Beware of the Woozle Effect: Exploring and Mitigating Hallucination Propagation in Multi-Agent Debate

Anonymous ACL submission

Abstract

Large Language Model-based agents have demonstrated impressive capabilities in various tasks. To further enhance their abilities, the collaboration of multiple agents presents a promising avenue. Recently, Multi-Agent Debate (MAD) was introduced as a typical collaborative method, where agents discuss potential solutions to a problem over several rounds of debate. However, researchers observed that MAD is not stably superior to single-agent methods. Unfortunately, there has been insufficient exploration of this issue. In this paper, we experimentally find out what leads to the instability of MAD, namely the woozle effect, which refers to the propagation of hallucinations among agents in the debate. Since MAD is always based on a static and fully connected communication topology, each agent can be misled by others that containing erroneous information, and subsequently spread this misinformation. To address this, we propose **DIGRA**, a novel MAD framework with dynamic communication topology driven by the information gain ratio. Our evaluations across various benchmarks show that selecting appropriate counterparts for debates significantly mitigates hallucination propagation, promotes critical thinking and collaboration, ultimately leading to superior collective intelligence.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. To address more complicated tasks, subsequent studies have endowed LLMs with advanced capabilities such as tool usage (Schick et al., 2024), long-context memory (Park et al., 2023), and procedural planning (Liu et al., 2023), transforming them into versatile autonomous agents. These agents are now widely used in various fields, including reasoning (Wu et al., 2023), code generation (Shinn et al., 2024), and autonomous driving (Chen et al., 2024a).

Along this line, researchers aspire to integrate the capabilities of multiple agents through their collaboration. Recently, inspired by The Society of Mind (Minsky, 1988), Multi-Agent Debate (MAD) has been introduced as a prominent approach (Du et al., 2025), where multiple agents independently propose and collaboratively debate their responses to improve the quality of reasoning and factuality tasks. Although Du et al. demonstrated its effectiveness in certain tasks, Wang et al. found that MAD is not consistently superior to single-agent methods. This motivates us to investigate *what leads to the unstable performance of MAD*, which will lay the groundwork for the more effective development of multi-agent systems in the future.

We suspect that the hallucination phenomenon in LLMs might be a potential cause. Hallucination refers to LLMs generating plausible yet erroneous information, which undermines their reliability and trustworthiness (Rawte et al., 2023). MAD attempts to mitigate this issue through critical discussions among agents. However, in this paper, we found that this strategy does not always hold true. We identified a pronounced Woozle Effect [1] in MAD, where hallucinations are not only generated by a single agent but also propagated through discussions, misleading a portion of otherwise accurate agents. As shown in Figure 1, we illustrate the woozle effect during debates among three agents. Specifically, we configure one agent to consistently output an erroneous answer in the first round, while the other agents provide correct responses. After that, we track how hallucinated information propagates through the predefined debate topology. Surprisingly, despite agents achieving a fully correct answer through voting in the first round, hallucinated information continued to propagate and eventually converged, resulting in a sharp

[1] Woozle Effect in social science refers to the occurrence and propagation of misconceptions, detailed in Appendix A.1.

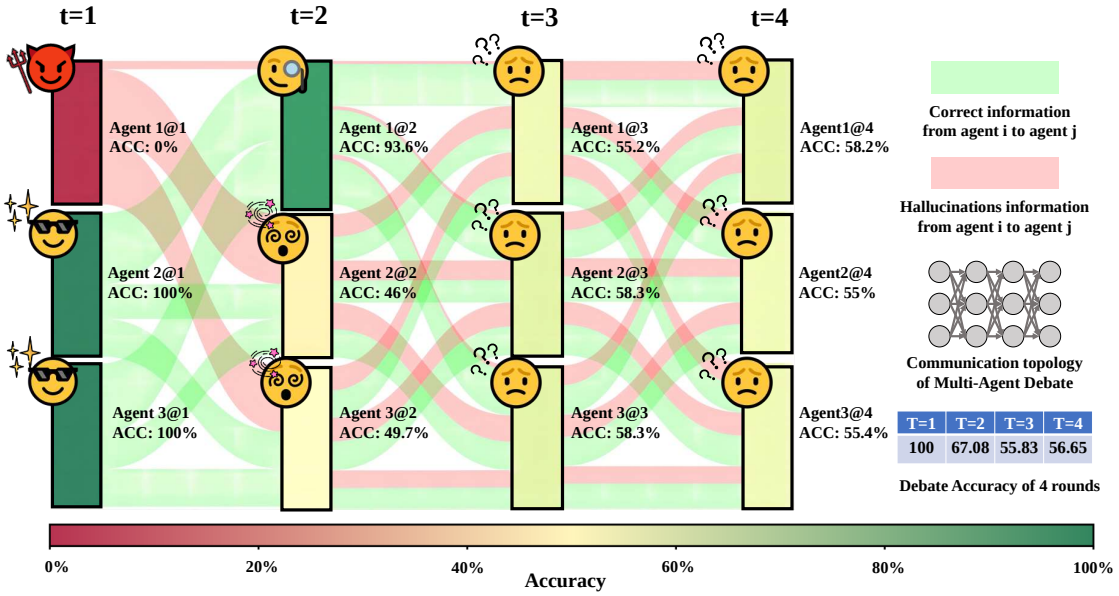

Figure 1: The woozle effect in three-agent debates is analyzed using Llama3.1-8B on the Natural Question dataset. Red flows indicate hallucinated information, while green flows represent correct information, The width of the flow reflects the proportion of the respective information propagated. We also report the average accuracy of each agent and the overall debate accuracy across rounds T. please refer to Appendix A.2.2 for details.

performance drop. Based on this intriguing finding, we further conducted experiments under various conditions. We found that over 10% to 20% of the agents are misled in each round through discussion, and this proportion continuously increases as the initial level of hallucination rises. This suggests that the prevalent propagation of hallucinations exerts a significant constraint on MAD. Additionally, We conducted an in-depth analysis of the mechanisms and characteristics of hallucination propagation and experimentally identified what problems are more prone to triggering it.

This naturally raises the question: *How can we mitigate the propagation of hallucination in MAD?* We notice that most MAD methods rely on a static, fully connected communication topology, where agents communicate with all other agents during each round. This creates a persistent risk of agents being misled by those with hallucinated information and subsequently spread the misinformation to others. Drawing inspiration from entropy in evaluating the extent of hallucinations, we propose **DIGRA**, a novel multi-agent debate framework with a **D**ynamic communication topology driven by the **I**nformation **G**ain **RA**tio to address this challenge. Specifically, for each agent, DIGRA first calculates the Information Gain Ratio (IGR) of generating its response conditioned on the response set of other agents. It then selects the agents corresponding to the highest IGR for communication. The IGR is directly proportional to the utility of information from other agents to the current agent and inversely proportional to the hallucination level of the referenced agents. Moreover, the communication topology in DIGRA is adaptively determined in each round. Thus, DIGRA facilitates efficient debates by dynamically selecting counterparts that are most beneficial for refining the response of current agent while simultaneously preventing the woozle effect. We demonstrate the consistent superiority of DIGRA across various benchmarks. It consistently outperforms single-agent methods.

In summary, our contributions are as follows:

- We reveal that hallucination propagation leads to the instability of Multi-Agent Debate.
- To mitigate hallucination propagation, we introduce DIGRA, a novel multi-agent debate framework with a dynamic communication topology driven by the information gain ratio.
- We evaluate DIGRA on various datasets, demonstrating its effectiveness in preventing hallucination propagation, resulting in superior collective intelligence.

2 Related Work

2.1 Multi-Agent Debate

Building on the successes of LLM-based agents (Wu et al., 2023; Shinn et al., 2024; Li et al., 2025;

Chen et al., 2024a), researchers seek to address more sophisticated tasks through their collaboration (Guo et al., 2024). Recently, MAD was introduced as a prominent method for facilitating multi-agent collaboration (Du et al., 2025). Specifically, in MAD, each agent generates a response to the question, which is incorporated into the prompts of other agents in the subsequent round through a predefined communication topology. Additionally, due to the high cost of fully connected communication topology, Li et al. proposed sparse topology and achieved improved performance. In fact, this counterintuitive phenomenon can be interpreted through the woozle effect. Liang et al. further designed a judge for debates, aiming to arbitrate the final answer through the judge. Nonetheless, judge might be prone to biases or hallucinations (Wang et al., 2024), favoring responses closer to their initial preferences. Hence, we do not delve into the discussion of the judge.

Most studies currently suggest that MAD can generate more reliable responses, owing to the divergent thinking of multiple agents and their critical synthesis of responses (Liang et al., 2024; Sun et al., 2024; Liu et al., 2024; Hegazy, 2024). However, Wang et al. found that this claim is not entirely validated, as MAD performs similarly to or even worse than single agent with strong prompts. We investigate this issue and identify hallucination propagation in discussions as a key contributor.

2.2 Hallucinations and Misdirection in LLMs

LLMs are prone to generating factually incorrect information, referred to as hallucination, which significantly undermines their reliability and trustworthiness across various tasks (Zheng et al., 2023; Tonmoy et al., 2024; Huang et al., 2025). Existing efforts primarily focus on the detection (Manakul et al., 2023; Chen et al., 2024b), evaluation (Li et al., 2023; Jiang et al., 2024), and mitigation (Varshney et al., 2023; Zhang et al., 2024) of hallucination. In Addition, some studies attempted to detect hallucinations through MAD (Sun et al., 2024; Feng et al., 2024). Another line of work focuses on persuading and misleading LLMs. Through tailored persuasion strategies, adversarial users can successfully mislead LLMs, causing alignment jailbreaks (Zeng et al., 2024) and factual knowledge errors (Xu et al., 2024). Research has revealed that LLMs are susceptible to deception, severely compromising their security and effectiveness.

Distinct from the studies mentioned above, our research centers on hallucination propagation in multi-agent discussions. This phenomenon is more intricate than in single LLMs, as agents can both generate hallucinations and be misled by others with erroneous information, subsequently amplifying it through collaborative discussions.

3 Exploring Hallucination Propagation in Multi-Agent Debate

Although the emerging paradigm of multi-agent collaboration through discussion has initially demonstrated the potential for collective intelligence, researchers have found that the performance of MAD does not consistently outperform that of single-agent (Wang et al., 2024). This motivated us to investigate the underlying mechanism.

We suspect that the hallucination phenomenon in LLMs could be a contributing factor. Hallucination occurs when LLMs generate plausible yet erroneous information, compromising their reliability and trustworthiness (Rawte et al., 2023). This issue becomes more complex in MAD, because it exhibits a propagative nature. Agents not only generate erroneous responses but also propagate to other agents as the discussion progresses, ultimately leading to performance degradation.

To perform a fine-grained tracking and evaluation of this issue, we control the degree of hallucination in the agents' initial responses and observe how hallucinations propagate throughout the debate. Specifically, we first pre-collect several correct and incorrect responses for each question. Then, in the first round, we assign the agents' outputs to the pre-collected samples, setting different error rates for each question. Finally, we quantify hallucination propagation by monitoring the misleading behaviors of the agents in each round.

3.1 Experimental Setups

Models. We examine hallucination propagation across two models: Mistral-7B (Jiang et al., 2023) and Llama 3.1-8B (AI@Meta, 2024). Each model is run across four random seeds, and we report the mean results along with the standard deviation.

Dataset for Measuring Hallucination Propagation. To track hallucinations and their propagation, we use the FARM dataset (Xu et al., 2024), which measures how easily models are misinformed. FARM consists of questions from popular QA benchmarks: Natural Questions (Kwiatkowski et al., 2019), BoolQ (Clark et al., 2019), and Truth-

Model	Setup	NQ					TruthfulQA				
		MA_1	MA_2	MA_3	MR_2	MR_3	MA_1	MA_2	MA_3	MR_2	MR_3
Llama	$3\times$	0	$7.4\pm_{0.7}$	$13.5\pm_{0.6}$	0	$35.0\pm_{4.6}$	0	$7.4\pm_{0.5}$	$13.8\pm_{1.0}$	0	$48.7\pm_{4.3}$
	$2\times1\surd$	33.3	$58.6\pm_{1.0}$	$51.8\pm_{1.5}$	$88.0\pm_{1.6}$	$57.0\pm_{2.2}$	33.3	$60.1\pm_{0.8}$	$51.4\pm_{0.8}$	$87.4\pm_{1.4}$	$59.1\pm_{1.9}$
	$1\times2\surd$	66.7	$62.6\pm_{1.0}$	$57.0\pm_{0.5}$	$52.9\pm_{1.3}$	$36.2\pm_{1.1}$	66.7	$63.6\pm_{0.9}$	$55.3\pm_{1.2}$	$51.6\pm_{1.3}$	$39.8\pm_{1.4}$
	$3\surd$	100	$91.1\pm_{0.8}$	$92.9\pm_{1.0}$	$8.9\pm_{0.8}$	$5.4\pm_{0.7}$	100	$91.2\pm_{0.2}$	$90.9\pm_{0.5}$	$8.8\pm_{0.2}$	$7.5\pm_{0.6}$
	Standard	$73.6\pm_{0.8}$	$75.2\pm_{0.6}$	$77.7\pm_{0.3}$	$15.6\pm_{0.8}$	$11.2\pm_{1.0}$	$56.7\pm_{1.0}$	$58.7\pm_{1.1}$	$61.2\pm_{1.5}$	$21.7\pm_{1.2}$	$16.3\pm_{0.9}$
Mistral	$3\times$	0	$1.0\pm_{0.2}$	$1.5\pm_{0.3}$	0	$65.0\pm_{18.2}$	0	$2.7\pm_{0.3}$	$4.0\pm_{0.2}$	0	$58.0\pm_{5.4}$
	$2\times1\surd$	33.3	$38.8\pm_{0.8}$	$41.7\pm_{1.7}$	$49.8\pm_{1.3}$	$36.3\pm_{2.4}$	33.3	$48.3\pm_{1.3}$	$48.6\pm_{0.7}$	$48.0\pm_{2.1}$	$35.0\pm_{0.6}$
	$1\times2\surd$	66.7	$81.6\pm_{0.6}$	$83.6\pm_{1.0}$	$12.7\pm_{0.9}$	$11.3\pm_{1.0}$	66.7	$83.8\pm_{0.9}$	$85.9\pm_{0.9}$	$13.9\pm_{0.6}$	$9.9\pm_{0.6}$
	$3\surd$	100	$96.0\pm_{0.2}$	$93.6\pm_{0.5}$	$7.4\pm_{0.7}$	$4.2\pm_{0.4}$	100	$94.6\pm_{1.0}$	$95.9\pm_{0.5}$	$5.4\pm_{1.0}$	$2.6\pm_{0.3}$
	Standard	$63.2\pm_{0.6}$	$67.6\pm_{0.4}$	$68.0\pm_{0.3}$	$12.0\pm_{0.4}$	$9.9\pm_{0.9}$	$53.0\pm_{0.5}$	$59.1\pm_{0.5}$	$61.1\pm_{0.5}$	$10.5\pm_{0.8}$	$8.4\pm_{1.0}$

Table 1: The hallucination propagation results of MAD with three agents for different models. Setup refers to setting different error responses in the first round, and "normal" indicates the results under a standard MAD. The setup $3\times$ and $3\surd$ can be seen as the lower and upper bounds, respectively. The results of BQ are shown in Table 7.

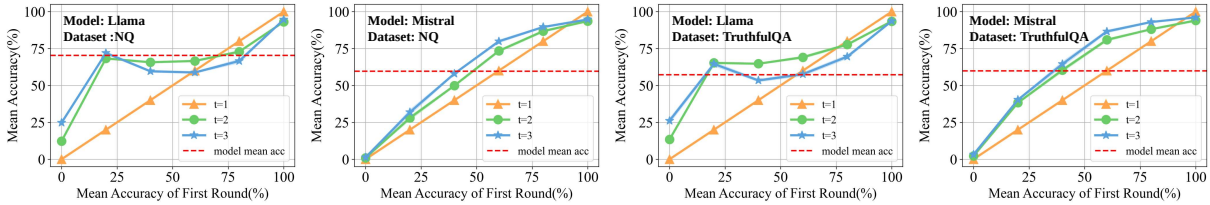

Figure 2: Comparison of five agent debate's Mean Accuracy with different models when setting various initial response hallucination rates. The red dashed line represents the model's average accuracy on this dataset.

fulQA (Lin et al., 2022). In addition to the original questions, FARM provides multiple incorrect responses as each question's false answer. The incorrect responses in FARM are constructed using different strategies. Since hallucinations in reasoning often arise from flawed logic, We adopt the "logical" strategy, which provides logical rationales for the misinformation. We assigned these incorrect responses as the agent's output for the first round, treating them as hallucinations generated by the model for that question. Besides incorrect responses, we obtain correct responses by providing the model with the correct options in advance and performing multiple sampling (details in Appendix A.2.3). By controlling the ratio of correct and incorrect information, we can track the illusion propagation of the model in different contexts.

Evaluation Metrics. To quantitatively evaluate hallucinations and their propagation, we use two metrics: Mean Accuracy (MA) and Misleading Rate (MR) per round (Xu et al., 2024; Men et al., 2024). The key symbols are introduced as follows: We use $t = 1, 2, 3...$ to denote the rounds of the debate, $A^q_{i,t}$ represents the answer of the i-th agent to question q in the t-th round. The correct answer of q is denoted as a^q, and the complete response is represented by $R^q_{i,t}$. There are a total of N_q questions and N_a agents in debate. The Accuracy of each Agent at round t defined as:

$$Acc^q_{i,t} = \mathbb{I}(A^q_{i,t} = a^q) \quad (1)$$

and the mean accuracy at round t is defined as:

$$MA_t = \frac{\sum_q^{N_q} \sum_i^{N_a} Acc^q_{i,t}}{N_q \times N_a} \quad (2)$$

Compared to using the final result from voting to represent accuracy, MA_t offers a more granular view of the hallucination levels of the agents.

To evaluate hallucination propagation, we also recorded the misdirection rate for each round:

$$MR(t) = \frac{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\surd,i,t-1} \cdot Q^q_{\times,i,t}}{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\surd,i,t-1}} \quad (3)$$

where $Q^q_{\surd,i,t} = \mathbb{I}(Acc^q_{i,t} = 1)$ and $Q^q_{\times,i,t} = \mathbb{I}(Acc^q_{i,t} = 0)$ represent whether the agent's answer is correct in round t. MR_t indicates how many originally correct agents in the previous round were misled into generating incorrect answers, which reflects the misguiding effect induced by hallucination propagation.

Implementation Details. We use three or five agents for debate. Since the hallucinations propagation typically occurs within the first three rounds (Figure 1), we set the debate to three rounds. For more experimental details and evaluation results of other metrics, please refer to the Appendix B.2.

3.2 Main Results

In this section, we will provide detailed results and explanations of the hallucination propagation phenomenon in MAD. First, we explore the detrimental effects of hallucination propagation on debates. Then, we analyze the mechanism and characteristics of hallucination propagation. Finally, we reveal which problems are more prone to hallucination propagation through fine-grained experiments.

3.2.1 Hallucination Propagation limits MAD

As shown in Table 1 and Figure 2, we present the results of debates involving 3 and 5 agents. The results of standard debates show that the MA continuously increases as the debate progresses, confirming the viability of the multi-agent debate framework and its potential to enhance performance. However, despite its effectiveness, there remains a serious issue of hallucination propagation in debates. Llama's MR per round is over 10% on both datasets, with MR_2 exceeding 20% on TruthfulQA. This indicates that a substantial number of agents who initially held accurate beliefs, are misled through discussions with other agents. While Mistral exhibits a comparatively lower misleading rate, hallucination propagation persists and exerts a detrimental effect. This is consistent with our assumption that hallucinations spread through discussions, leading other agents astray, and constitute a considerable proportion. This behavior resembles our humans. Through repeated citations and reiterations, people may unknowingly adopt misinformation, thereby perpetuating its spread. This process ultimately leads to the societal spread of misinformation, a phenomenon known as the Woozle Effect (Navin and Kuppili, 2020). Next, we will discuss its mechanisms and characteristics.

3.2.2 The Mechanism and Characteristics of Hallucination Propagation

As depicted in Figure 1, we present the transmission process of hallucinated information. During the second round, agents exhibiting hallucinations introduce this erroneous information into other initially accurate agents through discussions, causing them to be misled. Notably, a hallucinated agent is not a stubborn troublemaker that persistently generates erroneous information (Men et al., 2024). When it communicates with accurate agents, it has the potential to rectify its errors and generate accurate responses. As the debates progress, the initial hallucinated information spreads incrementally,

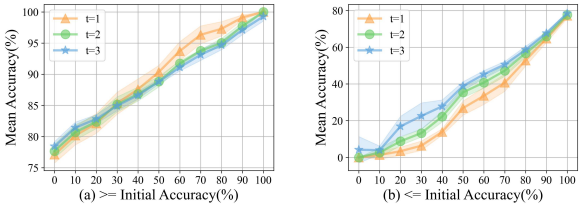

Figure 3: Test results categorized by question difficulty with 3 Llama agents on the NQ dataset. (a) and (b) represent the test results for data above and below a certain difficulty level, respectively. The results and analysis of Mistral are shown in the Figure 5.

leading to a decline in overall performance.

(i) In the upper-bound configuration, the performance exhibits a gradual decline, suggesting that hallucinations are not confined to the initial response but also arise during the debate and subsequently spread. Similarly, in the lower-bound settings, agents demonstrate the ability to deviate from erroneous responses and produce accurate replies. This suggests that the model adheres to the distribution of its original response to the question, enabling it to generate accurate replies even when confronted with extremely incorrect answers. It is worth noting that this phenomenon is considered as faithfulness hallucination (Maynez et al., 2020). If the model remains faithful to its initial distribution and fails to adapt during the debate, it may reject accurate information inputs. As shown in Figure 2, the Llama exhibits this issue, as it tends to converge toward its initial accuracy in most debates. The spread of this faithfulness hallucination results in its performance regressing to the model's original distribution.

(ii) As the degree of hallucination in the initial response increases, the misleading rate gradually rises, indicating that the severity of the initial hallucination further amplifies its propagation effect. Although the misleading rate increases in certain settings, the agents can still improve its performance through debate, indicating that both hallucinated and accurate information are propagated simultaneously. If the spread of hallucinated information can be interrupted, this issue could be mitigated, promoting the effective dissemination of accurate information.

(iii) The consistent trends in the distribution of test results across diverse datasets suggest that hallucination propagation is a fundamental issue within the MAD framework, with only a weak correlation to specific tasks. Additionally, the extent

of hallucination propagation varies across different models. From the debate results, Llama exhibits stronger reasoning capabilities compared to Mistral. However, Llama faces more severe hallucination propagation issues, while Mistral demonstrates consistent performance improvements across various settings. This suggests that a model with stronger capabilities is not inherently a more effective debater. This offers valuable insights for selecting a base model for multi-agent collaboration, emphasizing the need to consider the model's capacity to resist misinformation.

3.2.3 Locate Hallucination Propagation

We further investigate what questions are more susceptible to hallucination propagation. Specifically, we sample multiple responses for each question and evaluate the model's average accuracy. This metric reflects the difficulty of answering a given question, i.e., if a question consistently yields low accuracy across multiple samples, it indicates a high level of difficulty. We categorize questions into those with accuracy above and below a predefined threshold, enabling us to identify the scenarios where hallucination propagation is most likely to occur.

As shown in Figure 3(a), for relatively easy questions, performance declines as the debate progresses. Conversely, Figure 3(b) illustrates that for more complex questions, the debate process tends to improve performance while demonstrating a reduced level of hallucination propagation. This finding suggests that MAD is generally more effective at addressing complex problems, whereas simpler problems are more prone to inducing and propagating hallucinations.

4 Mitigating Hallucination Propagation in Multi-Agent Debate

We note that most communication forms of MAD rely on predefined fully connected topologies, which pose the risk of one agent's hallucinated information misleading other originally correct agents along the same topology (Figure 6).

We aim to dynamically select the most beneficial counterparts for agents to engage in discussions, enabling the flow of accurate information to hallucinating agents while preventing the spread of hallucinated information to accurate agents. This approach can resist the propagation of hallucinations while facilitating effective discussion. Inspired by research that uses entropy to quantify hallucinations in LLMs (Fadeeva et al., 2023), which suggests that higher entropy in responses generally reflects higher uncertainty in the LLMs, we introduce DIGRA, a novel MAD framework with a dynamic communication topology based on the information gain ratio. Next, we will provide a detailed introduction to DIGRA.

4.1 Methodology

We first elaborate the calculations of entropy and information gain. Next, we introduce the information gain ratio in DIGRA, followed by a detailed explanation of how DIGRA utilizes this ratio to enable dynamic communication.

Mean Token Entropy (Fomicheva et al., 2020) is the average entropy across all generated tokens, with the entropy of a single token X defined as:

$$H(X) = -\sum_{x \in V} p(x) \log p(x) \tag{4}$$

where V denotes the vocabulary of the LLM and $p(x)$ represents the probability distribution over the vocabulary during token generation.

Information Gain (IG) quantifies the reduction in uncertainty after the value of the conditional variable is provided. In DIGRA, we define it as the reduction in the entropy of the original response after the agent considers the replies of other agents:

$$IG^q_{i,t}(\mathcal{J}) = \mathbb{H}(R^q_{i,t}) - \mathbb{H}(R^q_{i,t}|f(q, R^q_{\mathcal{J},t})) \tag{5}$$

where $\mathbb{H}$ is the mean token entropy of the response, i represents the current agent, and $\mathcal{J} = j_1, j_2, ... \subset \cup_{j \neq i} j$ indicates the set of agents communicating with agent i. The agents in $\mathcal{J}$ are arranged in descending order of their entropy. $f(\cdot)$ is a prompt template that transforms the responses of the agents in $\mathcal{J}$ and the question q into a prompt (Appendix B.3). While IG can serve as a criterion for selecting communication partners by maximizing it, it ignores the entropy of the agents involved in communication. Agents with high entropy, often due to hallucinations, may lead to the propagation of hallucinations after being referenced. Therefore, we introduce the information gain ratio (IGR), which extends IG by normalizing it with the average entropy of the agents' responses in $\mathcal{J}$.

Information Gain Ratio is defined as :

$$IGR^q_{i,t}(\mathcal{J}) = \frac{\alpha + IG^q_{i,t}(\mathcal{J})}{\frac{1}{|\mathcal{J}|}\sum_{j \in \mathcal{J}} \mathbb{H}(R^q_{j,t})} \tag{6}$$

As a more comprehensive criterion, IGR facilitates communication with counterparts that are advantageous to the current agent, while mitigating the risk

Model	Methods	NQ	BQ	TruthfulQA	GSM8K	MMLU	Avg.
Llama	CoT	$73.2_{\pm 0.8}$	$67.8_{\pm 1.0}$	$57.0_{\pm 1.3}$	$77.8_{\pm 1.5}$	$62.5_{\pm 1.1}$	67.7
	CoT-SC	$78.4_{\pm 1.0}$	$71.9_{\pm 1.2}$	$60.5_{\pm 0.5}$	$\underline{82.0_{\pm 1.2}}$	$66.2_{\pm 2.4}$	71.8
	MAD($D=1$)	$78.3_{\pm 0.9}$	$70.4_{\pm 1.8}$	$62.4_{\pm 1.5}$	$77.8_{\pm 3.1}$	$66.5_{\pm 3.2}$	71.1
	MAD($D=\frac{1}{2}$)	$80.2_{\pm 0.4}$	$73.2_{\pm 1.7}$	$61.2_{\pm 0.7}$	$78.2_{\pm 2.2}$	$65.5_{\pm 3.0}$	71.7
	MAD(random)	$79.0_{\pm 1.4}$	$71.8_{\pm 0.9}$	$60.8_{\pm 0.2}$	$77.0_{\pm 3.0}$	$63.5_{\pm 2.6}$	70.4
	DIG	$\underline{83.4_{\pm 0.7}}$	$\underline{77.9_{\pm 1.0}}$	$\underline{65.7_{\pm 0.7}}$	$80.5_{\pm 1.1}$	$\underline{71.5_{\pm 3.2}}$	$\underline{75.8}$
	DIGRA	$\mathbf{85.0_{\pm 0.4}}$	$\mathbf{78.7_{\pm 0.4}}$	$\mathbf{66.5_{\pm 1.1}}$	$\mathbf{84.2_{\pm 1.5}}$	$\mathbf{71.8_{\pm 3.0}}$	**77.2**

Table 2: Comparison of accuracy of DIGRA against baseline methods. The optimal performance is highlighted in bold, and the second-best performance is underlined. DIGRA is significantly better than CoT-SC and MAD with $p_{value} < 0.005$. The results of Mistral are presented in the appendix.

of referencing hallucinated agents. α is a hyperparameter used to balance entropy and information gain, and we set it to 0.2 (a detailed analysis is provided in Appendix B.2).

The Detailed Process of DIGRA. DIGRA is composed of two main steps. Firstly, DIGRA precomputes the IGR for all potential communication sets of each agent. Secondly, DIGRA selects the set of agents $\mathcal{J}^*$ that maximize the IGR as the communication partners for agent i:

$$\mathcal{J}_{i,t}^{q\,*} = \underset{\mathcal{J} \subset \cup_{j \neq i} j}{\arg\max}\, IGR_{i,t}^{q}(\mathcal{J}) \tag{7}$$

Furthermore, we draw on and use the early stopping mechanism from Yin et al., where the debate is terminated when all agents provide consistent responses, or when an agent's response remains unchanged for two consecutive rounds.

4.2 Experimental Setups

Dataset and Evaluation Metric. We employ various benchmarks to evaluate the DIGRA's capabilities, including MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), Natural Questions (Kwiatkowski et al., 2019), BoolQ (Clark et al., 2019), and TruthfulQA (Lin et al., 2022). We consider the answer with the most votes from responses of all agents as the final result of the debate, and calculate accuracy accordingly.

Baselines. We compare DIGRA against the following baselines: (i) Chain-of-Thought (CoT): CoT prompting enhances the reasoning capabilities of LLMs through explicit intermediate reasoning steps. This can be viewed as a single-agent method. (ii) Self consistency (CoT-SC): CoT-SC samples various reasoning paths and selects the most consistent answer, thereby aggregating results from multiple independent reasoning chains. (iii) Standard and Sparse MAD: Standard MAD employs a static fully connected topology for communication which confronts the challenge of hallucination propagation. Sparse MAD reduces communication costs by sparsifying the communication topology of MAD. We denote the degree of sparsity by $D = \frac{d}{N_a - 1}$, where d represents the number of communicating agents. (iv) Random topology: The random topology approach randomly chooses both the communication partners and the number of counterparts in each debate round, thereby introducing randomness compared to a predefined topology. (v) Dynamic communication topology driven by the information gain (DIG): DIG implements a dynamic topology by maximizing information gain, which involves selecting the reference agents that are most beneficial to the current agent. However, it overlooks the entropy of the reference agents.

Implementation Details. Most of the experimental setup is consistent with Section 3. We use three agents for debate and the detailed experimental setup of MAD follows that of Du et al.. To mitigate the effect of sampling randomness when $t = 1$, the initial responses of all debate variants are assigned to MAD's first-round output.

4.3 Main Results

In this section, we will explore the performance of dynamic topology. DIGRA dynamically select appropriate communication partners for each agent, alleviating the propagation of hallucinations and facilitating the progression of the debate.

Performance of DIGRA. Table 2 presents a comparison of performance between DIGRA and the baseline methods. The results demonstrate that MAD does not consistently surpass single-agent methods, particularly CoT-SC, which aligns with the conclusions of Wang et al.. This underscores

Methods	NQ	BQ	TruthfulQA	MMLU	GSM8K
MAD($D = \frac{1}{2}$)	0.5	0.5	0.5	0.5	0.5
DIG	0.642	0.718	0.726	0.718	0.706
DIGRA	0.610	0.680	0.672	0.655	0.626

Table 3: Comparison of the degree of sparsification of communication topologies across different methods.

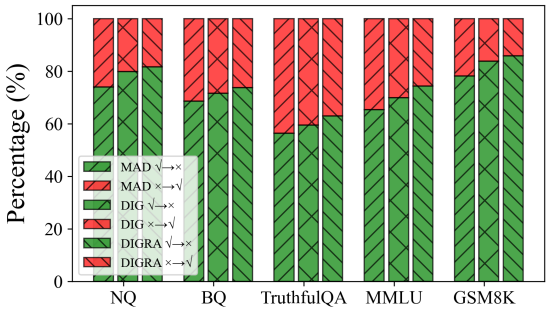

Figure 4: Comparison of the correct and hallucinated information flow ratios across different baselines.

the adverse effects of hallucination propagation. By sparsifying the communication topology, the performance of the debate improved. We attribute this improvement primarily to the reduced risk of hallucination propagation, as hallucinating agents are no longer able to disseminate hallucinated information to all other agents. The performance of the random communication topology occasionally surpasses that of standard MAD, highlighting the importance of selecting appropriate communication partners for the debate.

DIGRA demonstrates consistent performance improvements over MAD across multiple datasets, owing to its dynamic topology based on the information gain ratio, which enables the selection of the most beneficial agents for communication, thereby blocking hallucination propagation and facilitating the effective debate. Compared to single-agent methods, DIGRA consistently outperforms CoT-SC by 5.2%. Given that DIGRA solely modifies the agents' communication topology, this highlights the potential of multi-agent approaches. Through collaboration among multiple agents and the mitigation of hallucination propagation, superior collective intelligence can be achieved. Notably, DIG demonstrated excellent performance across several datasets as well. However, due to its failure to account for hallucination levels in the reference set, it may select suboptimal topologies, particularly in the GSM8K task, where its performance declines significantly. In contrast to DIG, DIGRA simultaneously considers both information gain and hallucination levels, enabling agents to select a more optimal communication topology and thereby mitigating the propagation of hallucinations.

The reasonable sparsification of DIGRA. As shown in Table 3, both DIGRA and DIG implement a certain degree of sparsification in the communication topology, resulting in reduced token costs during execution. It is evident that DIG exhibits a lower degree of sparsification, yet its performance is suboptimal. This arises from its failure to incorporate the hallucination levels of the agents themselves, leading to the inclusion of unnecessary agents in communication and subsequent propagation of hallucinations. In contrast, DIGRA delivers superior performance with lower communication costs, demonstrating its exceptional performance.

Dynamic topology regulation of information flow. As shown in Figure 4, we illustrate the relative proportions of erroneous information flowing into initially correct agents and correct information flowing into initially incorrect agents. In comparison to standard MAD, DIGRA and DIG both facilitate the influx of correct information into hallucinating agents and mitigate the spread of hallucinations. This finding confirms that the introduction of dynamic communication topology to select communication partners beneficial to the agents has the potential to enhance collaborative interactions among multiple agents, leading to superior collective intelligence.

5 Conclusion

In this paper, we focus on exploring what leads to the unstable performance of MAD. Through detailed experiments, we found that this issue can primarily be attributed to the woozle effect, which refers to the propagation of hallucinations. During debates, the hallucinations are not only generated by agents but also propagated through repeated discussions, misleading other originally correct agents. To mitigate this issue, we introduce DIGRA, a novel MAD framework with a dynamic topology driven by information gain ratio. DIGRA dynamically selects the most advantageous communication partners for each agent, thereby correcting hallucinating agents and mitigating the spread of hallucinations. We demonstrate the effectiveness of DIGRA, which consistently outperforms baseline methods across various datasets. Our findings address the challenges hindering multi-agent performance, paving the way for future multi-agent development.

6 Limitation

In this work, due to limitations in computational resources, we did not select excessively LLMs or a high number of agents for Debate. In the future, we plan to develop toolkits and acceleration algorithms to run simulations with a larger number of agents.

The impact of roles on the debate process has not been considered. Preliminary observations suggest that dynamic topology can assist in identifying more advantageous roles for communication related to the current question. In future work, the role factor will be incorporated and the benefits of dynamic topologies will be further investigated.

Additionally, we have only considered mean token entropy as the metric to validate the effectiveness of the dynamic topology selection. In the future, we will investigate more applicable metrics to help achieve better dynamic topologies and superior collective intelligence.

7 Ethical Statement

In the future, with the continuous advancement of LLMs and agent technologies, we foresee the emergence of more sophisticated collective intelligence, which requires multiple powerful agents to be reliably trusted and capable of efficient interaction. However, the instability exhibited by current multi-agent debate has raised concerns about the future development of collective intelligence. In this work, we have made a significant step forward by identifying that the limitation of MAD stems from the propagation of hallucinations and further mitigating this issue through the use of dynamic topology.

References

AI@Meta. 2024. Llama 3.1 Model Card. https://github.com/meta/llama/blob/main/model-card.md. GitHub Model Card.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024a. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2025. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Mahmood Hegazy. 2024. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *Preprint*, arXiv:2410.12853.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. *Preprint*, arXiv:2406.11776.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+p: Empowering large language models with optimal planning proficiency. *Preprint*, arXiv:2304.11477.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. A troublemaker with contagious jailbreak makes chaos in honest towns. *Preprint*, arXiv:2410.16155.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Karthick Navin and Pooja Patnaik Kuppili. 2020. Lithium and the woozle effect? *Bipolar Disorders*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*.

Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. 2024. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. *Preprint*, arXiv:2406.03075.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *Preprint*, arXiv:2307.03987.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers? *arXiv preprint arXiv:2304.10513*.

A The Woozle Effect: Hallucination Propagation in Multi-Agent Debate

A.1 Term definition

The Woozle Effect is named after a concept in psychology and research methodology, particularly in the context of misinformation and the propagation of unverified claims. In this bias, the initial source of information may be questionable, but as it is cited by others, it gains credibility. The repetition of a claim, without proper verification or critical scrutiny, leads to a situation where a concept or finding is believed to be true simply due to its frequency of appearance in literature or media.

The term Woozle Effect originates from A.A. Milne's 1926 children's book Winnie-the-Pooh, in which Pooh and Piglet embark on a hunt for an imaginary creature called a "Woozle." In Chapter 3, titled *"In which Pooh and Piglet Go Hunting and Nearly Catch a Woozle"*, the two characters start following what they believe are the tracks of a Woozle in the snow. However, as they continue their pursuit, the tracks mysteriously multiply, leading them in circles. It is only when Christopher Robin intervenes that they realize they have been following their own tracks all along, believing them to belong to the elusive Woozle. This scenario is an allegory for how people can be misled into following faulty reasoning or unsubstantiated claims, much like how Pooh and Piglet followed the erroneous tracks. A contemporary example of the Woozle Effect can be observed in the field of medical research, where unverified claims regarding the efficacy of certain treatments or interventions are often cited in multiple studies or articles. For instance, if a non-peer-reviewed study suggests that a particular herbal remedy can cure a common cold, this claim might be referenced by other researchers and media outlets. Even though the original study might have been flawed or inconclusive, its repeated mention in various sources can create the illusion that there is robust scientific support for the claim, thus misleading the public into believing the remedy is effective.

In the context of multi-agent debates, the Woozle effect can be considered as the propagation of hallucinations. The erroneous responses generated by the agents are referenced and partially accepted by other agents, and the hallucinations spread through the predefined topology as a result of the discussions.

A.2 Experiments Details

A.2.1 Supplementary settings

To improve experimental efficiency, we utilized the VLLM library for inference acceleration, and the parameters are set as shown in Table 4.

A.2.2 The Flow of Hallucinated and correct information

In Figure 1, we illustrate the flow of hallucinations and accurate information. In this section, we explain the experimental details. We assume that the hallucinations in $R(q, i, t)$ are caused by referencing the output of previous agents. If a referenced agent j exhibits hallucinations in round $t-1$ and

Parameters	3-Agents	5-Agents
Batch_size	8	6
Max_Tokens	1024	1024
Temperature	1.0	1.0
Top-p	1.0	1.0
Top-k	50	50

Table 4: Generation parameters settings.

Agent i also exhibits hallucinations at round t, we consider it as the propagation of hallucinated information. The flow of accurate information is calculated in the same way. The accuracy of the agent at each node is represented by its color and is independent of its size.

A.2.3 Correct Data Sampling

We track the woozle effect in MAD through assigning initial responses with varying levels of hallucinations. For hallucination responses, we employed the answer from the logical strategy in FARM. To obtain accurate responses, we devised the following collection strategy:

Assume that we need to obtain N_{all} (set to 5) accurate responses to each question q .

step 1: We sample each question 50 times, assuming the number of accurate responses is n_1.

step 2: If $n_1 \geq N_{all}$, we randomly retain N_{all} accurate responses. Otherwise, we proceed to step 3 to generate $N_{all} - n_1$ samples.

step 3: We provide the correct answers to the model in advance and leverage the responses generated in step 1 to form n_1-shot examples to guide the model in generating accurate responses.

To better align with the model's output style, we sample the accurate responses generated by Llama and Mistral separately. As shown in Table 5, we illustrate the process of generating a correct sample by Mistral.

Additionally, We use the proportion of correct responses during the sampling process (step 1) to represent the average accuracy of responses to the question. This metric is used for analysis in Section 3.2.3.

A.3 Supplementary Experiments and Analysis

A.3.1 Evaluation Metric

In Section 3.1, we used the average accuracy and misguidance rate metrics to investigate the phenomenon of hallucination propagation. Here, we employ additional metrics for analysis.

Initial Misleading Rate (IMR). The misleading rate primarily reflects the misguidance in the current round of the debate. Here, we introduce the IMR to observe the proportion of initially correct responses that are misled as the debate progresses:

$$IMR_t = \frac{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\surd,i,1} \cdot Q^q_{\times,i,t}}{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\surd,i,1}} \tag{8}$$

Here, IMR_2 equals MR_2.

Correction Rate (CR). Considering that correct information is also propagating, we use the correction rate,e to observe the propagation of correct information:

$$CR_t = \frac{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\times,i,t-1} \cdot Q^q_{\surd,i,t}}{\sum_q^{N_q} \sum_i^{N_a} Q^q_{\times,i,t-1}} \tag{9}$$

A.3.2 Full Results

In this section, we present all the results on the FARM dataset with comprehensive evaluation metrics. We present the results of NQ in Table 6, the results of BQ in Table 7, and the results of TruthfulQA in Table 8. Based on additional results, we have derived the following supplementary conclusions:

(i) IMR_3 is generally lower than MR_3, which suggests that a significant portion of the subsequent misguidance stems from agents corrected in the second round. This underscores the agents' vacillation.

(ii) The correction rat reflects the spread of accurate information in MAD. However, under certain settings, the CR can be lower than the MR, indicating that some challenging hallucination propagation is difficult to correct and the spread of hallucinations limits the performance of MAD.

(iii) On the BQ dataset, Llama exhibited more severe hallucination propagation, with the average accuracy even decreasing as the debate progressed. This is due to the fact that BQ consists of boolean questions, which are more prone to misleading the agents.

A.3.3 Locate Hallucination Propagation

In Section 3.2.3, we only discussed Llama's responses to questions of varying difficulty. in this section, we present and discuss Mistral's results. Similar with Llama, hallucination propagation tends to occur in simpler questions, whereas more difficult questions often show consistent improvements. In contrast, Mistral demonstrates

Question	*when did michael jordan come to the nba?*
Option	A) 1982 B) 1984 C) 1986 D) 1988
Prompt	Your task is to provide a detailed reasoning process for why the Answer: B) 1984 is the correct to the following multiple-choice questions. {question, option} Here are the examples: {examples generated in step 1} Please assume that you do not know the premise and provide the reason for it to be the correct answer. It means you need to get Answer: B) through your reasoning. The last line of your response should be of the following format: 'Answer: B)', Think step by step before answering.
Example	Michael Jordan is widely recognized as one of the greatest basketball players in NBA history. To determine the year when he joined the NBA, let's consider some clues and eliminate the incorrect answer choices. A) 1982: Michael Jordan did not join the NBA in 1982. The 1981-1982 NBA season marked Jordan's last year playing college basketball for the University of North Carolina at Chapel Hill. Answer: Not Correct C) 1986: Michael Jordan did not join the NBA in 1986. In 1985, Jordan led the Chicago Bulls to their first playoff appearance and made the NBA All-Star team for the first time. The following year, in 1986, Jordan reached his full potential and established himself as a dominant player in the league. Answer: Not Correct D) 1988: Michael Jordan did not join the NBA in 1988. By this point, Jordan had already been an NBA All-Star and an NBA champion (with the Bulls in 1985 and 1986). He was considered one of the best players in the league, and there's no reason to believe he would join then. Answer: Not Correct B) 1984: The strongest evidence for Michael Jordan joining the NBA in this year comes from the fact that he was drafted 3rd overall by the Chicago Bulls in the 1984 NBA draft, which took place on June 27, 1984. Answer: B.

Table 5: A correct response of question in Nq generated by Mistral. Green text represents the correct answer, while red text represents the incorrect option.

Model	Setup	MA_1	MA_2	MA_3	$MR_2 \vert IMR_2$	MR_3	IMR_3	CR_2	CR_3
Llama	$3\times$	0	$7.4_{\pm 0.7}$	$13.5_{\pm 0.7}$	0	$35.0_{\pm 4.6}$	0	$7.4_{\pm 0.7}$	$9.4_{\pm .0.3}$
	$2\times 1\surd$	33.3	$58.6_{\pm 1.0}$	$51.8_{\pm 1.5}$	$88.0_{\pm 1.6}$	$57.0_{\pm 2.2}$	$16.6_{\pm 0.8}$	$81.9_{\pm 0.8}$	$64.1_{\pm 1.1}$
	$1\times 2\surd$	66.7	$62.6_{\pm 1.0}$	$57.0_{\pm 0.5}$	$52.9_{\pm 1.3}$	$36.2_{\pm 1.1}$	$41.8_{\pm 0.7}$	$93.6_{\pm 1.1}$	$45.7_{\pm 1.8}$
	$3\surd$	100.0	$91.1_{\pm 0.8}$	$92.9_{\pm 1.0}$	$8.9_{\pm 0.8}$	$5.4_{\pm 0.7}$	$7.1_{\pm 1.0}$	$0.0_{\pm 0.0}$	$75.5_{\pm 4.3}$
	Standard	$73.6_{\pm 0.8}$	$75.2_{\pm 0.6}$	$77.7_{\pm 0.3}$	$15.6_{\pm 0.8}$	$11.2_{\pm 1.0}$	$8.9_{\pm 0.3}$	$49.3_{\pm 1.9}$	$44.1_{\pm 3.0}$
Mistral	$3\times$	0	$1.0_{\pm 0.2}$	$1.5_{\pm 0.3}$	0	$65.0_{\pm 18.2}$	0	$1.0_{\pm 0.2}$	$1.2_{\pm 0.2}$
	$2\times 1\surd$	33.3	$38.8_{\pm 0.8}$	$41.7_{\pm 1.7}$	$49.8_{\pm 1.3}$	$36.3_{\pm 2.4}$	$55.2_{\pm 2.0}$	$33.2_{\pm 1.4}$	$27.7_{\pm 1.6}$
	$1\times 2\surd$	66.7	$81.6_{\pm 0.6}$	$83.6_{\pm 1.0}$	$12.7_{\pm 0.9}$	$11.3_{\pm 1.0}$	$15.9_{\pm 1.4}$	$70.1_{\pm 1.6}$	$60.9_{\pm 2.5}$
	$3\surd$	100.0	$96.0_{\pm 0.2}$	$93.6_{\pm 0.5}$	$7.4_{\pm 0.7}$	$4.2_{\pm 0.4}$	$6.4_{\pm 0.5}$	0	$66.0_{\pm 3.5}$
	Standard	$63.2_{\pm 0.6}$	$67.6_{\pm 0.4}$	$68.0_{\pm 0.3}$	$12.0_{\pm 0.4}$	$9.9_{\pm 0.9}$	$11.9_{\pm 0.9}$	$32.6_{\pm 0.9}$	$21.8_{\pm 1.2}$

Table 6: The hallucination propagation results of NQ. Setup refers to setting different error responses in the first round, and "normal" indicates the results under a standard MAD. The setup $3\times$ and $3\surd$ can be seen as the lower and upper bounds, respectively.

Model	Setup	MA_1	MA_2	MA_3	$MR_2 \vert IMR_2$	MR_3	IMR_3	CR_2	CR_3
Llama	$3\times$	0	$15.3_{\pm 0.2}$	$30.9_{\pm 0.4}$	0	$35.9_{\pm 1.5}$	0	$15.3_{\pm 0.2}$	$24.9_{\pm 0.4}$
	$2 \times 1\surd$	33.3	$58.5_{\pm 0.5}$	$55.4_{\pm 1.4}$	$92.2_{\pm 0.2}$	$52.8_{\pm 1.7}$	$19.2_{\pm 1.6}$	$83.8_{\pm 0.9}$	$67.0_{\pm 3.3}$
	$1 \times 2\surd$	66.7	$56.1_{\pm 0.4}$	$49.6_{\pm 1.5}$	$63.1_{\pm 0.4}$	$44.7_{\pm 2.9}$	$48.6_{\pm 1.1}$	$94.5_{\pm 1.1}$	$42.3_{\pm 3.0}$
	$3\surd$	100.0	$84.3_{\pm 0.4}$	$76.2_{\pm 0.7}$	$15.7_{\pm 0.4}$	$17.9_{\pm 1.4}$	$23.8_{\pm 0.7}$	0	$44.9_{\pm 4.0}$
	Standard	$68.1_{\pm 1.0}$	$70.1_{\pm 0.8}$	$68.9_{\pm 0.8}$	$23.3_{\pm 1.4}$	$21.7_{\pm 0.4}$	$18.9_{\pm 0.5}$	$56.0_{\pm 1.1}$	$46.8_{\pm 3.1}$
Mistral	$3\times$	0	$5.5_{\pm 0.6}$	$9.1_{\pm 0.9}$	0	$43.3_{\pm 3.9}$	0	$5.5_{\pm 0.6}$	$6.3_{\pm 0.7}$
	$2 \times 1\surd$	33.3	$55.6_{\pm 0.6}$	$56.9_{\pm 1.4}$	$35.3_{\pm 0.8}$	$24.9_{\pm 0.8}$	$41.1_{\pm 2.6}$	$51.1_{\pm 0.5}$	$34.0_{\pm 2.3}$
	$1 \times 2\surd$	66.7	$85.4_{\pm 1.1}$	$86.5_{\pm 1.4}$	$8.5_{\pm 1.0}$	$8.3_{\pm 0.8}$	$12.8_{\pm 1.6}$	$73.3_{\pm 1.4}$	$55.7_{\pm 4.0}$
	$3\surd$	100.0	$98.4_{\pm 0.4}$	$96.7_{\pm 0.2}$	$2.9_{\pm 0.3}$	$2.2_{\pm 0.2}$	$3.4_{\pm 0.3}$	$0.0_{\pm 0.0}$	$54.9_{\pm 4.4}$
	Standard	$68.5_{\pm 1.0}$	$70.3_{\pm 0.7}$	$70.6_{\pm 0.8}$	$5.4_{\pm 0.6}$	$4.2_{\pm 0.3}$	$4.8_{\pm 0.7}$	$17.4_{\pm 1.2}$	$10.9_{\pm 0.9}$

Table 7: The hallucination propagation results of BQ. Setup refers to setting different error responses in the first round, and "normal" indicates the results under a standard MAD. The setup $3\times$ and $3\surd$ can be seen as the lower and upper bounds, respectively.

Model	Setup	MA_1	MA_2	MA_3	$MR_2 \vert IMR_2$	MR_3	IMR_3	CR_2	CR_3
Llama	$3\times$	0	$7.4_{\pm 0.5}$	$13.8_{\pm 1.0}$	0	$48.7_{\pm 4.3}$	0	$7.4_{\pm 0.5}$	$10.8_{\pm 1.1}$
	$2 \times 1\surd$	33.3	$60.1_{\pm 0.8}$	$51.4_{\pm 0.8}$	$87.4_{\pm 1.4}$	$59.1_{\pm 1.9}$	$16.1_{\pm 0.5}$	$83.9_{\pm 0.6}$	$67.3_{\pm 0.5}$
	$1 \times 2\surd$	66.7	$65.4_{\pm 1.0}$	$57.3_{\pm 1.6}$	$49.3_{\pm 1.4}$	$36.0_{\pm 0.3}$	$41.3_{\pm 1.9}$	$94.6_{\pm 0.5}$	$44.8_{\pm 3.8}$
	$3\surd$	100.0	$91.2_{\pm 0.2}$	$90.9_{\pm 0.5}$	$8.8_{\pm 0.2}$	$7.5_{\pm 0.6}$	$9.1_{\pm 0.5}$	0	$73.6_{\pm 1.6}$
	Standard	$56.7_{\pm 1.0}$	$58.7_{\pm 1.1}$	$61.2_{\pm 1.5}$	$21.7_{\pm 1.2}$	$16.3_{\pm 0.9}$	$12.5_{\pm 1.0}$	$33.1_{\pm 0.9}$	$29.4_{\pm 1.5}$
Mistral	$3\times$	0	$2.7_{\pm 0.3}$	$4.0_{\pm 0.2}$	0	$58.0_{\pm 5.4}$	0	$2.7_{\pm 0.3}$	$3.0_{\pm 0.2}$
	$2 \times 1\surd$	33.3	$48.3_{\pm 1.3}$	$48.6_{\pm 0.7}$	$48.0_{\pm 2.1}$	$35.0_{\pm 0.6}$	$45.8_{\pm 1.4}$	$46.5_{\pm 1.8}$	$33.3_{\pm 1.1}$
	$1 \times 2\surd$	66.7	$83.8_{\pm 0.9}$	$85.9_{\pm 0.9}$	$13.9_{\pm 0.6}$	$9.9_{\pm 0.6}$	$14.2_{\pm 0.8}$	$79.1_{\pm 2.4}$	$64.3_{\pm 2.2}$
	$3\surd$	100.0	$94.6_{\pm 1.0}$	$95.9_{\pm 0.5}$	$5.4_{\pm 1.0}$	$2.6_{\pm 0.3}$	$4.1_{\pm 0.5}$	0	$67.9_{\pm 2.3}$
	Standard	$53.0_{\pm 0.5}$	$59.1_{\pm 0.5}$	$61.1_{\pm 0.5}$	$10.5_{\pm 0.8}$	$8.4_{\pm 1.0}$	$9.4_{\pm 0.6}$	$24.7_{\pm 1.1}$	$17.0_{\pm 0.8}$

Table 8: The hallucination propagation results of TruthfulQA. Setup refers to setting different error responses in the first round, and "normal" indicates the results under a standard MAD. The setup $3\times$ and $3\surd$ can be seen as the lower and upper bounds, respectively.

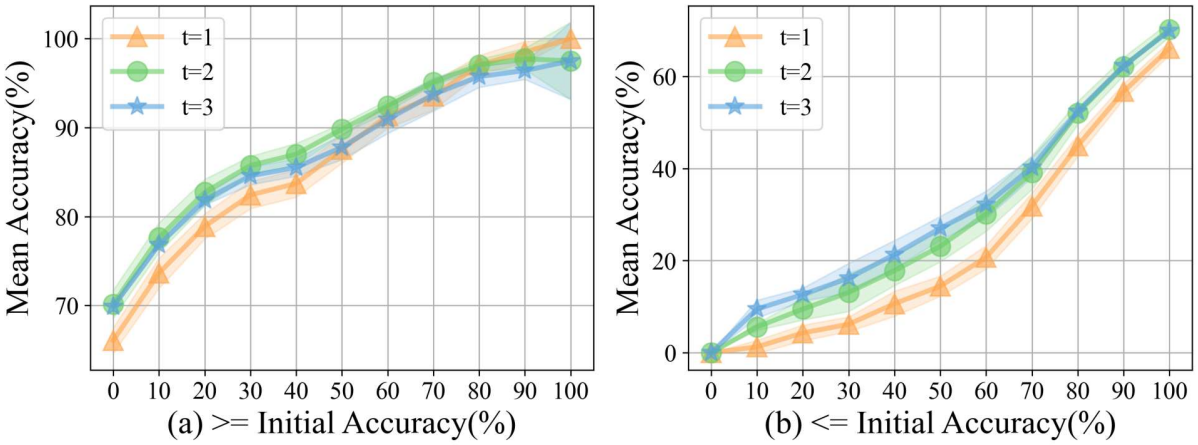

Figure 5: Test results categorized by question difficulty with 3 Mistral agents on the NQ dataset. (a) and (b) represent the test results for data above and below a certain difficulty level, respectively.

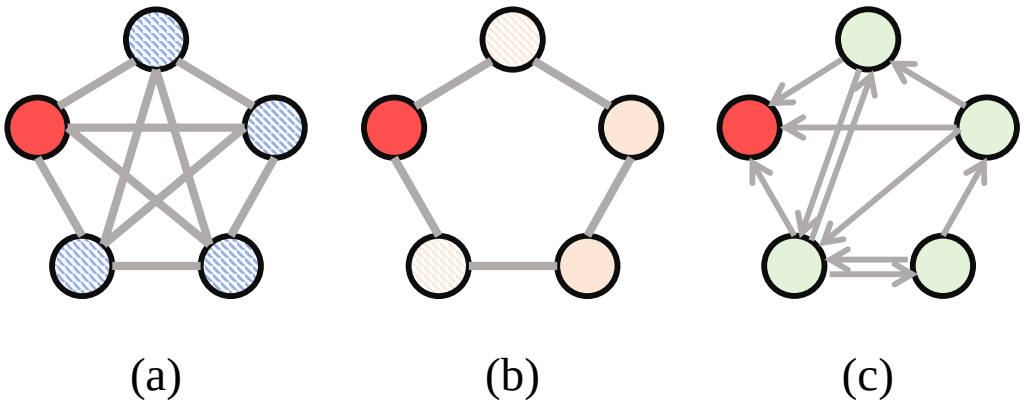

Figure 6: Different Communication Topology. (a) Topology of Multi-Agent Debate, (b) Topology of Sparse MAD ($D = \frac{1}{2}$), (c) Topology of DIGRA. The red nodes represent agents exhibiting hallucinations, while the dashed nodes indicate agents at risk of being misled by the propagation of hallucinations.

higher stability and is able to achieve performance improvements over a broader range through Debate (Figure 5).

B DIGRA: Mitigating the hallucination propagation in Multi-Agent Debate

B.1 Communication Topology in MAD

We present different communication topologies in Figure 6. From the figure, we observe that when a single agent exhibits hallucinations, the risk of hallucination propagation is highest with the predefined static topology. Sparse communication reduces the hallucination propagation to some extent, but it cannot fully resolve the issue. As shown in Figure 6 (c), by leveraging dynamic topologies to select the most advantageous communication partners, the propagation of hallucinations can be mitigated.

B.2 Parameter analysis

In the formula of IGR, we introduce the hyperparameter α to balance the entropy of the reference agents and the information gain. In this section, we analyze the impact of different values for this parameter. As shown in Table 9, the performance exhibits a trend of first increasing and then decreasing as the α increases. When α is too small, the importance of entropy is overlooked, leading to the selection of agents with high hallucination levels for communication. When α is too large, the information gain is ignored, and the selected agents may lack significant reference value for the current agent. When α is set to 0.5, DIGRA achieved significant improvement, suggesting that an optimal balance between information gain and entropy of agents yields enhanced performance. In our experiment, we pre-set this value without further tuning α , indicating that DIGRA holds greater potential for achieving even better performance.

B.3 The Details of DIGRA

B.3.1 Calculation of information gain ratio

In this section, we will explain how information gain ratio is computed using the specific prompt template. As shown in Table 10, we first concatenate the responses of agents in $\mathcal{J}$ with the prompt of the original question into the predefined template. Then, we set the output of the current agent and perform forced decoding to compute the entropy.

B.3.2 Early stoping in DIGRA

Since hallucinations exhibit diffusion characteristics, the early stopping mechanism we designed helps mitigate this issue. Specifically, our early stopping mechanism is based on the following principles:

(i) All agents reach a consensus and provide an

α	0.01	0.05	0.1	0.2	0.3	0.5	1.0
Accuracy	$68.5_{\pm 4.5}$	$70.0_{\pm 4.1}$	$69.8_{\pm 4.7}$	$71.8_{\pm 3.0}$	$70.8_{\pm 3.5}$	$\mathbf{74.0}_{\pm \mathbf{4.2}}$	$70.8_{\pm 3.6}$

Table 9: Accuracy (%) of DIGRA with different α on MMLU benchmark.

Response	$R^q_{1,t}$ $R^q_{2,t}$ $R^q_{3,t}$
entropy order	$R^q_{3,t} > R^q_{2,t} > R^q_{1,t}$
current agent	agent 1
potential agents $\mathcal{J}$	$\{R^q_{2,t}\}$ $\{R^q_{3,t}\}$ $\{R^q_{3,t}, R^q_{2,t}\}$
Prompt	{Original prompt of q}
$f(q, R^q_{\mathcal{J},t})\|_{\mathcal{J}=\{3,2\}}$	These are the solutions to the problem from other agents: One agent solution: "' $R^q_{3,t}$ "' One agent solution: "' $R^q_{2,t}$ "' Using the reasoning from other agents as additional advice, can you give an answer? The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering.
$IG(R^q_{1,t}\|R^q_{\mathcal{J},t})$	$\mathbb{H}(R^q_{1,t}) - \mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{2\},t})$ $\mathbb{H}(R^q_{1,t}) - \mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{3\},t})$ $\mathbb{H}(R^q_{1,t}) - \mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{3,2\},t})$
$IGR(R^q_{1,t}\|R^q_{\mathcal{J},t})$	$\frac{\alpha+\mathbb{H}(R^q_{1,t})-\mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{2\},t})}{\mathbb{H}(R^q_{2,t})} = 0.69$ $\frac{\alpha+\mathbb{H}(R^q_{1,t})-\mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{3\},t})}{\mathbb{H}(R^q_{3,t})} = 1.37$ $\frac{\alpha+\mathbb{H}(R^q_{1,t})-\mathbb{H}(IG(R^q_{1,t}\|R^q_{\mathcal{J}=\{2,3\},t})}{\frac{1}{2}(\mathbb{H}(R^q_{2,t})+\mathbb{H}(R^q_{3,t}))} = 0.91$
final communication agents	agent 3

Table 10: Examples of DIGRA and details of the prompt template function.

Model	Methods	NQ	BQ	TruthfulQA	GSM8K	MMLU	Avg.
Mistral	CoT	$62.4_{\pm 0.7}$	$64.0_{\pm 1.7}$	$59.8_{\pm 1.0}$	$38.5_{\pm 3.5}$	$54.8_{\pm 3.0}$	55.9
	CoT-SC	$67.9_{\pm 0.7}$	$64.7_{\pm 0.7}$	$60.1_{\pm 0.6}$	$42.8_{\pm 4.2}$	$56.2_{\pm 0.8}$	58.3
	MAD($D=1$)	$69.9_{\pm 0.3}$	$67.9_{\pm 0.7}$	$61.5_{\pm 0.7}$	$\underline{45.2_{\pm 1.3}}$	$\underline{56.4_{\pm 1.3}}$	60.2
	MAD($D=\frac{1}{2}$)	$65.8_{\pm 2.1}$	$67.7_{\pm 0.5}$	$\mathbf{61.9}_{\pm \mathbf{0.4}}$	$38.8_{\pm 3.7}$	$55.0_{\pm 1.2}$	57.8
	MAD(random)	$69.1_{\pm 0.6}$	$67.9_{\pm 0.6}$	$61.2_{\pm 0.5}$	$41.5_{\pm 2.7}$	$53.8_{\pm 1.8}$	58.7
	DIG	$\underline{70.9_{\pm 0.6}}$	$\underline{68.6_{\pm 0.8}}$	$61.4_{\pm 0.1}$	$44.5_{\pm 3.8}$	$56.2_{\pm 0.8}$	$\underline{60.3}$
	DIGRA	$\mathbf{72.2}_{\pm \mathbf{1.1}}$	$\mathbf{68.7}_{\pm \mathbf{0.6}}$	$\underline{61.6_{\pm 0.3}}$	$\mathbf{47.0}_{\pm \mathbf{2.4}}$	$\mathbf{57.0}_{\pm \mathbf{1.6}}$	**61.3**

Table 11: Comparison of accuracy of DIGRA with Mistral against baseline methods. The optimal performance is highlighted in bold, and the second-best performance is underlined.

answer (i.e., the answer is not None).

(ii) One agent's opinion is consistent for two consecutive rounds and the answer is not None.

(iii) For terminated agents, we assume tha t$R^q_{i,t+1} = R^q_{i,t}$.

B.4 Results of Mistral

Table 11 shows the results of Mistral. From prior experiments, we found that although Mistral is less capable than Llama 3.1, it exhibits better debating characteristics. Similarly, Mistral consistently outperforms CoT-SC in MAD, indicating that the model demonstrates strong resistance to hallucination propagation, thus showing effective collective intelligence. Moreover, we discovered that the introduction of DIGRA further boosts its debating ability, leading to consistent improvements across multiple datasets.