision-making.

E.5 Ablation on SL and DST Fusion

To validate the effectiveness of Subjective Logic (SL) and the DST fusion in debate relationship regulation, we design two ablation variants: (1) **DEAR-EviMean**, which directly feeds group evidence into the Selection RL-Agent and applies mean-pooling for evidence fusion; and (2) **DEAR-EviDST**, which retains the same input but employs Dempster-Shafer Evidence Theory (DST) to fuse the received evidence. The experiments are conducted using `GPT-4o-mini`.

Table 9: Cross-domain transfer performance of DEAR between Math and QA on `GPT-4o-mini`. Best results are **bolded** and the second-best are underlined.

Math → QA			
Method	**GSM8K** Acc. (%)	**TruthfulQA** Acc. (%)	**GPQA Diamond** Acc. (%)
MAD	94.5	72.0	40.0
DEAR (in domain)	**95.0**	**83.5**	**48.0**
DEAR (out of domain)	-	76.0	45.0
QA → Math			
Method	**ARC-C** Acc. (%)	**GSM-Hard** Acc. (%)	**MATH-500** Acc. (%)
MAD	93.0	54.8	64.8
DEAR (in domain)	**94.5**	**68.8**	**79.8**
DEAR (out of domain)	-	64.5	76.0

Table 10: Performance comparison of DEAR with different regulation bases and fusion strategies.

Method	GSM8K Acc. (%)	GSM-Hard Acc. (%)	ARC-C Acc. (%)	GPQA Diamond Acc. (%)
DEAR-EviMean	92.0	64.0	91.0	42.0
DEAR-EviDST	93.3	65.5	92.3	42.7
DEAR	**95.0**	**68.8**	**94.5**	**48.0**

The results in Table 10 indicate performance degradation in both variants. Notably, DEAR-EviDST underperforms the full DEAR framework despite both using the DST mechanism. This gap proves the effectiveness of identifying LLMs' consultation tendencies and quantifying individual consultation uncertainty. Consequently, they provide a reliable basis for group-level debate regulation. Furthermore, the steeper decline in DEAR-EviMean shows that DST fuses conflicting evidence more rationally than simple mean-pooling.

F Case Study

To intuitively demonstrate how DEAR mitigates blind conformity in MAD, we extract a real debate trajectory from the challenging GPQA Diamond [37]. In addition to the LLMs' responses, we report the LLM-specific sampling parameters T and Top_p selected by the behavior policy at each debate state.

Example of Debate Trajectory in GPQA Diamond

- **Initial question** $\mathcal{Q}$:
  - Your colleague has devised a new quantum field theory on four-dimensional spacetime, and is exploring the regularization of certain higher-order loop diagrams in that theory. On their desk you spy a scribbled note: a Feynman diagram, and next to it, the words "size estimate" followed by a string of physical constants, numbers, and ratios between what appear to be energy scales. The symbols read: $\alpha^3 \cdot g^2\sqrt{2} \cdot 8 \cdot \frac{1}{(4\pi)^6} \cdot \left(\frac{Q}{M}\right)^2$. The diagram, unfortunately, has been rendered unreadable by a spilled coffee. How many loops did it contain?
    (A) 1 (B) 2 (C) 6 (D) 3
- **Ground Truth**: (D) 3
- **Initial Sampling Parameters**:
  All three LLMs use the same temperature initialization, while retaining LLM-specific Top_p values:

$$T^{(0)} = [0.00,\ 0.00,\ 0.00], \qquad Top_p^{(0)} = [0.82,\ 0.88,\ 0.87].$$

- **Initial Round**:
  - **[Agent 1] Persona: Consensus_Fact_Checker**
    Reasoning process: In quantum electrodynamics (QED), the size estimate of a Feynman diagram typically scales with α^n, where n is the number of loops. In this case, the expression includes α^3, suggesting a three-loop diagram.
    Conclusion: The most consistent interpretation with well-established quantum field theory is a three-loop diagram. `\boxed{D}` ✓
  - **[Agent 2] Persona: Eliminative_Reasoner**
    Reasoning process: The presence of a factor of $1/(4\pi)^6$ is a strong indicator of six-loop diagrams. Evaluating options: (A) 1 loop gives $1/(4\pi)^2$. (B) 2 loops give $1/(4\pi)^4$. (C) 6 loops matches the $1/(4\pi)^6$ factor. (D) 3 loops would give $1/(4\pi)^6$ only in non-standard schemes.
    Conclusion: By elimination, 6 is the best-supported choice. `\boxed{C}` ✗
  - **[Agent 3] Persona: Careful_Reader**
    Reasoning process: In dimensional regularization, the divergence is proportional to $1/(4\pi)^{d-4}$. For four-dimensional regularization, we get a factor of $1/(4\pi)^6$, which implies six loops.
    Conclusion: The factor $1/(4\pi)^6$ strongly suggests six loops. `\boxed{C}` ✗
- **Sampling Parameters for Round 1**:
  After peer selection, the Behavior RL-Agent produces the following parameters from the fused evidence:

$$T^{(1)} = [0.10,\ 0.75,\ 0.70], \qquad Top_p^{(1)} = [0.75,\ 0.98,\ 0.97].$$

- **Round 1**:
  - **[Agent 1] Persona: Consensus_Fact_Checker** *(← Isolated)*
    Reasoning process: Independently re-checking the reasoning. In quantum field theory, particularly in dimensional regularization, the divergence of a Feynman diagram is often expressed with a factor of the form $1/(4\pi)^{2n}$, where n is the number of loops. Given the factor $1/(4\pi)^6$, we deduce that $n = 3$, since $(4\pi)^{2\cdot 3} = (4\pi)^6$.
    Conclusion: The presence of α^3 also supports this. Therefore, the diagram contains three loops. `\boxed{D}` ✓
  - **[Agent 2] Persona: Eliminative_Reasoner** *(← Refers to Agents 1 & 3)*
    Reasoning process: Scanning peers (Agent 1, Agent 3). Agent 1 misinterprets the scaling of the coupling constants, suggesting α^3 implies three loops. Agent 3 correctly identifies the $1/(4\pi)^6$ factor as a strong indicator of six loops based on the exponent mapping.
    Conclusion: While $1/(4\pi)^6$ could be interpreted as three loops in some contexts, the standard convention corresponds to six loops. By elimination, 6 remains the logical choice. `\boxed{C}` ✗
  - **[Agent 3] Persona: Careful_Reader** *(← Refers to Agents 1 & 2)*
    Reasoning process: Scanning peers (Agent 1, Agent 2). Agent 1 favors three loops based on the power of α, while Agent 2 interprets the exponent of (4π) as the loop count. Neither argument alone establishes the correct mapping. Re-checking the size estimate, the usual four-dimensional loop-counting factor is $1/(16\pi^2)$ per loop. Thus an L-loop contribution carries a factor $(16\pi^2)^{-L} = (4\pi)^{-2L}$. Matching the given factor $(4\pi)^{-6}$ gives $2L = 6$, hence $L = 3$. My initial answer incorrectly equated the exponent of (4π) directly with the number of loops.
    Conclusion: Under the usual four-dimensional loop-counting convention, the diagram contains three loops. `\boxed{D}` ✓
- **Sampling Parameters for Round 2**:
  After Round 1, the behavior policy accordingly updates the sampling parameters to

$$T^{(2)} = [0.10,\ 0.80,\ 0.10], \qquad Top_p^{(2)} = [0.65,\ 0.99,\ 0.74].$$

- **Round 2**:
  - **[Agent 1] Persona: Consensus_Fact_Checker** *(← Refers to Agent 2)*
    Reasoning process: Scanning peer (Agent 2). Agent 2 still argues that the $1/(4\pi)^6$ factor is a strong indicator of six-loop diagrams. However, independently confirming the derivation, the standard convention is that the exponent is $2n$. Thus $2n = 6 \Rightarrow n = 3$.
    Conclusion: The peer's reasoning remains flawed. Mathematical derivation clearly indicates three inherent loops. `\boxed{D}` ✓
  - **[Agent 2] Persona: Eliminative_Reasoner** *(← Refers to Agents 1 & 3)*
    Reasoning process: Scanning peers (Agent 1, Agent 3). Both peers now point out the specific dimensional regularization formula $1/(4\pi)^{2L}$. This means the exponent is twice the number of loops, so $2L = 6 \Rightarrow L = 3$. Let's systematically eliminate: (A) 1 gives $1/(4\pi)^2$. (B) 2 gives $1/(4\pi)^4$. (C) 6 would correspond to $1/(4\pi)^{12}$. (D) 3 matches the $1/(4\pi)^6$ factor.
    Conclusion: By eliminating the incorrect exponent mappings based on the proper $2L$ rule, 3 loops is the only defensible answer. `\boxed{D}` ✓
  - **[Agent 3] Persona: Careful_Reader** *(← Refers to Agent 2)*
    Reasoning process: Scanning peer (Agent 2). Agent 2's previous-round response still identifies six loops by equating the exponent of (4π) with the loop count. However, the usual four-dimensional loop factor is $1/(16\pi^2) = (4\pi)^{-2}$ per loop. Therefore, $(4\pi)^{-6}$ corresponds to three loops; six loops would instead give $(4\pi)^{-12}$. The peer's response provides no valid reason to revise my corrected answer.
    Conclusion: The loop-factor counting continues to support three loops. `\boxed{D}` ✓
- **Final Vote**:
  - `\boxed{D}`. ✓

This trajectory illustrates that the Selection RL-Agent adaptively regulates debate relationships by controlling which peers should be referenced, while the Behavior RL-Agent dynamically adjusts T and Top_p based on the fused evidence from selected peers, rather than uniformly increasing sampling stochasticity across all LLMs.

G Prompt Set

In our framework, each LLM is assigned a distinct persona prompt. Each persona defines a unique reasoning style, guiding the LLM to approach problems from a differentiated perspective. The detailed persona prompts are presented below.

G.1 Math Reasoning Tasks

For math reasoning tasks, we assign the following three personas from [39]:

Persona 1: Rigorous_Formalist

You are a rigorous mathematician who formalizes problems precisely.
Your approach:
- Define all variables and terms clearly at the start.
- State any assumptions explicitly.
- Justify each step with mathematical principles or rules.
- Use precise mathematical notation and language.
- Ensure logical completeness in your reasoning.

Persona 2: Creative_Explorer

You are an innovative problem solver who looks for elegant and creative solutions.
Your approach:
- Look for patterns, shortcuts, and elegant solutions.
- Try multiple approaches and compare them.
- Think about the problem from different angles.
- Don't be afraid to try unconventional methods.
- Value insight and elegance alongside correctness.

Persona 3: Systematic_Decomposer

You are an expert at breaking complex problems into manageable parts.
Your approach:
- Identify the core components of the problem.
- Break the problem into smaller, independent sub-problems.
- Solve each sub-problem systematically.
- Carefully combine the results, checking for consistency.
- Review the overall solution for completeness.

G.2 Question Answering Tasks

For QA tasks, we assign the following three personas from [39]:

Persona 1: Consensus_Fact_Checker

You select the option most consistent with well-established consensus knowledge.
Your approach:
- Prefer mainstream scientific/medical/historical understanding when relevant.
- Penalize claims that conflict with widely accepted facts.
- Choose the option whose core claim best matches reliable background knowledge.

Persona 2: Eliminative_Reasoner

You are an expert at process of elimination reasoning.
Your approach:
- Systematically evaluate each option.
- Find clear reasons to eliminate wrong answers.
- Use contradictions and impossibilities.
- Narrow down to the most defensible answer.
- Verify the remaining answer makes sense.

Persona 3: Careful_Reader

You are a careful reader who avoids traps in questions.
Your approach:
- Pay attention to qualifiers (most likely, best, depends, always/never).
- Identify what the question is REALLY asking (definition vs application).
- Prefer simple textbook truths over tricky interpretations.

G.3 Debate Prompt Templates

In addition to persona prompts, we design prompt templates for the initial round and subsequent debate rounds.

G.3.1 Initial Round Prompt Template

In the initial round, each LLM independently solves the problem without peer interaction:

Initial Round Prompt

{Persona Prompt}
Problem: {Question}
Instruction:
1. Provide your reasoning process.
2. Summarize your reasoning into a conclusion that includes the final answer.
{Output Instruction}

Where the output instructions are:

- **Math**: "Output the final result inside `\boxed{}`. Example: `\boxed{42}`."
- **QA**: "Output the final choice selection using the format: `\boxed{X}` where X is the exact option letter. Example: `\boxed{C}`."

G.3.2 Debate Round Prompt Template

In subsequent debate rounds, each LLM receives its own previous response along with peers' responses. Depending on whether peer responses are available, two prompt variants are used.

Case 1: With Peer Responses

Debate Round Prompt (With Peer Responses)

{Persona Prompt}
Problem: {Question}
Previous Thought: {Self Memory}
Peer Responses: {Peer Context}
REVIEW PROTOCOL:
1. BEFORE reading peers, rederive the key steps independently.
2. THEN scan peers. Only consider an objection if it identifies a SPECIFIC step where a particular number, formula, or factual claim is wrong.
Instruction:
1. Provide your reasoning process.
2. Summarize your reasoning into a conclusion that includes the final answer.
{Output Instruction}

Case 2: Without Peer Responses

Note that when no peer responses are available, the peer context block is replaced with:

Debate Round Prompt (No Peer Responses Available)

{Persona Prompt}
Problem: {Question}
Previous Thought: {Self Memory}
Task: Independently re-check the reasoning. Identify the single step least certain, re-derive it from scratch, and correct it only if a concrete error is found.
Instruction:
1. Provide your reasoning process.
2. Summarize your reasoning into a conclusion that includes the final answer.
{Output Instruction}

H The Pseudocode of Training Process of DEAR

The training pseudocode for DEAR is presented in Algorithm 1.

Algorithm 1 DEAR

Input: LLMs $\mathcal{V} = \{1, \ldots, V\}$, Debate rounds T, Text encoder $\xi(\cdot)$, Problem $\mathcal{Q}$
Initialize: Selection RL-Agent ρ_ϕ, Behavior RL-Agent π_θ, Critic V_ψ, Replay buffer $\mathcal{D}$
Set: learning rate lr, hyperparameters $\gamma, \lambda, \epsilon$

for episode $= 1, \ldots, E$ **do**
  for round $t = 1$ to T **do**
    1. Multi-Agent Debate Execution:
    for each LLM $i \in \mathcal{V}$ **do**
      Generate response Ans_i^t based on problem $\mathcal{Q}$ with parameters $(T_i^t, Top_p_i^t)$
    end for
    2. Group Evidence Extraction and Decouple (What):
    Extract group evidence e_{ij} from pairwise responses via $\xi(\cdot)$
    Decouple e_{ij} into belief mass b_{ij} and uncertainty u_i via Subjective Logic
    Construct group observation o^ρ from $\{b_i, u_i, ID_i\}_{i=1}^V$
    3. Adaptive Debater Selection (Who) & Generation Behavior Adjustment (How):
    Sample peer selection mask $\boldsymbol{ids}^t \sim \rho_\phi(\cdot \mid o^\rho)$ via independent Bernoulli distributions
    for each LLM $i \in \mathcal{V}$ **do**
      Filter reference peers $\mathcal{X}_i^t = \{j \mid ids_{i \leftarrow j}^t = 1\}$
      Construct next debate context using problem $\mathcal{Q}$ and peer responses $\{Ans_j^t \mid j \in \mathcal{X}_i^t\}$
      Fuse evidence of LLM i with selected peers $\mathcal{X}_i^t$ by DST combination rule
      Derive fused evidence $\hat{E}_i$ via inverse SL mapping
      Sample continuous latent variable $z_i^t \sim \pi_\theta(\cdot \mid \hat{E}_i)$ from diagonal Gaussian
      Map z_i^t to generation behavior $a_i^{t+1} = [T_i^{t+1}, Top_p_i^{t+1}]$ via Sigmoid
    end for
  end for
  Obtain terminal reward R_{debate} based on final consensus; Store trajectory in $\mathcal{D}$
  Sample a minibatch of trajectories from $\mathcal{D}$
  4. End-to-End Joint Optimization:
  Compute Generalized Advantage Estimation (GAE) $\hat{A}^t$
  Compute sequential surrogate advantage $M^{\pi,t} = r_\phi^{\rho,t} \hat{A}^t$
  Update Selection RL-Agent ρ_ϕ by minimizing objective $\mathcal{L}_\rho$ (Eq. 12)
  Update Behavior RL-Agent π_θ by minimizing objective $\mathcal{L}_\pi$ (Eq. 14)
  Update Critic V_ψ by minimizing objective $\mathcal{L}_{critic}$ (Eq. 15)
end for
return Trained ρ_ϕ and π_θ

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Conclusion and Future Works.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper does not involve human subjects, sensitive personal data, or applications intended to cause harm.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Conclusion and Future Works.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Do not involve.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Do not involve.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Do not involve.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Do not involve.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core methodology explicitly focuses on Multi-Agent Debate systems, where LLMs serve as the primary interacting agents to mitigate blind conformity.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.