IMPD-MACD: A Comprehensive Multi-Perspective Cognitive Fusion Approach for LLM Hallucination Detection

Anonymous ACL submission

Abstract

In recent years, researchers have observed that LLMs frequently generate false content ("hallucinations") with highly confident tones when answering questions, a phenomenon that severely undermines model reliability. Therefore, timely and accurate detection of LLM-generated hallucinations is crucial. However, existing methods face multiple challenges in hallucination detection: (1) Single-agent methods lack comprehensive identification of different types of hallucinations, and some methods are difficult to apply to black-box models; (2) Multi-agent methods lack clear division of labor and deep interaction, and combined with inherent biases and overconfidence issues, their detection effectiveness is insufficient in complex scenarios. In response, we propose IMPD-MACD, which enhances agent division of labor and collaboration through multiple perspectives and stances, not only significantly improving detection accuracy but also more comprehensively covering different types of hallucinations, thereby enhancing LLM reliability and practicality in diverse scenarios. Extensive experiments demonstrate that our method significantly outperforms the current SOTA (state-of-the-art) approaches across multiple metrics. The project and its associated dataset will be publicly released.[1]

1 Introduction

In recent years, large language models (LLMs), such as the GPT series (Brown et al., 2020) and LLaMA (Touvron et al., 2023), have achieved significant breakthroughs in the field of Natural Language Processing (NLP). However, due to the constraints arising from the sources and quality of the training data, LLMs often produce incorrect answers with high confidence, a phenomenon generally referred to as "hallucination" (Ji et al., 2023). Recently, researchers have proposed various approaches to identifying these hallucinations, including Chain-of-Thought (CoT) (Wei et al., 2022), internal activation-based detection (Azaria and Mitchell, 2023), Self-Reflection (Shinn et al., 2024), Self-CheckGPT (Manakul et al., 2023), InterrogateLLM (Yehuda et al., 2024), and multi-agent debate (Du et al., 2023). Although such efforts represent noteworthy progress, key challenges remain: (1) The traditional single-agent method, such as internal activation-based methods are not applicable to black-box models like GPT, and Self-Correction or Diversified-Sampling method (Fang et al., 2024; Valmeekam et al., 2023; Huang et al., 2023; Stechly et al., 2023) often encounter excessive confidence or repetitive biases. (2) In addition, existing single-agent approaches exhibit different strengths and weaknesses when detecting hallucinations of logic, factual, or context inconsistency.

Although multi-agent methods have partially mitigated the aforementioned issues, they have not fully resolved these challenges and furthermore face the following additional limitations: **(1) They lack differentiated role assignments and rely on agents that operate in relative isolation, rendering cross-agent collaboration minimal and constraining the full realization of multi-perspective advantages (Fang et al., 2024). (2) This limitation hinders comprehensive detection of those three categories of hallucinations. (3) Moreover, most such methods remain focused on closed-ended question answering, leaving their effectiveness in more complex semi-open question answering (semi-open Q&A) scenarios (Appendix A) uncertain.**

Inspired by real-world debates, we propose a novel model, IMPD-MACD (Integrated Multi-Perspective and Diverse Stance Multi-Agent Collaborative Debate Model). This model achieves comprehensive hallucination detection in complex

[1] https://anonymous.4open.science/r/Semi-Open-HaluQA-and-IMPD-MACD-4883 (Please note that when you directly copy this address, a space may be mistakenly added between "-" and "HaluQA". Remove the space, and the address should be accessible.)

scenarios through two key mechanisms: first, by assigning distinct perspectives and roles to individual agents, and second, by leveraging multi-agent synergistic capabilities through multiple rounds of interaction. Our experimental results on two semi-open Q&A datasets validate the effectiveness of this approach. To address the inherent biases and overconfidence commonly observed in single-agent approaches, we implement a dual enhancement strategy: assigning diverse stances to individual agents while incorporating a dedicated Information-Gathering Agent. This integration of multiple perspectives and external knowledge significantly enhances the model's objectivity and accuracy, particularly in complex hallucination detection scenarios.

Building upon the HaluEval dataset (Li et al., 2023), we manually annotated and released a new dataset, Semi-Open-HaluQA, and conducted systematic evaluations of IMPD-MACD on both datasets. Experimental findings reveal that, compared with CoT, Self-Reflection,MAD and MADR our proposed method achieves 15.5%, 24%, 14% and 17% improvements in Accuracy, respectively, along with corresponding increases of 0.0789, 0.1292, 0.123 and 0.115 in F1_Score, while also demonstrating robust performance across multiple other metrics. Furthermore, due to its multi-perspective debate mechanism, our method attains faster inference speeds relative to the baselines. Overall, the principal contributions of this paper include:

(1) This study proposes Semi-Open-HaluQA, a semi-open Q&A dataset specifically developed for evaluating and detecting hallucinations. (2) We propose the IMPD-MACD framework, designed to more effectively identify and address erroneous or misleading information generated by LLMs. (3) Furthermore, we present a multi-perspective interactive debate method integrated with Information-Gathering Agent. (4) Experimental evaluations conducted on the Semi-Open-HaluQA and HaluEval datasets demonstrate that our framework significantly enhances the accuracy of hallucination detection compared to existing methods.

2 Related Work

LLMs frequently exhibit factual inaccuracies or fabricated content when generating natural language text (Brown et al., 2020; Ji et al., 2023). To address this issue, researchers and industry practitioners have proposed numerous detection and mitigation strategies.

2.1 Traditional Single-Agent Method

Existing single-agent hallucination detection methods can be framed under three layers—data, model, and application (Liu et al., 2024b). Data-layer approaches, such as cleaning training data beforehand, are resource-intensive and inapplicable to pretrained models. At the model layer, Azaria and Mitchell (2023) suggest extracting LLM's specific activation values to train a hallucinatory-output classifier, but this approach requires access to internal states—unsuitable for black-box models like GPT. The application layer often leverages the LLM itself (Fang et al., 2024); self-correction techniques, including Chain-of-Thought (Wei et al., 2022) and Self-Reflection (Shinn et al., 2024), promote transparent reasoning yet offer limited sensitivity to factual or contextual inaccuracies. Diversified sampling methods, such as Self-CheckGPT (Manakul et al., 2023) and InterrogateLLM (Yehuda et al., 2024), respectively compare multiple candidate answers or examine reconstructed queries for reliability but may fail to capture all types hallucinations in semi-open Q&A tasks. To address multi-category detection, Hu et al. (2024) adopt a diversified-sampling-based approach that still relies on human intervention or knowledge validation, impeding full automation.

2.2 Multi-Agent Method

Irving et al. (2018) proposed "AI Safety via Debate", positing that multiple agents debating reasoning flaws can guide humans toward valid arguments. Building on this, Du et al. (2023) employed multi-agent debates for hallucination detection in LLMs, where agents solve questions independently before converging on a unified conclusion. Kim et al. (2024) proposed MADR method with two-agent cyclical feedback refinement. Sun et al. (2024) proposed a Markov chain-based debate framework for the same purpose. However, these debate-based methods lack fine-grained role assignments and remain untested on semi-open Q&A datasets reflecting real-world scenarios. To address these gaps, we designate distinct roles and stances among agents, encourage multi-perspective interaction to boost objectivity, and incorporate external knowledge to strengthen hallucination detection.

3 Method

This chapter provides a detailed introduction to the proposed IMPD-MACD approach. Before commencing multi-agent collaborative debates, external knowledge retrieval is incorporated to enhance the model's internal reasoning capabilities. During the debate, potential hallucinations are systematically identified and corrected through multi-role settings, contrasting stances, and a judging mechanism. Figure 1 illustrates the overall process framework of this method.

3.1 Information-Gathering Agent

To mitigate the risk of hallucination in multi-agent reasoning and debate, this study employs a RAG (Retrieval-Augmented Generation) approach for external knowledge retrieval prior to multi-agent interactions, while classifying agents into different stance groups so that each group retrieves information specifically relevant to its respective viewpoint. Specifically, an Information-Gathering Agent is designed to integrate externally retrieved objective information with the agents' internal representations (e.g., parameterized knowledge or short-term working memory) (Li et al., 2024). This integration enables the evaluation of generated outputs for potential biases or errors within a more comprehensive knowledge space. Implementation details are provided in Appendix B.

3.2 Multi-Agent Debate Framework

In this subsection, we undertake an in-depth exploration of the IMPD-MACD framework's overall design and operational mechanisms.

3.2.1 Multi-Perspective Agent Roles and Agent Grouping

Once the Information-Gathering Agent has completed gathering external knowledge, IMPD-MACD formally initiates the multi-agent collaborative debate. This debate process encompasses three core stages—role allocation, inter-group debate, and judge evaluation—whose detailed functionalities can be found in Appendix C.1.

First, the system divides all agents into two teams: the Pro Team, which asserts that "no hallucination exists in the original answer", and the Con Team, which contends that "the original answer may contain hallucinations". Within each team, four distinct agent roles are further designated to establish a multi-perspective, multi-role debating environment:

- **Logic Agent:** Focuses on identifying logical inconsistencies in the answer (i.e., logical hallucinations).
- **Context Consistency Agent:** Aims to detect mismatches between the answer and its context (i.e., context inconsistency hallucinations).
- **Factual Agent:** Primarily utilizes its internal parameter knowledge to evaluate factual consistency and additionally leverages retrieved external information to verify factual accuracy (i.e., factual hallucinations).
- **Comprehensive Agent:** Comprehensively consolidates and integrates the judgments and opinions from the other three agents within the same team, generates summaries of each debate round and an overall stance, **and** provides independent evaluations regarding the presence of hallucinations in both the question and the answer.

3.2.2 Multi-Round Interactions Among Agents

In multi-round debates, the nature of interactions between agents is fundamental to the final debate quality. To formally describe this process, we first define the set of participating agents in the debate system:

$$\mathcal{A} = \{A_1, A_2, \ldots, A_n\} \tag{1}$$

These agents are systematically divided into two opposing teams, with no overlap between them:

$$\mathcal{A}_{\text{team1}} \cup \mathcal{A}_{\text{team2}} = \mathcal{A}, \quad \mathcal{A}_{\text{team1}} \cap \mathcal{A}_{\text{team2}} = \varnothing \tag{2}$$

The quality of interaction at each round t is denoted as Q_t, and the final debate quality Q_{final} is expressed as:

$$Q_{\text{final}} = \frac{1}{T} \sum_{t=1}^{T} Q_t \tag{3}$$

where T denotes the total number of rounds. Due to the inherent context window limitations of LLMs (MaxLen), the context $\mathcal{C}_t(A_i)$ received by each agent A_i at round t must satisfy the following constraint:

$$\|\mathcal{C}_t(A_i)\| \leq \mathsf{MaxLen} \tag{4}$$

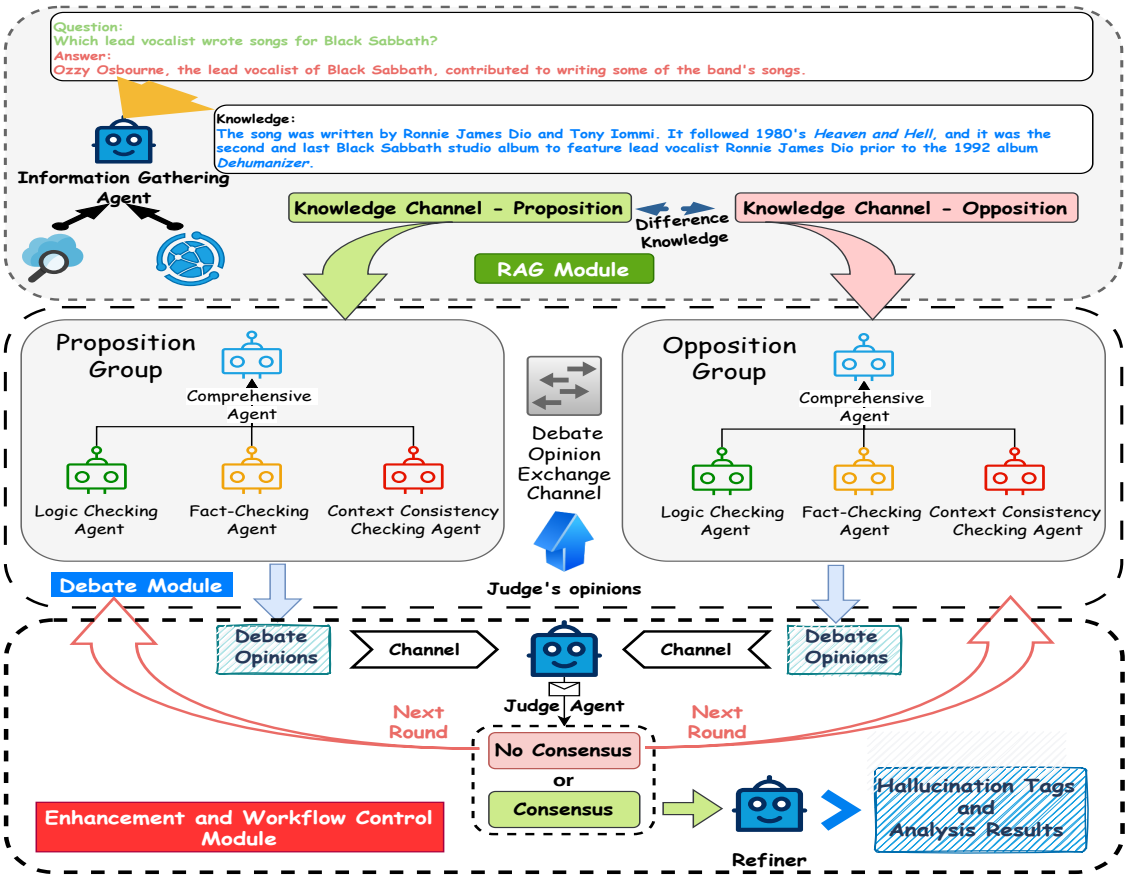

Figure 1: The proposed IMPD-MACD framework

To address this constraint effectively, a Comprehensive Agent implements a history summarization task ($Summarize$ denotes the mapping from specific inputs to corresponding outputs for the Comprehensive Agent):

$$Summarize(H, L) \rightarrow \tilde{H} \tag{5}$$

Where L represents the constraint length of historical information. The compression ratio γ of this summarization is expressed as:

$$\gamma = \frac{|\tilde{H}|}{|H|} \tag{6}$$

At each round t, the summarized history is computed by combining current team content with previous opponent information:

$$\tilde{H}_t = Summarize(H_t^{\text{team1}} \cup H_{t-1}^{\text{team2}}, L) \tag{7}$$

Concurrently, a Judge Agent executes a specialized feedback function ($JudgeSummarize$ denotes the mapping from specific inputs to corresponding outputs for the Judge Agent):

$$JudgeSummarize(H_t, \alpha) \rightarrow F_t \tag{8}$$

where α controls the feedback length $|F_t|$. To mitigate hallucinations in Agents, strict access restrictions are imposed on historical information for each agent A_i:

$$\mathcal{H}_t(A_i) = \tilde{H}_t(A_i) \cup \tilde{H}_{t-1}(\text{Opp}(A_i)) \tag{9}$$

Here, $\text{Opp}(A_i)$ represents the historical summary of the opposing agent or team. Empirically, the probability of hallucination in a single-turn interaction is directly correlated with the amount of irrelevant information I_{irr}:

$$P(\text{hallucination}) = g(I_{\text{irr}}) \tag{10}$$

Therefore, it is essential to implement this stringent access restriction mechanism to effectively minimize the amount of irrelevant information, thereby reducing the likelihood of hallucinations in Agents during debates and ensuring both the quality and reliability of the debate process.

3.2.3 Multi-Round Debate Process

During the formal debate phase, we can mathematically formalize the multi-round interaction process between Pro and Con teams. Let's define the debate state at round t as:

$$S_t = \{Q, A_0, \mathcal{K}_{\text{ext}}, \mathcal{K}_{\text{param}}, H_{t-1}\} \quad (11)$$

where Q represents the input question, A_0 is the initial answer, $\mathcal{K}_{\text{ext}}$ and $\mathcal{K}_{\text{param}}$ represent external and parameter knowledge respectively, and H_{t-1} captures previous debate history.

Each agent's perspective at round t can be formalized through a perspective function [2]:

$$V_t^{\text{Agent}} = f_{\text{perspective}}(S_t, \mathcal{K}_{\text{team,Agent}}) \quad (12)$$

The interaction quality Q_t between teams is measured by aggregating individual agent interactions:

$$Q_t = \sum_{i=1}^{4} w_i \cdot q_t^i(A_i, \text{Opp}(A_i)) \quad (13)$$

where w_i represents agent weights and q_t^i measures interaction quality between agents (Appendix C.2).

The teams' conclusions are consolidated through a consolidation function ($Consolidate$ denotes the mapping from specific inputs to corresponding outputs for the Comprehensive Agent):

$$C_t^{\text{team}} = Consolidate(\{V_t^i\}_{i=1}^4, \tilde{H}_t) \quad (14)$$

The debate process continues until either reaching maximum rounds T_{max} or achieving convergence defined [2] by (See Appendix C.3 for details):

$$J_t = f_{\text{judge}}(C_t^{\text{team1}}, C_t^{\text{team2}}, Q_t) \leq \epsilon \quad (15)$$

The probability of hallucination decreases exponentially over debate rounds according to:

$$P(\text{hallucination})_t = g(I_{irr}(t)) \cdot (e^{-\lambda t} + \phi \mathcal{H}(t)) \quad (16)$$

where λ represents the learning rate (an intrinsic model characteristic quantifying knowledge acquisition capability rather than a training hyperparameter) and $g(I_{\text{irr}}(t))$ captures the impact of irrelevant information as defined in previous Equation (10). (See Appendix C.4 for more details)

Within this rigorous framework, the iterative debate process facilitates a quantifiable reduction in model uncertainties, as demonstrated by Equation (16). The adversarial-cooperative dynamics support the ongoing refinement of responses by continuously integrating new opinions and knowledge, thereby maintaining tractability in the detection of hallucinations. Consequently, when Multi-Agents are engaged in debate, this approach progressively diminishes errors arising from inherent biases and overconfidence.

3.2.4 Judge Agent and Refiner Agent

To ensure fairness and efficiency across multiple rounds of interaction, the system incorporates a Judge Agent and a Refiner Agent. After round t, the Judge Agent evaluates arguments via Equation (15), where J_t represents the judgment score based on logical consistency, factual accuracy, coherence, cross-reference validity (Q_t), and comprehensive assessment.

Debate continues to round $t+1$ if consensus isn't reached:

$$J_t > \epsilon \quad (17)$$

At conclusion (round T), the Refiner Agent generates report [2] R:

$$R = f_{\text{refine}}(\{C_t^{\text{team1}}, C_t^{\text{team2}}, J_t, Q_t\}_{t=1}^T) \quad (18)$$

Hallucination assessment quantifies:

$$H_{\text{assess}}(R) = \{(h_i, p_i, E_i) \mid i \in \text{Types}\} \quad (19)$$

where h_i represents hallucination type i, p_i its probability from J_t, and E_i the supporting evidence, ensuring traceable detection results.

4 Experiment

To evaluate the effectiveness of the IMPD-MACD model in detecting hallucinations in semi-open questions, we then conducted multiple experiments on Semi-Open-HaluQA and the HaluEval dataset. During evaluation, we compare the model's output predictions (0 indicating the presence of hallucinations in the answer, 1 indicating the absence of hallucinations) with the ground truth labels in the dataset, based on which we calculate Accuracy, Precision, Recall, and F1 scores to comprehensively measure the model's detection capabilities.

[2] $f_{\text{perspective}}$, f_{judge} and f_{refiner} denote the mapping relationships between inputs and outputs for their respective agents.

Method	Metrics (Semi-Open-HaluQA)				Metrics (HaluEval)			
	Accuracy	Precision	Recall	F1_Score	Accuracy	Precision	Recall	F1_Score
Single-Agent Method								
CoT	0.665	0.605	**0.950**	0.739	0.660	0.638	0.740	0.685
Self-Reflection	0.580	0.547	0.930	0.689	0.635	0.617	0.710	0.666
Multi-Agent Method								
MAD	0.680	0.664	0.730	0.695	0.630	0.667	0.520	0.584
MADR	0.650	0.610	0.830	0.703	0.580	0.563	0.720	0.632
IMPD-MACD (Ours)	**0.820**	**0.827**	0.810	**0.818**	**0.795**	**0.798**	**0.790**	**0.794**

Table 1: Performance comparison on Semi-Open-HaluQA and HaluEval datasets. The best values are highlighted in green and blue.

4.1 Dataset

In subsequent experiments, we select 600 high-quality samples from Semi-Open-HaluQA's 1,800 Q&A entries. Using random seed 42, we then draw 200 samples each from HaluEval and Semi-Open-HaluQA for model evaluation. As for the HaluEval dataset, its basic structure is similar to that of Semi-Open-HaluQA, but its answers are explicitly categorized as correct or hallucinatory. Therefore, using random seed 2025, we randomly choose either the correct answer or the hallucinatory answer for the corresponding question and provide relevant annotations. This dataset will be open-sourced along with the model; Additional details can be found in Appendix D.

4.2 Baselines

In this study, we select CoT (Wei et al., 2022), Self-Reflection (Shinn et al., 2024), MAD (Du et al., 2023), and MADR (Kim et al., 2024) as the comparison methods. To ensure fairness, the agents used in these methods, as well as in our proposed multi-perspective agent approach, uniformly adopt the same gpt-4o-mini model. However, due to substantial differences in workflow and task responsibilities, strict consistency in the number of agents or debate rounds cannot be maintained. Consequently, we use the original MAD paper's default parameters when testing MAD. Since MAD is not directly applicable to the Semi-Open-HaluQA usage scenario, we make modifications to fit this study's requirements, detailed in Appendix E.1.

4.3 Performance Comparison

Table 1 presents the performance of all the compared methods on the Semi-Open-HaluQA and HaluEval datasets. Among these, the best results achieved by each method in the HaluEval dataset for a given metric are marked in blue bold, and the best results in the Semi-Open-HaluQA dataset for a given metric are marked in green bold.

In the HaluEval dataset, our method exhibits a significant advantage over other compared approaches across all metrics. Notably, on this two datasets the most critical Accuracy metric improves by at least 13.5%, and it also surpasses the second-best method on the Semi-Open-HaluQA dataset by 14%. This result indicates that a multi-perspective, multi-agent debate framework enables the model to more accurately distinguish hallucinatory answers from correct ones.

In LLM hallucination detection, Single-Agent approaches leverage their streamlined architectures and focused reasoning to effectively identify potential hallucinations. However, in application scenarios where a low false-positive rate is paramount, Multi-Agent approaches show superior detection accuracy and more robust misclassification control.

Further analysis reveals that CoT, Self-Reflection and MADR methods achieve relatively high Recall on the Semi-Open-HaluQA dataset but have relatively low Accuracy; however, their performance on the HaluEval dataset is more balanced. The possible reason is that the questions and answers in Semi-Open-HaluQA are more complex, making it difficult for models to fully parse the semantics and identify nuanced errors. As a result, they tend to predict most answers as having "no hallucination," leading to high Recall but lower Accuracy and Precision. This issue is less pronounced in the simpler HaluEval dataset. Therefore, in more complex application scenarios, increasing the model's complexity for single agents and enabling efficient interactions among multiple agents both serve as vital approaches to enhancing detection performance.

From a stability perspective, due to the "razor effect", complex models add "additional entities", which do not necessarily improve task-switching adaptability and can hinder it. Complex models are more sensitive to task variations, while simpler models tend to demonstrate consistent performance across diverse tasks. Consequently, the simplest CoT model shows the smallest performance gap between the two datasets. Our model ranks second in stability, indicating strong generalization without compromising robustness akin to simpler models.

4.4 Ablation Experiment

To clarify the specific contribution of each module in our method, Table 2 presents the model's performance on both datasets after the removal of various modules. Specifically, "w/o EK" indicates the removal of the Information Gathering Agent, "w/o Logic" indicates the removal of the Logic Agent, "w/o Fact" indicates the removal of the Factual Agent, "w/o Comprehensive" indicates the removal of the Comprehensive Agent, and "w/o Consistency" indicates the removal of the Context Consistency Agent. (More analysis in Appendix E.2)

IMPD-MACD-w/o EK: After removing the Information-Gathering Agent, the model shows a marked decrease in both the Accuracy and Recall metrics. This indicates that in the complete method, the various agents are already capable of thoroughly understanding and utilizing the retrieved external knowledge, thereby effectively enhancing hallucination detection performance. Consequently, introducing correct and relevant external knowledge is crucial for accurately identifying hallucinations.

IMPD-MACD-w/o Logic: When the Logic Agent is removed, the Accuracy metric drops significantly. Given that the answers in this experiment's dataset are generated by GPT-4, this phenomenon implies that among the types of hallucinations produced by a LLM, logical hallucinations are quite prominent; the absence of a Logic Agent makes it more difficult to identify such hallucinations.

IMPD-MACD-w/o Comprehensive: As shown in Table 2, removing the Comprehensive Agent does not lead to a clear decrease in any metric. We believe this may be due to the Judge Agent partially sharing the responsibilities of multi-round interaction and conclusion consolidation, thus diminishing the importance of the Comprehensive Agent. However, in longer contextual question-answering tasks, the Comprehensive Agent's ability to integrate and coordinate multiple perspectives would presumably play a more crucial role.

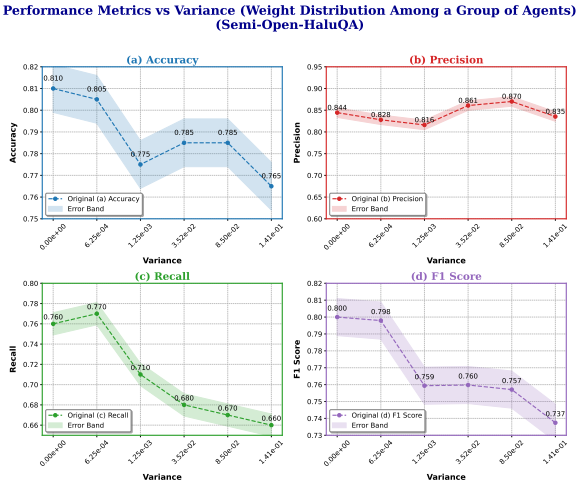

Figure 2: Influence of agent weight distribution on performance indicators in Semi-Open-HaluQA dataset

4.5 Hyper-Parameter Analysis

In this study, multiple intelligent agents were introduced within each faction, necessitating an evaluation of their relative contributions to the overall decision-making process. To address this, each agent in our model was assigned a corresponding weight (w_i), and we examined variations in these weights—specifically the variance among them—to assess the effectiveness of different weighting approaches and their impact on model performance.

$$\mathrm{Var}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(w_i - \bar{w}\right)^2 \tag{20}$$

Figure 2 and Figure 3 display how the model's performance metrics vary under different variance settings. Notably, when the variance is relatively small, the model demonstrates superior results in terms of Accuracy, Recall, and F1 scores. This finding indicates that, when each group contains only a limited number of agents, it is advisable to maintain a relatively balanced distribution of decision-making authority among them so as to optimize overall performance.

In addition, we further examined the impact of different numbers of debate rounds on model performance. Figure 10 and Figure 11 (Appendix E.3) show how the IMPD-MACD model performs under varying debate rounds on the Semi-Open-HaluQA

Method	Metrics (Semi-Open-HaluQA)				Metrics (HaluEval)			
	Accuracy	Precision	Recall	F1_Score	Accuracy	Precision	Recall	F1_Score
IMPD-MACD	**0.820**	**0.827**	**0.810**	**0.818**	**0.795**	**0.798**	**0.790**	**0.794**
w/o EK	0.760	0.781	0.740	0.755	0.715	0.809	0.570	0.661
w/o Logic	0.781	0.809	0.740	0.760	0.685	0.740	0.570	0.644
w/o Fact	0.785	0.802	0.750	0.775	0.740	0.817	0.630	0.711
w/o Comprehensive	0.810	0.816	0.810	0.813	0.735	0.774	0.720	0.746
w/o Consistency	0.780	0.784	0.760	0.784	0.730	0.767	0.660	0.710

Table 2: Performance of IMPD-MACD and its ablated versions on Semi-Open-HaluQA and HaluEval datasets.

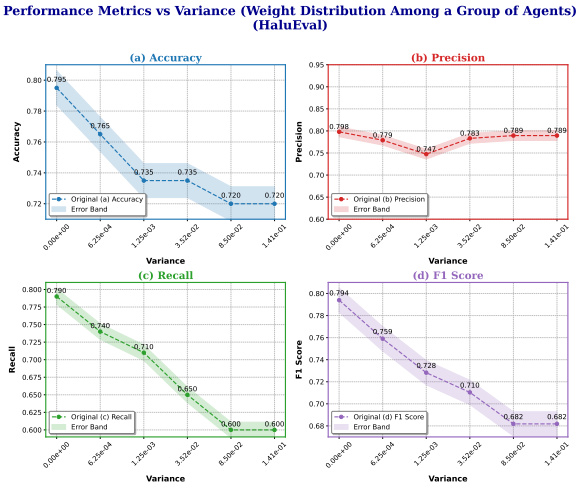

Figure 3: The influence of agent weight distribution on performance indicators in HaluEval dataset

and HaluEval datasets, respectively. The results indicate that the model often reaches optimal performance in the middle rounds on both datasets. The underlying reason could be that, after a moderate number of rounds, most contentious points have been thoroughly discussed and begin to converge. Continuing to increase the number of debate rounds may instead introduce redundant or imprecise information, thereby reducing overall performance (As shown in Equation (16)). Therefore, setting an appropriate number of debate rounds in a multi-agent debate framework is crucial for ensuring both accuracy and efficiency.

4.6 Cross-Model Robustness Evaluation

To comprehensively evaluate the applicability and robustness of the proposed method in hallucination detection across diverse LLM environments, this subsection presents a systematic analysis in which we substitute the underlying LLMs within the IMPD-MACD framework while preserving its core algorithmic architecture. (More robustness evaluation results, see Appendix E.4, Table 4)

The experimental results in Table 3 demonstrate the remarkable performance stability of the IMPD-MACD framework across different underlying LLMs. Specifically, the framework's hallucination detection performance maintains its efficacy without significant degradation when switching between different foundation models, and notably exhibits an improving trend as the parameter scale of the underlying LLMs increases. Our experiments reveal a positive correlation between the parameter scale of the underlying LLMs and detection performance, which not only validates the robustness of our proposed method but also highlights its scalability in LLM environments. These findings provide strong empirical evidence for both the effectiveness and generalizability of the IMPD-MACD method in hallucination detection tasks.

Model	Metrics	
	Accuracy	**F1_Score**
GPT-3.5-Turbo	0.780	0.766
GPT-4o-mini	0.820	0.818
gemini-1.5-flash	0.830	0.823
moonshot-v1-8k	0.840	0.835
claude-3-haiku-20240307	0.855	0.856

Table 3: IMPD-MACD Performance Under Different Base Agent Models (Simplified Table)

5 Conclusion

This study focuses on detecting hallucination in LLMs within semi-open Q&A scenarios. The multi-perspective, multi-agent interactive debate framework proposed in this paper provides an effective solution for identifying hallucinations; Meanwhile, this paper presents an empirical formula for quantifying the probability of agent hallucinations in multi-agent debates; Experimental results show that our approach significantly outperforms SOTA approaches on multiple evaluation metrics.

Limitations

Despite proposing the novel IMPD-MACD framework for hallucination detection in LLMs, this framework still exhibits significant limitations. The design involving multiple agents and multi-turn interactions leads to a substantial increase in token consumption, and when processing longer contexts, the required inference time is considerably extended. Furthermore, although the summaries generated by the Refiner agent can effectively pinpoint specific hallucinated content, most of the detected hallucinations remain at a relatively coarse granularity at the sentence level, which is insufficient for hallucination detection at the token-level granularity. And more important, even with the introduction of more stringent interaction and verification mechanisms, there remains a risk that multi-agent systems may introduce new erroneous information during the hallucination detection process. This could potentially lead to the further propagation and exacerbation of errors in the generated text. Future research will focus on overcoming these limitations to enhance the accuracy and efficiency of hallucination detection.

Future Work

Looking ahead, we anticipate extending this approach to more complex Q&A contexts so that the outputs generated by LLMs in longer-context or fully open Q&A scenarios can attain higher reliability and accuracy.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *ArXiv*, abs/2305.14325.

Yi Fang, Moxin Li, Wenjie Wang, Hui Lin, and Fuli Feng. 2024. Counterfactual debating with preset stances for hallucination elimination of llms. *ArXiv*, abs/2406.11514.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. https://openreview.net/forum?id=SBbJICrglS#all. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Mengya Hu, Rui Xu, Deren Lei, Yaxi Li, Mingyu Wang, Emily Ching, Eslam Kamal, and Alex Deng. 2024. Slm meets llm: Balancing latency, interpretability and consistency in hallucination detection. *ArXiv*, abs/2408.12748.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *ArXiv*, abs/2307.10236.

Geoffrey Irving, Paul Francis Christiano, and Dario Amodei. 2018. Ai safety via debate. *ArXiv*, abs/1805.00899.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. 55(12).

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *ArXiv*, abs/2402.07401.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

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*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection. *ArXiv*, abs/2405.01593.

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*, pages

17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

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

Ze-Yuan Liu, Peng-Jiang Wang, Xiao-Bin Song, Xin Zhang, and Ben-Ben Jiang. 2024b. Survey on hallucinations in large language models. *Journal of Software*, 36:1–34.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

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*, 36.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *ArXiv*, abs/2310.12397.

Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. 2024. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. *ArXiv*, abs/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. *ArXiv*, abs/2401.01313.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *ArXiv*, abs/2310.08118.

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. *ArXiv*, abs/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 (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2024. Debunc: Mitigating hallucinations in large language model agent communication with uncertainty estimations. *ArXiv*, abs/2407.06426.

A Semi-open Questions

In this paper, we focus on semi-open questions. Semi-open questions are those where the answer is not explicitly given in the question, yet there exists a correct or standard answer. These types of questions typically require reasoning, querying, or knowledge retrieval to determine the answer, differing from completely unconstrained open-ended questions. To better understand the uniqueness of semi-open questions, this paper first introduces the characteristics of open questions and closed-ended questions, and then compares their similarities and differences with semi-open questions.

Open questions. It allows respondents to answer completely freely, with researchers providing no options. Respondents can freely express personalized views or describe situations. Such questions are typically used in exploratory research to help understand deeper emotions, needs, or opinions. Figure 4 below is a simple example.

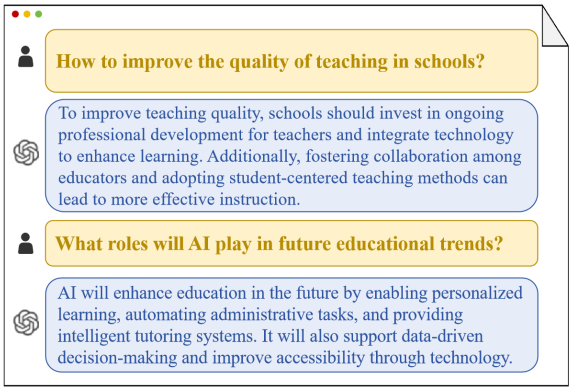

Figure 4: An example of open questions

Closed-ended questions. Closed-ended questions are those where a clear correct answer can be found within the question itself. These types of questions are usually presented in the form of multiple-choice or summary questions, where the correct answer can be directly identified from the question.

Semi-open questions. Semi-open questions combine characteristics of both open questions and closed-ended questions. They guide or limit the scope of responses to some extent, but do not strictly confine the answer like closed-ended questions. The following Figure 6 is a simple example. Semi-open questions usually have the following characteristics:

- Reasoning or Querying: Respondents usually need to search, reason, or query external resources to find the correct answer.
- Verifiability of the Answer: Although the answer is not directly given, once found, it is unique and verifiable.
- Presence of a Standard Answer: Even though not all information is provided in the question, there usually exists a standard answer recognized as correct.

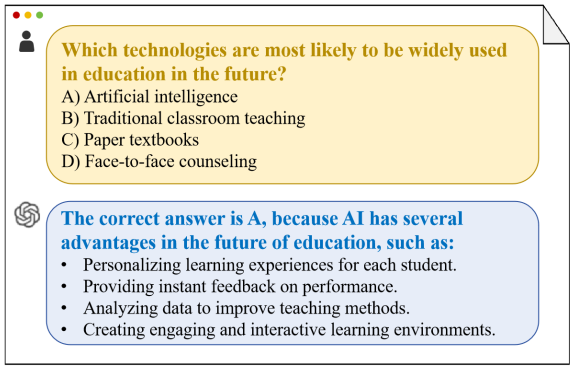

Figure 5: An example of closed-ended questions

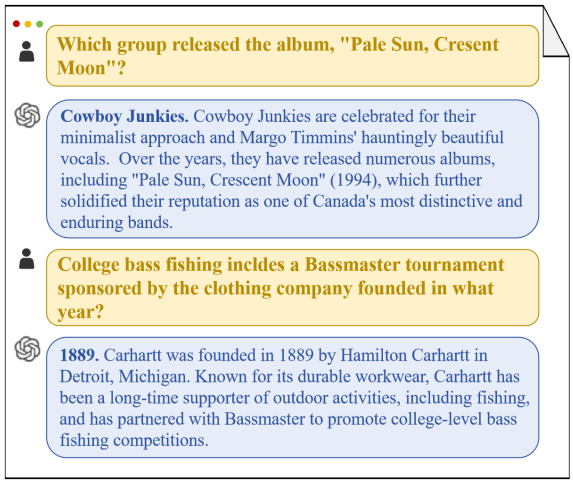

Figure 6: An example of semi-open questions

B Information-Gathering Agent

To more intuitively depict the role of Information-Gathering Agent in this framework, let Q represent the input question, D represent the external document corpus, d_j be the candidate document retrieved from the document corpus, and $P_r(d_j \mid Q)$ represent the retrieval probability distribution (or similarity) function. Each document is combined

with Q to serve as input to the generative model, which produces candidate output y.So its output distribution can be expressed as a weighted sum over all candidate documents:

$$p_\theta(y|Q) = \sum_{d_j \in D} p_r(d_j|Q) \times p_\theta(y|Q, d_j) \quad (21)$$

among:

- The $p_r(d_j|Q)$ is provided by the retrieval model, which measures the relevance of document d_j to question Q.
- The $p_\theta(y|Q, d_j)$ is provided by the generative model (i.e., the subsequent multi-perspective multi-agent debate section), which is used to measure the probability of generating an output y given the question Q and document d_j.

Through the aforementioned weighting method, external retrieval information is naturally integrated with the model's internal representations, achieving a "retrieval + generation" synergy. This reduces potential biases during the generation phase that may arise from solely relying on the internal model. This agent consists of the following four basic steps: question decomposition, Google search, knowledge acquisition, and knowledge refinement. These steps together form a systematic framework for effectively extracting and organizing information. The overall framework diagram of this part is shown in the Figure 7 below.

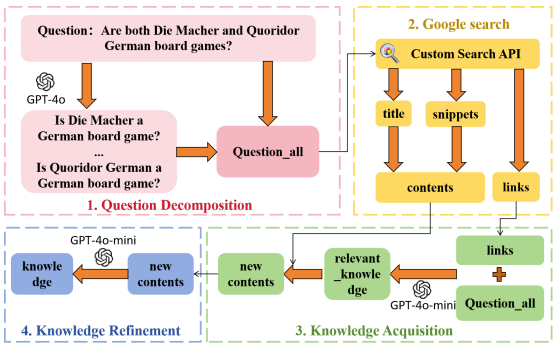

Figure 7: Information-Gathering Agent flow frame diagram

Question Decomposition. The first step in knowledge acquisition is question decomposition, which aims to break down complex questions into smaller, more specific question units and extract relevant keywords to better consolidate knowledge. This process utilizes the GPT-4o model, interacting with the model to decompose the question into multiple sub-questions and generate corresponding search keywords. This approach ensures the precision and relevance of the search.

Google search. After completing question decomposition, we use the Google Search API to search the decomposed keywords and the original question separately, in order to obtain titles, web links, and summaries related to each keyword. It is important to note that this API search method cannot directly access the most useful knowledge within the webpages, so we save the links to the webpages for later use.

Knowledge Acquisition. The main task of knowledge acquisition is to extract specific knowledge related to the question from the search results. This step uses the GPT-4o-mini model, combining the question and links to extract the most relevant knowledge fragments from the webpages. These fragments are then combined with the titles and summary knowledge from the previous retrieval step. This process ensures the relevance and accuracy of the information.

Knowledge Refinement. Since each question generates more than one query link, the acquired knowledge fragments may be redundant. Therefore, the purpose of the knowledge refinement step is to summarize and deduplicate this knowledge, extracting the most core and accurate information. This step also uses the GPT-4o-mini model to summarize the input knowledge fragments, producing concise knowledge statements. This process ensures the simplicity and accuracy of the final knowledge.

C Supplementary Information on the IMPD-MACD Approach

C.1 Details of Each Agent

This section will sequentially analyze each type of agent in the multi-agent collaboration framework, covering their structural characteristics, functional positioning, and prompt design. By dissecting and analyzing these agents in detail, one can gain a deeper understanding of the principles of coordination among the modules, thereby achieving better overall performance in hallucination detection.

Comprehensive Agent. In IMPD-MACD approach, the Comprehensive Agent takes on the core responsibility of integrating and deeply analyzing

the viewpoints and opinions of other agents, aiming to identify potential hallucination areas in the answers. The analysis report output by this agent includes the relevance between the question and the answer, traceable evidence information, and a comprehensive explanation of the identified issues. This provides more targeted input for subsequent agents.

In terms of prompt design, it is first necessary to clarify the types of hallucinations that this task focuses on in the hallucination definition explanation. This enables the Comprehensive Agent to identify different dimensions such as logical hallucinations, factual hallucinations, and contextual inconsistencies.Secondly, by presenting positive and negative examples, the agent is shown which responses can be considered "hallucination-free" and in which situations there are potential signs of hallucination.Finally, in the analysis guidelines section, the Comprehensive Agent needs to actively search for possible points of hallucination from the Pro perspective and question statements lacking dataset support from the Con perspective. This approach helps avoid overlooking any potential sources of hallucination.

In practice, the Comprehensive Agent compares the Pro/Con analysis results it generates with the external knowledge base and the problem background, and outputs them in a structured manner. On this basis, subsequent agents can use the information provided to more efficiently examine possible hallucinations.

Context Consistency Agent. In IMPD-MACD approach, the Context Consistency Agent is primarily responsible for examining the internal logical coherence of the answers and their alignment with the questions, with particular attention to whether there is "local contradiction" or "self-contradictory" content in the contextual semantics. To achieve this goal, the agent can extract consistency-related parts from the analysis results of other agents (opposing agents) and further conduct secondary verification or provide additional explanations, thereby ensuring a comprehensive review of the answer's context and narrative flow.

In specific implementation, the prompt design for the Context Consistency Agent focuses particularly on the following aspects:Firstly, ensuring that the answer remains closely related to the question itself, avoiding errors caused by missing information or deviation from the topic.Secondly, checking whether the answer is internally coherent to prevent contradictions or logical breaks.Thirdly, integrating points raised by other agents in the multi-agent debate environment to comprehensively assess the completeness and consistency of the answer.Fourthly, providing appropriate explanations or executing necessary corrections when the Judge Agent questions the consistency.Through these means, the Context Consistency Agent plays a crucial role in the multi-round interaction process. Once an inconsistency or potential conflict in the answer is detected, it can promptly alert the relevant modules and mark possible hallucinations, making the system's overall responses more consistent and reliable.

Factual Agent. The Factual Agent is primarily responsible for examining whether the objective facts in the answer align with existing knowledge and for extracting searchable entities or key facts from the answer to identify and locate factual hallucinations. Its internal structure typically includes two key modules: fact extraction and verification.On one hand, it extracts key information from the answer text that can be used for queries and matches it with external knowledge sources or databases. On the other hand, it assesses the relevance and reliability of this information against authoritative knowledge to determine whether there are factual errors or omissions in the answer.

In prompt design, this agent considers both Pro and Con analytical directions. On the one hand, it identifies and verifies assertions supported by ample evidence to ensure that the facts in the answer have sufficient external data or literature to back them up. On the other hand, if the answer content is found to lack support in the knowledge base or if inconsistencies with external evidence are detected, the system raises questions and flags these potential factual hallucinations.Thus, the Factual Agent assumes the role of "objectivity assurance" within the entire model framework. Through continuous verification and monitoring processes, it effectively identifies and detects factual hallucinations present in the generated answers.

Logic Agent. The Logic Agent is primarily responsible for examining whether the reasoning process in the answer possesses sufficient rationality, including the completeness of the argument chain and the close connection between premises and conclusions. Its core function lies in the layered dissection of the answer's reasoning chain to identify

potential flaws or unreasonable transition points. Through this multi-level analysis, the Logic Agent can identify possible reasoning errors in the answer and assess whether the premises and conclusions are indeed logically supportive of each other.

Pro team Logic Agent Prompt

```
As a Fact Checker supporting NO HALLUCINATION position, verify:
Question: {context['question']}
Answer: {context['answer']}

Note: Although opponent_text contains the last round of analysis for
multiple agents, you only need to consider the analysis of agents
whose functions are aligned with yours.

{opponent_text}{judge_text}

Focus on:
1. Identifying supported claims
2. Matching knowledge evidence
3. Latest opponent arguments and their factual claims
4. Judge's latest feedback on factual accuracy
5. Counter-evidence to opponent's fact-checking

Please provide a concise response focusing on the most critical points or findings.

Format:
SUPPORTED FACTS:
[List key supported facts]

CONCLUSION:
[FACTUAL/NOT FACTUAL]

Attention: Your answer does not need to contain too much irrelevant analysis,
just give the main points and the most important answers.
```

Figure 8: Pro-team Agent Prompt Design (Using the Logic Agent as an Example)

In specific implementation, the Logic Agent analyzes from both positive and negative directions. In the positive (Pro) direction, the agent searches for reasonable reasoning paths within the answer, evaluates whether there are appropriate logical connections between premises and conclusions, and tracks the validity of the reasoning chain to ensure that the answer has a solid reasoning foundation.In the negative (Con) direction, the agent identifies various potential logical fallacies, including circular reasoning, false causality, or other common reasoning flaws. It also tracks reasoning gaps in the answer and questions statements that clearly lack transitions or evidential support.Through these methods, the Logic Agent provides the Judge with more comprehensive reference opinions, offering more substantial grounds for determining whether there are logical hallucinations in the answer.

Judge Agent. In this multi-agent collaboration framework, the Judge Agent plays a crucial role by synthesizing the outputs of various agents and making the final judgment and decision. During the analysis process, it first summarizes and organizes the opinions from multiple sources. Then, using pre-established evaluation criteria, it provides a clear conclusion on whether the answer contains hallucinations. If major issues are detected in the answer, it suggests appropriate corrections.The judge's decision typically includes information such as the overall conclusion, weighted analysis, consensus status, and the final winning side. This provides the entire system with consistent and traceable output.

Con team Logic Agent Prompt

```
As a Fact Checker supporting HALLUCINATION EXISTS position, verify:
Question: {context['question']}
Answer: {context['answer']}

{opponent_text}{judge_text}

Focus on:
1. Identifying unsupported claims
2. Finding knowledge gaps
3. Latest opponent arguments and their factual errors
4. Judge's latest feedback on factual concerns
5. Refutation of opponent's fact-checking

Please provide a concise response focusing on the most critical points or findings.

Format:
UNSUPPORTED CLAIMS:
[List key unsupported claims]

CONCLUSION:
[FACTUAL/NOT FACTUAL]

Attention: Your answer does not need to contain too much irrelevant analysis,
just give the main points and the most important answers.
```

Figure 9: Con-team Agent Prompt Design (Using the Logic Agent as an Example)

When making a judgment, the judge needs to consider the weight coefficients assigned to each agent in the system, synthesizing various viewpoints and making selections and judgments based on the strength of arguments and the quality of evidence, ensuring that each argument is thoroughly evaluated. The credibility of external knowledge and the completeness of the internal reasoning chain are also taken into account to avoid hasty conclusions when information is incomplete or logic is not rigorous. If the majority of agents reach the same conclusion, the reliability of the answer can be determined with relative certainty. However, if there are significant disagreements, the judge has the duty to facilitate further rounds of debate, or to output an "uncertain" conclusion when consensus cannot be reached.

By balancing different perspectives and sources of evidence, the judge entity ultimately provides a unified and transparent decision for the system's hallucination detection, laying a solid foundation for subsequent applications or interpretative analysis.

Prompt design. Within the framework of multi-agent collaboration, the prompt design for each agent must maintain adversarial qualities while achieving a certain degree of coordination to op-

timize the identification of hallucinations. First, ensure that the output content is structured, facilitating reading and integration by other agents or subsequent modules. When providing different types of opinions (Pro/Con), use clear labels or fields to distinguish them.Secondly, adversarial design in the prompts is crucial. Each agent, during the analysis process, needs to have both Pro and Con modes of thinking. They should be able to question the opponent's viewpoints or conclusions and incorporate opponent information in their outputs for rebuttal or supplementation, thus encouraging diverse and in-depth arguments.

In addition, in scenarios involving multi-round interactions and decision-making mechanisms, prompts must have contextual awareness. By allowing each agent to reference previously generated historical information (such as opponent arguments or feedback from the Judge) and providing a sufficient context window, agents can continue to track and relate to prior discussion content in subsequent debates, thereby enhancing the continuity and accuracy of the reasoning process. Lastly, adhering to the principle of simplicity is an indispensable part of prompt engineering. To avoid unnecessary lengthy analysis, it is important to highlight the core arguments and establish key metrics or a checklist of issues to help each agent quickly focus on the parts most prone to hallucination.

Through the above design principles, each agent can maintain an independent and adversarial analysis style while being able to coordinate and share information when necessary. This ultimately allows for the precise identification of hallucinations in multi-perspective, multi-role interactions. This systematic prompt engineering approach is highly versatile in complex question-answering and reasoning scenarios and can provide a feasible reference model for subsequent applications and improvements in other fields.

C.2 Interaction Quality between Agents

The overall framework for interaction quality assessment is defined as:

$$Q_t = \sum_{i=1}^{4} w_i \cdot q_t^i(A_i, \text{Opp}(A_i)) \tag{22}$$

In the LLM inference process, we first construct a unified input vector: $X_i = [A_i; \text{Opp}(A_i); h_t]$, where the historical information (Defined by (9)) tuple $h_t = (\mathcal{H}_t(A_i), \mathcal{H}_t(\text{Opp}(A_i)))$. The conditional probability is calculated through multi-layer attention mechanism:

$$P(q_t^i|X_i) = \int P(q_t^i|Z_i)P(Z_i|X_i)dZ_i \tag{23}$$

Due to the fact that Z_i contains sufficient information to fully confirm q_t^i, we can consider $P(q_t^i|Z_i)$ to be approximately equal to 1 in application. where Z_i represents the latent semantic space, computed via attention mechanism:

$$P(Z_i|X_i) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{24}$$

The interaction quality assessment comprises three dimensions q_{sem}^i, q_{info}^i and q_{time}^i:

$$q_{\text{sem}}^i = \text{softmax}(\text{Cos}(A_i, \text{Opp}(A_i))) \tag{25}$$

$$q_{\text{info}}^i = \frac{\mathcal{H}_e(A_i|h_t)}{\mathcal{H}_{\text{e\_max}}^t} \tag{26}$$

In information value assessment, semantic units $s_j \in \mathcal{S}$ are hierarchically divided into: core arguments, key phrases, and basic tokens. The conditional entropy is calculated as:

$$\mathcal{H}_e(A_i|h_t) = -\sum_{s_j \in \mathcal{S}} P(s_j|h_t)\log_2 P(s_j|h_t) \tag{27}$$

with conditional probability (γ is the compression ratio defined by equation (6)):

$$P(s_j|h_t) = \frac{\text{count}(s_j, \{A_i, \text{Opp}(A_i)\}) + \gamma}{\sum_k \text{count}(s_k, \{A_i, \text{Opp}(A_i)\}) + |\mathcal{S}|\gamma} \tag{28}$$

The temporal efficiency is measured by (We use debate rounds as a substitute for interaction time):

$$q_{\text{time}}^i = e^{-\beta \Delta t_i} \tag{29}$$

Where β is the temporal accumulation coefficient. The complete synthesis calculation is:

$$q_t^i = \int q_{\text{sem}}^i \cdot q_{\text{info}}^i \cdot q_{\text{time}}^i \cdot P(q_t^i|X_i)dq_t^i \tag{30}$$

In practical applications, we adopt the approximation: $q_t^i \approx q_{\text{sem}}^i \cdot q_{\text{info}}^i \cdot q_{\text{time}}^i$.

This assessment mechanism possesses the following theoretical properties: consistency

($\lim_{|h_t|\to\infty} \text{Var}(q_t^i) \to 0$), boundedness ($q_t^i \in [0,1]$), and temporal monotonicity ($\Delta t_i > \Delta t_j \implies q_{\text{time}}^i < q_{\text{time}}^j$). Through the integration of LLMs, multi-level semantic unit analysis, and probabilistic reasoning, we achieve precise quantification of debate interaction quality.

C.3 Judge Score

The Judge Agent employs a multi-dimensional scoring system to evaluate debate quality. The scoring process consists of the following components:

Input Processing: The debate content is first processed through tokenization and embedding:

$$X_t = \text{Tokenize}([C_t^{\text{team1}}; C_t^{\text{team2}}; Q_t]) \tag{31}$$

According to the Transformer encoder:

$$E_t = W_E X_t + \text{PE}(X_t) \tag{32}$$

Feature Extraction: Contextual features are extracted using multi-head attention:

$$Z_t = \text{MultiHeadAttention}(E_t) \tag{33}$$

Dimension Scoring: Four core dimensions ($i = 1, 2, 3, 4$) are evaluated:

$$s_i = \text{MLP}_i(Z_t) \tag{34}$$

where:

- s_1: Logic Score (reasoning rigor)
- s_2: Fact Score (information accuracy)
- s_3: Coherence Score (contextual consistency)
- s_4: Comprehensive Score (overall quality)
- MLP represents the layer following Transformer Attention

Each score satisfies $s_i \in [0, 1]$.

Final Scoring: A confidence score modulates the final judgment:

$$c_t = \sigma(\text{MLP}_{\text{confidence}}(Z_t)) \tag{35}$$

The final judgment score combines dimensional scores with weights:

$$J_t = (\sum_{i=1}^{4} w_i s_i) \cdot c_t \tag{36}$$

subject to:

- $\sum_{i=1}^{4} w_i = 1$
- $w_i \geq 0$
- $J_t \in [0, 1]$

This scoring mechanism ensures comprehensive evaluation of debate quality through multiple dimensions, with weights w_i adjustable based on specific requirements.

C.4 Hallucination Rate of Agents as a Function of Rounds

The following provides a detailed breakdown of Equation (16):

- A_i: Agent i in the system
- h_t: Interaction history up to time t
- S: Set of possible states
- $P(s_j|h_t)$: Probability of state s_j given history h_t

The fundamental information measures are defined as follows (Equation (27)):

$$\mathcal{H}_e(A_i|h_t) = -\sum_{s_j \in S} P(s_j|h_t) \log_2 P(s_j|h_t) \tag{37}$$

The effective information is given by the following equation:

$$\begin{aligned} I(A_i, A_j|h_t) =& \mathcal{H}_e(A_i|h_t) + \mathcal{H}_e(A_j|h_t) \\ & - \mathcal{H}_e(A_i, A_j|h_t) \end{aligned} \tag{38}$$

Calculation of Irrelevant Information:

$$\begin{aligned} I_{irr}(t) = & \sum_{i=1}^{4} w_i \mathcal{H}_e(A_i|h_t) - \eta \sum_{i \neq j} I(A_i, A_j|h_t) \\ & + \beta t + \gamma(1-\rho)\mathcal{H}_{rel}(t) \end{aligned} \tag{39}$$

$$\mathcal{H}_{rel}(t) = \sum_{i=1}^{4} \mathcal{H}_e(A_i|h_t) - \mathcal{H}_{prior} \tag{40}$$

The probability framework is characterized by:

$$P(\text{hallucination})_t = g(I_{irr}(t)) \cdot (e^{-\lambda t} + \phi \mathcal{H}(t)) \tag{41}$$

with supporting functions:

$$g(I_{irr}) = \eta \tanh(I_{irr}) \tag{42}$$

$$\mathcal{H}(t) = \frac{1}{1 + e^{-\gamma(t-t_0)}} \tag{43}$$

System parameters and their constraints:

- $\rho \in (0, 1)$: Inter-agent correlation coefficient.
- $w_i > 0$: Agent weights, $\sum_{i=1}^{4} w_i = 1$.
- $\eta \in (0, 1]$: Mutual information impact factor. In the experiment, we set $\eta = 0.3$.
- $\beta > 0$: Temporal accumulation coefficient. In the experiment, we set $\beta = 0.05$.
- $\gamma > 0$: Compression ratio.
- $\lambda > 0$: Learning rate.
- $\phi \in (0, 1)$: Entropy influence coefficient. In the experiment, we set $\phi = 0.5$.
- $t_0 > 0$: Critical round point, which represents a crucial parameter to be identified.

System characteristics:

1. Rounds Evolution:
    - Early phase ($t < t_0$): Dominated by $e^{-\lambda t}$.
    - Transition phase ($t = t_0$): Balanced effects.
    - Late phase ($t > t_0$): Influenced by $\mathcal{H}(t)$.
2. Implementation Guidelines:
    - Set $\mathcal{H}_{prior} \approx \log_2(n)$ for n-class problems. Therefore, $\mathcal{H}_{prior} \approx 1.0$ in our method.
    - Adjust w_i based on agent performance.
    - While we designed our experiments and optimized the model based on our derived formula (Equation (16)), we have not conducted systematic validation experiments for this formula. Nevertheless, the experimental results demonstrate that the observed trends in agent hallucination with respect to debate rounds and irrelevant information align with our theoretical predictions.

D The Semi-Open-HaluQA Dataset

The Semi-Open-HaluQA dataset comprises a series of Q&A entries spanning multiple factual domains—such as people, history, geography, and literature—to evaluate and analyze large language models' accuracy when answering factual questions as well as any potential hallucinations. Its core elements include: question, model answer, knowledge (authoritative information), and label (a binary annotation used to assess the correctness of the answer). All questions are meticulously selected from the HaluEval dataset to ensure that the content is objective and verifiable, and the model answers are generated by GPT-4. To better preserve context and external information, this dataset contains 1,800 Q&A pairs and their corresponding manual annotations. In addition, we merge the knowledge collected by the Information-Gathering Agent during model execution into the original HaluEval dataset's "old_knowledge" field to create the new dataset.

Based on these diverse questions and corresponding answers, researchers can conduct analyses on multiple levels: Firstly, it can be used to evaluate whether models can accurately call external knowledge and output the answers consistent with objective facts; Secondly, it can test whether models possess the ability to recognize and correct errors when faced with interference information similar to "hallucinated" answers. Thirdly, in interpretability studies, by comparing model outputs with authoritative knowledge, researchers can delve into the potential sources of bias in large models and their impact on Q&A conclusions. There are also plans to expand the types of questions and areas of knowledge further to meet the needs for broader evaluation of large language models in multi-disciplinary and multi-language environments in the future. To ensure data compliance and personal privacy security, the dataset has desensitization treatment and legal compliance review before release, with any information that may involve personal privacy removed.

When annotating each Q&A record, annotators first refer to authoritative sources such as encyclopedias, news reports, and academic literature to determine the most concise and accurate answer for the question. Then, the model output is compared with this correct answer and assigned a binary label: if the response matches the correct answer in terms of key information, it is labeled as 1; otherwise, it is

labeled as 0. Since some questions involve tracing and researching time, personal background, or factual details, the annotation process employs cross-verification and double-checking mechanisms to ensure the consistency and accuracy of the labels. The following example illustrates a typical record in the dataset:

- Question: "Are Anita Shreve and Elizabeth Jane Howard the same nationality?"
- Answer: "No, Anita Shreve was American, and Elizabeth Jane Howard was British."
- Knowledge: "Anita Shreve (born 1946) is an American writer. Elizabeth Jane Howard ... was an English novelist."
- Label: 1

E Experimental Supplement

E.1 Adjustments to the MAD Method

Universality: The new system is no longer limited to processing specific types of biographical data. Instead, by using more standardized input-output formats, it supports the processing and evaluation of various types of question-and-answer data.This universality makes the system easier to integrate with other platforms or tools, expanding its applicability across a wider range of use cases.

Error handling and recovery mechanism: The new system has been comprehensively enhanced in terms of error handling, offering a task retry mechanism of up to three attempts: If the API call fails, the system will attempt again after 20 seconds. If a round of processing fails to complete all tasks, the system will automatically recover the unfinished entries and refill the queue, and then initiate a new round of processing to ensure that the system can maintain stable operation even under adverse conditions such as network instability or API limitations.

Evaluation method: The evaluation mechanism has been expanded from a simple binary "correct/incorrect" judgment to a multi-dimensional comprehensive evaluation. While calculating the basic accuracy, the new system further introduces key metrics such as Precision, Recall, and F1 Score, and retains the detailed evaluation process and results.By recording and analyzing error cases, researchers can have a more comprehensive understanding of the system's potential defects, providing an objective basis for subsequent improvements.

Task processing mechanism: The new system achieves a structural upgrade from serial to parallel by a introducing queue and a multi-threaded processing mechanism.The system is configured with multiple worker threads (defaulting to 30), which retrieve pending entries from the task queue for concurrent processing. In the event of a processing failure, the task is placed back into the queue to ensure that each piece of data is processed.In order to further ensure data consistency in a multi-threaded environment, the new system also employs the thread lock mechanism, which can significantly improves throughput efficiency, especially when handling large-scale datasets.

E.2 Ablation Experiment

IMPD-MACD-w/o Fact: After removing the Factual Agent, there is no obvious change in the other metrics apart from Recall. We hypothesize that in the absence of a dedicated agent responsible for external fact verification, the multi-agent adversarial mechanism in IMPD-MACD may make the model relatively "over-cautious," thereby leading it to classify more answers as "containing hallucinations." This causes an increase in Recall, whereas Accuracy does not show a significant change. In a larger-scale dataset, the removal of this role might result in a more noticeable decline in Accuracy.

IMPD-MACD-w/o Consistency: After removing the Context Consistency Agent, although all metrics decline slightly, the drop is not pronounced. This suggests that the proportion of contextual-inconsistency hallucinations in current outputs of LLMs is relatively low. At the same time, such inconsistencies may be scattered throughout different parts of an answer, making it difficult for a single consistency checking agent to capture all of them.

E.3 Impact of Debate Rounds Hyperparameter

To determine the optimal number of debate rounds, we designed and conducted a series of systematic experiments.The results are detailed in Figure 10 and Figure 11. From the experimental data, it can be observed that as the number of debate rounds increases, the model's performance metrics (such as accuracy, recall, and F1 score) show significant improvement initially.However, when the number of debate rounds exceeds a certain threshold, there

Model	Metrics			
	Accuracy	Precision	Recall	F1_Score
GPT-3.5-Turbo	0.780	0.818	0.720	0.766
GPT-4o-mini	0.820	0.827	0.810	0.818
gemini-1.5-flash	0.830	0.859	0.790	0.823
moonshot-v1-8k	0.840	0.862	0.810	0.835
claude-3-haiku-20240307	0.855	0.852	0.860	0.856

Table 4: IMPD-MACD Performance Under Different Base Agent Models

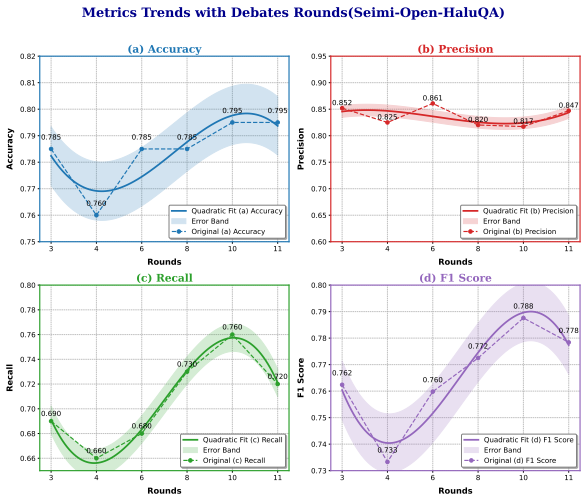

Figure 10: Evaluation Metrics vs. Debate Rounds on Semi-Open-HaluQA Dataset

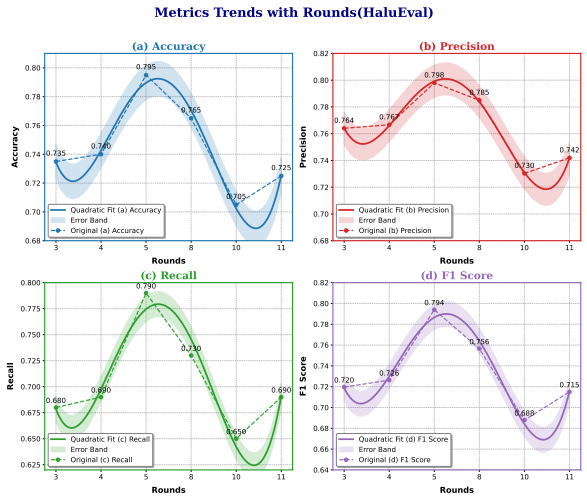

Figure 11: Evaluation Metrics vs. Debate Rounds on HaluEval dataset

is a tendency for performance to decline.This phenomenon indicates that there is an optimal number of debate rounds, which can effectively enhance model performance while avoiding resource wastage and potential misjudgments caused by excessive rounds.Additionally, the varying sensitivity of different datasets to the number of debate rounds also reflects the impact of data complexity on the choice of the optimal number of rounds.In summary, determining the appropriate number of debate rounds is crucial for enhancing the performance of a multi-agent collaboration framework in hallucination detection tasks.Additionally, this finding provides empirical evidence for optimizing the interaction design of multi-agent systems in the future, emphasizing the importance of adjusting the number of debate rounds in different application scenarios. Future research could further explore methods for dynamically adjusting the number of debate rounds to adapt to more complex and variable question-and-answer environments.

E.4 The Supplement of Cross-Model Robustness Evaluation

Table 4 presents the performance of the IMPD-MACD method on multiple core metrics under different base LLMs. As training knowledge continues to evolve and model parameter scales expand, the IMPD-MACD framework exhibits a steady upward trend across all evaluation metrics. This observation not only indicates that larger-scale LLMs can more fully unlock the method's reasoning and decision-making capabilities, but also underscores its robustness and generalizability when adapting to models of varying sizes and knowledge coverage. Equally noteworthy, the sustained performance gains suggest a positive synergy between the IMPD-MACD framework and more expressive models, effectively leveraging the extensive knowledge embedded in large models to achieve significant and stable improvements in hallucination detection tasks.