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

Anonymous ACL submission

Abstract

LLMs increasingly suffer from hallucinations that undermine their reliability and real-world utility. Existing single-agent detectors struggle to cover diverse hallucination types and are inapplicable to black-box models, while current multi-agent approaches lack clear role allocation and deep interaction, limiting the benefits of multi-perspective reasoning. Moreover, hallucination probability rises sharply beyond a critical debate "window" tied to the number of interaction rounds. We introduce IMPD-MACD, an Integrated Multi-Perspective, Diverse-Stance Multi-Agent Collaborative Debate framework that assigns specialized roles, injects external knowledge, and coordinates multi-round debates to jointly detect logical, factual, and context inconsistencies. A Poisson-based hallucination model formally characterizes the round-dependent window; a hierarchical visibility constraint and sliding summarization mechanism suppress irrelevant context, effectively delaying this window. A Judge-Agent monitors debate quality in real time and terminates deliberation before hallucination risk surges. Extensive experiments on multiple semi-open Q&A benchmarks show that IMPD-MACD consistently outperforms state-of-the-art detectors and optimally balances debate depth with accuracy, demonstrating strong generalization and robustness. We also release Semi-Open-HaluQA, a new dataset for hallucination evaluation, to foster future research.

1 Introduction

Large language models (LLMs), such as the GPT series (Brown et al., 2020) and LLaMA (Touvron et al., 2023), have achieved remarkable progress in natural language processing. However, they still often generate plausible but incorrect or unsupported content, commonly known as hallucination (Ji et al., 2023). This problem is especially critical in semi-open question answering, where the answer is not explicitly contained in the input but can be verified through external evidence; a detailed definition is provided in Appendix A.

Existing hallucination detection methods have made substantial progress, including Chain-of-Thought prompting (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). Nevertheless, important limitations remain. Single-agent methods are often unsuitable for black-box models, or suffer from overconfidence and repetitive reasoning biases (Fang et al., 2024; Valmeekam et al., 2023; Huang et al., 2023; Stechly et al., 2023; Varshney et al., 2023). Moreover, factual, logical, and contextual hallucinations usually require different forms of evidence and reasoning, making them difficult to capture with a homogeneous detector (Tonmoy et al., 2024; Lee et al., 2022; Atanasova et al., 2020).

Multi-agent methods mitigate some of these issues by introducing multiple reasoning paths, but existing debate frameworks still have limited role differentiation and weak cross-agent collaboration (Fang et al., 2024). As a result, they may fail to fully exploit multi-perspective reasoning for comprehensive hallucination detection (Liang et al., 2024; Yoffe et al., 2024). In addition, most prior studies focus on closed-ended question answering, leaving their effectiveness in more complex semi-open Q&A scenarios underexplored (Liang et al., 2024; Liu et al., 2024a; Longpre et al., 2021).

To address these challenges, we propose IMPD-MACD, an Integrated Multi-Perspective and Diverse-Stance Multi-Agent Collaborative Debate framework. Inspired by real-world debate, IMPD-MACD organizes agents into opposing Pro and Con teams and assigns each agent a specialized role for factual, logical, contextual, or comprehensive analysis. This design encourages complementary reasoning and adversarial verification, enabling the

system to expose hallucinations from multiple perspectives rather than relying on a single judgment path.

A key challenge in multi-round debate is that more interaction is not always beneficial. While early rounds can improve evidence coverage, excessive debate may introduce redundant context and new hallucinations. We therefore model the round-dependent hallucination risk and identify a critical "hallucination window". Based on this insight, IMPD-MACD incorporates external knowledge injection, hierarchical visibility constraints, sliding summarization, and a Judge-Agent convergence controller. These mechanisms preserve useful cross-agent interaction while suppressing irrelevant information growth and terminating the debate before hallucination risk increases.

We further construct Semi-Open-HaluQA based on HaluEval (Li et al., 2023) and evaluate IMPD-MACD on both datasets. Experiments against CoT (Wei et al., 2022), Self-Reflection (Shinn et al., 2024), Self-CheckGPT (Manakul et al., 2023), MAD (Du et al., 2023), MADR (Kim et al., 2024), and MC-MAD (Sun et al., 2024) show that our method achieves strong hallucination detection performance across different backbone models. In summary, the main contributions of this study are as follows:

(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 agent interaction method incorporating hierarchical visibility constraints and a sliding summarization mechanism, which effectively improves information accuracy and coordination efficiency during multi-agent interactions. (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 often generate factual inaccuracies or fabricated content, commonly known as hallucinations (Brown et al., 2020; Ji et al., 2023). Existing hallucination detection methods can be broadly divided into single-agent and multi-agent approaches.

2.1 Traditional Single-Agent Method

Single-agent methods address hallucination detection from the data, model, or application layer (Liu et al., 2024b). Data-level methods, such as training-data cleaning, are costly and difficult to apply to already deployed LLMs. Model-level methods, such as activation-based hallucination classifiers (Azaria and Mitchell, 2023), require access to internal states and are therefore unsuitable for black-box models. Application-level methods instead query the LLM itself (Fang et al., 2024). Chain-of-Thought and Self-Reflection improve reasoning transparency (Wei et al., 2022; Shinn et al., 2024), while diversified sampling methods such as Self-CheckGPT and InterrogateLLM estimate reliability by comparing multiple generations or reconstructed queries (Manakul et al., 2023; Yehuda et al., 2024). However, these methods may still inherit the model's over-confidence or fail to jointly capture factual, logical, and contextual hallucinations in semi-open Q&A. Hu et al. (Hu et al., 2024) further explore multi-category hallucination detection, but their approach still depends on human intervention or external validation, limiting full automation.

2.2 Multi-Agent Method

Multi-agent debate was first introduced as a way to expose reasoning flaws through adversarial interaction (Irving et al., 2018). Following this idea, Du et al. (Du et al., 2023) apply multi-agent debate to improve factuality and reasoning, Kim et al. (Kim et al., 2024) propose MADR with two-agent cyclical refinement, and Sun et al. (Sun et al., 2024) model debate with a Markov-chain-based framework. Despite their effectiveness, existing debate methods usually rely on homogeneous agents, weak role specialization, and unrestricted interaction histories. They are also rarely evaluated in semi-open Q&A scenarios that require external evidence and fine-grained inconsistency detection. In contrast, IMPD-MACD introduces role-specialized agents, opposing stances, external knowledge injection, and controlled information sharing to support more reliable hallucination detection.

3 Method

This section presents IMPD-MACD, a black-box hallucination detection framework for semi-open Q&A. Given a question Q and an initial answer A_0, the goal is to determine whether A_0 contains hal-

lucinations and to provide an evidence-grounded judgment. As illustrated in Figure 1, the framework consists of three main components: external knowledge gathering, multi-perspective stance-based debate, and hallucination-window-aware interaction control.

3.1 Information-Gathering Agent

Before the debate, the Information-Gathering Agent retrieves and summarizes external evidence relevant to the input question and the initial answer. This retrieval-augmented process reduces the reliance on the parametric knowledge of the backbone LLM and provides a shared evidence base for subsequent factual verification (Li et al., 2024). In practice, the agent decomposes the question, retrieves relevant evidence, and refines the retrieved content into concise knowledge statements. Detailed implementation is provided in Appendix B.

3.2 Multi-Perspective Agent Roles and Grouping

After external knowledge is collected, IMPD-MACD initiates a stance-based multi-agent debate. Agents are divided into two opposing teams: the Pro team argues that the original answer is hallucination-free, while the Con team argues that hallucinations may exist. This opposition encourages both supporting evidence and counter-evidence to be examined.

Each team contains four role-specialized agents: a **Factual Agent** for evidence-based fact verification, a **Logic Agent** for reasoning consistency, a **Context Consistency Agent** for checking alignment between the question and answer, and a **Comprehensive Agent** for aggregating role-specific opinions into a team-level conclusion. This design decomposes hallucination detection into factual, logical, and contextual perspectives while preserving team-level coordination. Detailed role definitions and prompt designs are provided in Appendix C.

3.3 Dual-Mechanism Poisson Process Modeling for Hallucination Probability

Multi-round debate can improve hallucination detection by exposing the answer to diverse evidence and opposing viewpoints. However, excessive interaction may also introduce redundant context and new hallucinations. We model this trade-off through a round-dependent hallucination risk. At debate round r, the risk is affected by two competing factors: early external knowledge suppresses hallucination, while increasing context length introduces long-context noise:

$$\lambda_r = \underbrace{\lambda_0 e^{-\beta r}}_{\text{knowledge suppression}} + \underbrace{\kappa n_r^2}_{\text{context noise}}, \quad (1)$$

where λ_0 denotes the initial hallucination propensity, β controls the decay of the knowledge benefit, κ is the context-noise coefficient, and n_r is the visible context length at round r.

This formulation suggests a hallucination window: debate is beneficial before context noise dominates, but excessive rounds may reduce reliability. Therefore, IMPD-MACD controls interaction through hierarchical visibility, sliding summarization, and Judge-Agent early stopping. The full derivation is provided in Appendix D.1.

3.4 Multi-Round Interactions Among Agents

To prevent multi-round debate from accumulating excessive irrelevant context, IMPD-MACD adopts a hierarchical visibility constraint with sliding summarization. Agents with the same role but opposite stances are allowed to read each other's complete output, enabling direct professional confrontation. In contrast, outputs from different-role agents are compressed by the Judge Agent into concise summaries before being shared. This design preserves the most relevant disagreement signals while reducing redundant cross-role information.

At the end of each round, the Judge Agent summarizes the current debate state and broadcasts a refined summary to guide the next round. The Judge only keeps the complete debate traces of the most recent rounds, while older information is compressed into summaries. This sliding mechanism ensures that each agent receives sufficient context for reasoning without exceeding the model context window or introducing excessive irrelevant information.

Under this mechanism, IMPD-MACD controls the effective context length exposed to each agent and delays the point at which context noise dominates the benefit of knowledge injection. Detailed visibility rules, context-length analysis, and the derivation of the effective hallucination-window boundary $r^{\star}_{\text{eff}}$ are provided in Appendix D.2.

3.5 Multi-Agent Debate Process

In the formal debate phase, IMPD-MACD proceeds in multiple rounds. At each round, agents

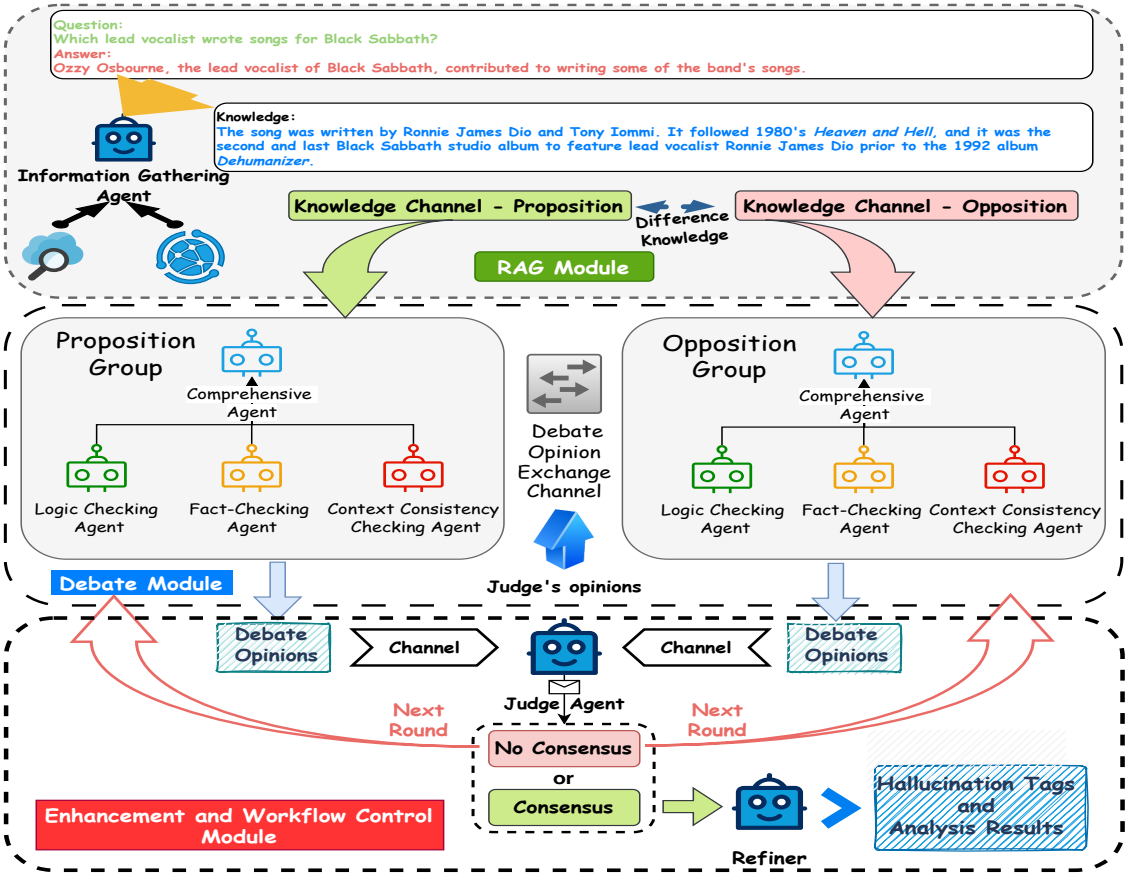

Figure 1: The proposed IMPD-MACD framework

receive the original question, the initial answer, the retrieved external knowledge, and the visible debate history controlled by the hierarchical visibility mechanism. Each role-specialized agent first generates its own analysis from the assigned stance, and the Comprehensive Agent then aggregates the role-level outputs into a team-level conclusion.

After both Pro and Con teams submit their conclusions, the Judge Agent evaluates the current debate state and determines whether the process should continue. The debate stops when the Judge score satisfies the convergence criterion defined in Section 3.6. If the condition is not satisfied, the debate continues to the next round until reaching the maximum round controlled by the hallucination-window mechanism.

When the debate terminates, the system outputs a final hallucination judgment together with an evidence-grounded report summarizing the key Pro and Con arguments. The complete mathematical formulation of the debate state, agent perspective function, team consolidation, and stopping process is provided in Appendix D.3.

3.6 Unified Quality Assessment and Convergence Judgment

The Judge Agent is responsible for assessing debate quality and determining whether the debate should continue. Since IMPD-MACD targets black-box LLMs, the Judge Agent relies only on agents' textual outputs, retrieved evidence, and debate-history summaries, without accessing internal model states.

At each round, the Judge Agent evaluates the Pro and Con conclusions from four dimensions: interaction quality, factual accuracy, logical coherence, and convergence trend. Interaction quality reflects whether opposing agents provide relevant and informative arguments; factual accuracy checks whether key claims are supported by the retrieved external knowledge; logical coherence examines whether the reasoning is complete and free from irrelevant or contradictory content; and convergence trend measures whether the debate has become stable across recent rounds.

The final Judge score is computed as:

$$J_r = \left(\sum_{i=1}^{4} w_i s_i \right) \xi_r, \tag{2}$$

where s_i denotes the score of the i-th evaluation dimension, w_i is its weight, and ξ_r is the Judge Agent's confidence based on the internal consistency of the evaluation. The debate converges when:

$$J_r \leq \epsilon \text{and } \mathrm{Var}\left(\{J_{r'}\}_{r'=\max(1,r-\ell)}^{r} \right) \leq \sigma_{\text{thresh}} \tag{3}$$

where ϵ is the convergence threshold, ℓ is the historical scoring window length, and σ_{thresh} is the stability threshold. If the condition is not satisfied, the debate proceeds to the next round. The detailed scoring formulation, including semantic relevance, information value, temporal efficiency, and prompt-based Judge implementation, is provided in Appendix D.4.

Since this paper focuses on hallucination detection, we treat the post-detection Refiner Agent as a supplementary component and describe it in Appendix E.

4 Experiments

We evaluate IMPD-MACD on Semi-Open-HaluQA and HaluEval for hallucination detection in semi-open Q&A. For each question–answer pair, the model predicts whether the answer contains hallucinations, and the prediction is compared with the ground-truth binary label. We report Accuracy, Precision, Recall, and F1 score as the main evaluation metrics. In addition, we further evaluated the IMPD-MACD model's performance in correcting hallucinatory statements using the same set of metrics.

4.1 Experimental Setup

We evaluate IMPD-MACD on two hallucination detection benchmarks: Semi-Open-HaluQA and HaluEval (Li et al., 2023). Semi-Open-HaluQA is our constructed semi-open Q&A dataset, where each answer is annotated with a binary hallucination label. For evaluation, we sample 200 instances from Semi-Open-HaluQA and 200 instances from HaluEval. More details about dataset construction, sampling, and annotation are provided in Appendix F.

We compare IMPD-MACD with representative single-agent and multi-agent baselines, including CoT (Wei et al., 2022), Self-Reflection (Shinn et al., 2024), SelfCheckGPT (Manakul et al., 2023), MAD (Du et al., 2023), MADR (Kim et al., 2024), and MC-MAD (Sun et al., 2025). Unless otherwise specified, all methods use GPT-4o-Mini (OpenAI et al., 2024) as the backbone model for a fair comparison. Since different debate-based baselines follow different interaction protocols, we use their default settings and adapt MAD to the semi-open Q&A setting; implementation details are provided in Appendix I. We further evaluate cross-backbone robustness using GPT-3.5-Turbo (Brown et al., 2020), Gemini-1.5-Flash (Team et al., 2024), and DeepSeek-V3 (DeepSeek-AI et al., 2025).

4.2 Performance Comparison

Table 1 reports the performance of all compared methods on Semi-Open-HaluQA and HaluEval. The best results on Semi-Open-HaluQA and HaluEval are highlighted in bold green and bold blue, respectively. Overall, IMPD-MACD achieves the best or near-best performance across most backbone models and datasets, demonstrating the effectiveness of multi-perspective debate with external knowledge and controlled interaction.

On Semi-Open-HaluQA, IMPD-MACD consistently outperforms both single-agent and multi-agent baselines across different foundation models. Compared with single-agent methods, it achieves an average F1 improvement of over 10%, indicating that semi-open hallucination detection benefits from role-specialized agents and stance-based verification. On HaluEval, IMPD-MACD also surpasses most baselines, except under the DeepSeek-V3 backbone where MC-MAD obtains a slightly higher F1 score. This exception suggests that the benefit of multi-round debate may depend on the backbone model and that excessive interaction can introduce additional noise.

Further analysis shows that CoT, Self-Reflection, and MADR often obtain high recall but relatively low accuracy on Semi-Open-HaluQA, suggesting a tendency to classify more answers as non-hallucinatory. In contrast, IMPD-MACD provides more balanced precision and recall, showing stronger control over both false positives and false negatives. Across different backbone models, our method also exhibits relatively stable performance, indicating that the proposed role specialization, external evidence integration, and Judge-controlled interaction improve robustness in complex hallucination detection scenarios.

Method	Metrics (Semi-Open-HaluQA)				Metrics (HaluEval)			
	Accuracy	Precision	Recall	F1_Score	Accuracy	Precision	Recall	F1_Score
Foundation Model: GPT-4o-Mini								
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	0.660
SelfCheckgpt	0.735	0.753	0.700	0.725	0.635	0.737	0.420	0.535
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
MC-MAD	0.785	0.828	0.720	0.770	0.765	**0.812**	0.690	0.746
IMPD-MACD (Ours)	**0.820**	**0.827**	0.810	**0.818**	**0.795**	0.798	**0.790**	**0.794**
Foundation Model: GPT-3.5-Turbo								
Single-Agent Method								
CoT	0.525	0.513	**0.970**	0.671	0.535	0.519	**0.940**	0.669
Self-Reflection	0.570	0.542	0.900	0.677	0.665	0.639	0.760	0.694
SelfCheckgpt	0.720	0.729	0.700	0.714	0.610	0.690	0.400	0.506
Multi-Agent Method								
MAD	0.720	0.712	0.740	0.726	0.625	0.662	0.510	0.576
MADR	0.680	0.653	0.770	0.704	0.615	0.597	0.710	0.648
MC-MAD	0.670	0.905	0.380	0.535	0.675	0.787	0.480	0.596
IMPD-MACD (Ours)	**0.780**	**0.818**	0.720	**0.766**	**0.790**	**0.809**	0.760	**0.784**
Foundation Model: Gemini-1.5-Flash								
Single-Agent Method								
CoT	0.555	0.531	**0.950**	0.681	0.535	0.523	0.800	0.632
Self-Reflection	0.655	0.622	0.790	0.696	0.570	0.606	0.400	0.482
SelfCheckgpt	0.655	0.691	0.560	0.619	0.580	0.625	0.400	0.488
Multi-Agent Method								
MAD	0.630	0.691	0.470	0.560	0.535	0.621	0.180	0.279
MADR	0.375	0.353	0.300	0.324	0.360	0.365	0.380	0.373
MC-MAD	0.775	0.910	0.610	0.731	0.725	**0.826**	0.570	0.675
IMPD-MACD (Ours)	**0.830**	**0.850**	0.790	**0.823**	**0.805**	0.785	**0.840**	**0.812**
Foundation Model: Deepseek-V3								
Single-Agent Method								
CoT	0.590	0.550	**0.990**	0.707	0.580	0.550	0.880	0.677
Self-Reflection	0.655	0.610	0.860	0.714	0.685	0.703	0.640	0.670
SelfCheckgpt	0.750	0.719	0.820	0.766	0.665	0.726	0.530	0.613
Multi-Agent Method								
MAD	0.730	0.735	0.720	0.727	0.685	0.761	0.540	0.632
MADR	0.450	0.473	0.860	0.610	0.520	0.511	0.950	0.664
MC-MAD	0.805	0.851	0.740	0.791	**0.785**	0.782	**0.790**	**0.786**
IMPD-MACD (Ours)	**0.825**	**0.828**	0.820	**0.824**	0.755	**0.787**	0.700	0.741

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

4.3 Ablation Study

To examine the contribution of each component, we remove the Information-Gathering Agent (w/o EK), Logic Agent (w/o Logic), Factual Agent (w/o Fact), Comprehensive Agent (w/o Comprehensive), and Context Consistency Agent (w/o Consistency), respectively. Table 2 reports the results on both datasets.

The full IMPD-MACD consistently achieves the best performance, showing that all components contribute to hallucination detection. Removing the Information-Gathering Agent causes a clear performance drop, especially on HaluEval, indicating the importance of external evidence for factual verification. Removing the Logic Agent also substantially degrades performance, suggesting that reasoning verification is critical for detecting non-factual hallucinations. The Factual and Context Consistency Agents further improve robustness by checking evidence support and question-answer alignment, while the Comprehensive Agent provides useful team-level aggregation. Although removing the Comprehensive Agent leads to a smaller drop than removing other agents, its degradation on HaluEval shows that team-level integration still improves final judgment reliability. Overall, the ablation results indicate that IMPD-MACD benefits from the combination of evidence grounding, specialized verification, and coordinated aggregation rather than from any single component alone. Detailed

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. (Foundation Model: GPT-4o-Mini)

analysis of each ablation variant is provided in Appendix G.

4.4 Hyper-Parameter Analysis

We analyze two key factors that affect IMPD-MACD: the agent weight distribution and the number of debate rounds. For agent weights, we vary the variance among role weights w_i and observe that balanced weights generally lead to better performance. This suggests that no single agent should dominate the team-level decision when the number of agents is limited. Detailed weight-sensitivity results are provided in Appendix H.

More importantly, we evaluate how the number of debate rounds affects hallucination detection. As shown in Figures 2 and 3, performance first improves and then declines as the number of rounds increases. This trend supports our hallucination-window hypothesis: moderate debate improves evidence coverage and resolves disagreements, while excessive interaction introduces redundant or misleading context and degrades reliability. The optimal round number also differs across datasets, indicating that debate length should be adjusted according to task complexity.

4.5 Evaluation of Hallucination Correction

We further evaluate the optional Refiner Agent for post-detection answer correction. The Refiner performs a two-round correction process after IMPD-MACD detects hallucinations, and the corrected responses are evaluated using post-correction accuracy and an average quality score from 0 to 10. All experiments use GPT-4o-Mini as the backbone model, and the automatically corrected answers are additionally checked by human annotators for correctness.

Table 3 shows that the correction mechanism achieves high post-correction performance on

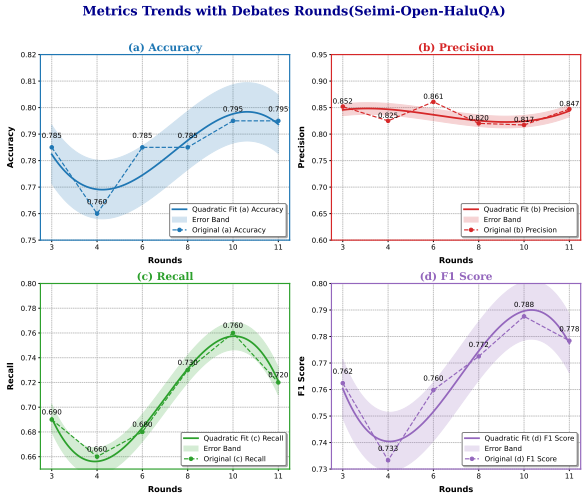

Figure 2: Evaluation metrics vs. debate rounds on Semi-Open-HaluQA.

Dataset	Acc.	Prec.	Rec.	F1	Avg. Quality
Semi-Open-HaluQA	0.882	1.000	0.882	0.938	9.376
HaluEval	0.807	1.000	0.807	0.893	8.807

Table 3: Correction performance and average quality score after response correction using IMPD-MACD.

hallucination-only samples, with F1 scores of 0.938 on Semi-Open-HaluQA and 0.893 on HaluEval. The average quality scores are also high on both datasets, indicating that the corrected responses are generally reliable.

Dataset	Acc.	Prec.	Rec.	F1
Semi-Open-HaluQA	0.842	0.911	0.899	0.905
HaluEval	0.875	0.911	0.939	0.925

Table 4: Overall post-correction performance of IMPD-MACD on samples with an initial hallucination rate of 50%.

Table 4 further shows that IMPD-MACD improves the overall reliability of samples with an initial hallucination rate of 50%, achieving over

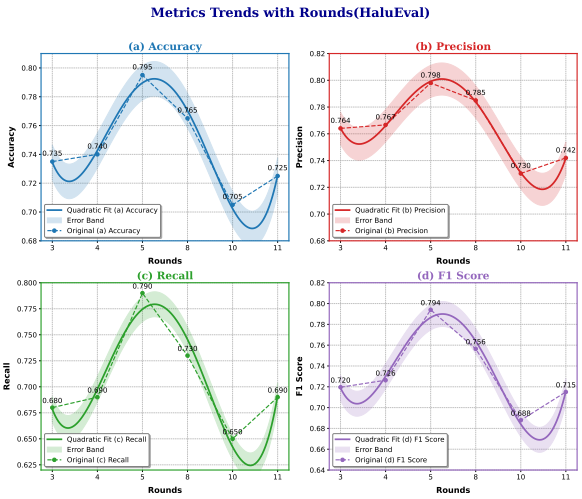

Figure 3: Evaluation metrics vs. debate rounds on HaluEval.

0.84 accuracy and over 0.90 F1 on both datasets. These results suggest that the Refiner Agent can serve as a useful supplementary component after hallucination detection.

4.6 Additional Cross-Model Robustness Evaluation

We further evaluate the robustness of IMPD-MACD across additional foundation models while keeping the framework unchanged. All experiments use the same evaluation data as the ablation study.

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
Deepseek-V3	0.825	0.828	0.820	0.824
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 5: IMPD-MACD Performance Under Different Foundation Agent Models

As shown in Table 5, IMPD-MACD maintains stable performance across different foundation models, with F1 scores ranging from 0.766 to 0.856. This result suggests that the proposed role-specialized debate structure is not tightly coupled with a specific backbone model and can be transferred to different LLM environments. Meanwhile, the performance differences across models indicate that the underlying model quality still affects role-specific reasoning, evidence interpretation, and final judgment. Overall, these results further support the cross-model generalization ability of IMPD-MACD.

4.7 Model Efficiency Evaluation

We further evaluate model efficiency in terms of token consumption and inference time. All methods are tested with GPT-4o-Mini on the same 200 Semi-Open-HaluQA samples, and multi-agent methods are executed in parallel whenever possible.

Method	Avg token (k token/min)	Avg running time (s)	Total token (k)
CoT	23.322	64.310	1499.838
Self-Reflection	115.663	510.460	59041.335
Self-Checkgpt	15.931	1618.610	25786.076
MAD	290.338	543.870	157906.128
MADR	106.700	638.500	68127.950
MC-MAD	15.304	3886.770	59483.128
IMPD-MACD	284.380	632.810	179958.508

Table 6: Comparison of token consumption and efficiency for different approaches, evaluated with GPT-4o-Mini.

As shown in Table 6, multi-agent approaches generally incur higher token consumption compared to single-agent methods. Notably, although MC-MAD demonstrates lower token consumption than MAD, its inference speed is significantly reduced due to inherent structural limitations. While IMPD-MACD exhibits the highest token consumption among all models, its inference speed remains comparable to other multi-agent frameworks. Overall, IMPD-MACD strikes a favorable balance between token usage, computational efficiency, and accuracy, thereby offering a practical advantage over existing baseline models.

5 Conclusion

This paper presents IMPD-MACD, a multi-perspective and diverse-stance collaborative debate framework for black-box hallucination detection in semi-open Q&A. By combining external knowledge retrieval, role-specialized Pro/Con agents, hierarchical visibility, sliding summarization, and Judge-controlled early stopping, IMPD-MACD improves the detection of factual, logical, and contextual hallucinations while reducing debate-induced noise. We further introduce a hallucination-window perspective to explain why debate depth should be controlled. Experiments on Semi-Open-HaluQA and HaluEval show that IMPD-MACD outperforms strong single-agent and multi-agent baselines across most settings, and ablation results confirm the importance of external knowledge and specialized agents. We also introduce Semi-Open-HaluQA as a new benchmark for evaluating hallucination detection in semi-open question answering.

Limitations

This work has several limitations. First, IMPD-MACD relies on multiple LLM agents, external retrieval, and multi-round interaction, so it requires more token consumption and inference time than single-agent detectors. Second, its effectiveness depends on the quality of retrieved external knowledge; incomplete, outdated, or noisy evidence may mislead the debating agents. Finally, the hallucination-window model is used as an interaction-control intuition rather than a complete causal theory of hallucination generation, and the optional Refiner Agent is treated only as a supplementary component rather than the main focus of this paper.

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, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, and 1 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

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.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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.

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.

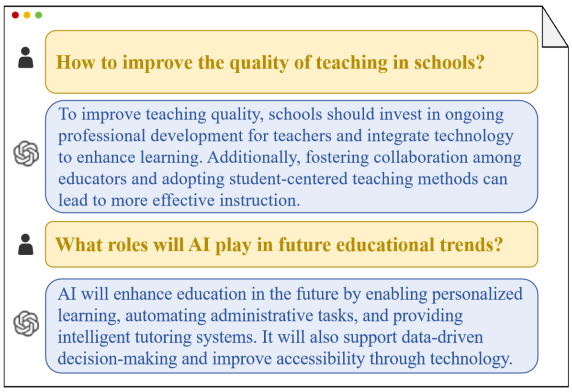

Figure 4: An example of open questions

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. Figure 5 below is a simple example.

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.

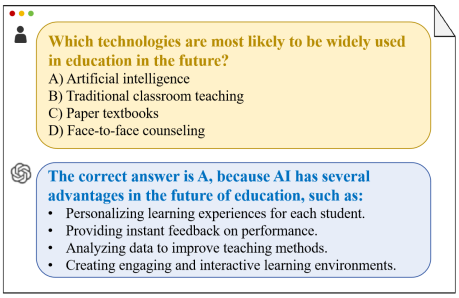

Figure 5: An example of closed-ended questions

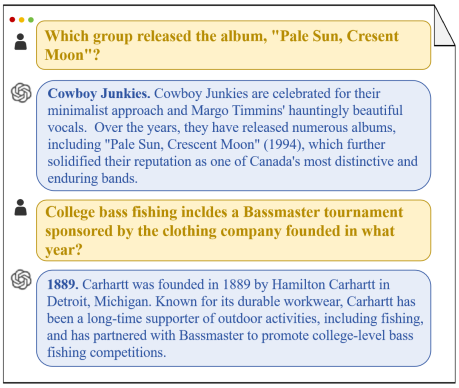

Figure 6: An example of semi-open questions

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:

$$p_\theta(y|Q) = \sum_{d_j \in D} p_r(d_j|Q) \times p_\theta(y|Q, d_j) \quad (4)$$

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.

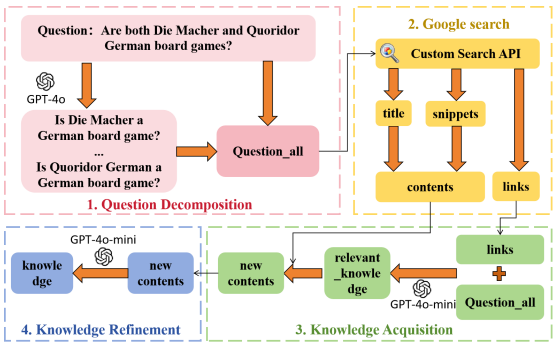

Figure 7: Information-Gathering Agent flow frame diagram

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 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.

C.1 Agent Grouping and Debate Structure

In IMPD-MACD, all debating agents are organized into two opposing teams. The Pro Team argues that "no hallucination exists in the original answer", while the Con Team argues that "the original answer may contain hallucinations". This stance-level opposition encourages the system to examine both supporting evidence and counter-evidence.

Within each team, four specialized agents are deployed: the Logic Agent, the Context Consistency Agent, the Factual Agent, and the Comprehensive Agent. The Logic Agent focuses on reasoning validity and logical consistency; the Context

Consistency Agent checks whether the answer is aligned with the question and internally coherent; the Factual Agent verifies factual claims using both parameterized knowledge and retrieved external evidence; and the Comprehensive Agent integrates the outputs of the other three agents into a team-level conclusion. The two teams then submit their conclusions to the Judge Agent, which evaluates the debate state and decides whether the process should continue or converge.

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.

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)

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.

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

Judge Agent Prompt

```
"You are the **Judge Agent** in the IMPD-MACD framework.\n"
"Responsibilities:\n"
"1. Read both teams' consolidated conclusions, external knowledge, and interaction-quality score, then perform a four-dimension evaluation;\n"
"2. Compute the round score J_r and decide whether the debate continues or converges;\n"
"3. Produce a concise summary (≤{L_sum} characters) for cross-agent visibility in the next round;\n"
"4. Return results strictly in the specified JSON schema—no extra text.\n\n"
"[Input schema (sent as the user message)]\n"
"{\n"
'  "Q": "<original user question>",\n'
'  "A0": "<initial answer>",\n'
'  "C_pro": "<Pro team consolidated conclusion>",\n'
'  "C_con": "<Con team consolidated conclusion>",\n'
'  "K_ext": "<external knowledge from Information-Gathering Agent>",\n'
'  "Q_r": <float in [0,1]>,\n'
'  "J_history": [<J_1>, ..., <J_{r-1}>],\n'
'  "r": <current round>,\n'
'  "L_sum": <max characters>,\n'
'  "epsilon": 0.30,\n'
'  "sigma_thresh": 0.0025,\n'
'  "w": [0.25,0.25,0.25,0.25]\n'
"}\n\n"
```

```
"[Evaluation dimensions] S1 Interaction Quality (Q_r); S2 Factual Accuracy;
                         S3 Logical Integrity / Irrelevance Penalty; S4 Convergence Trend.\n"

"Composite score: J_r = (w1·S1 + w2·S2 + w3·S3 + w4·S4) × confidence ξ_r (two decimals).\n"
"Convergence rule: if J_r ≤ epsilon AND Var(J_history∪{J_r}) ≤ sigma_thresh → decision=\"converged\"; else \"continue\".\n\n"

"[Summary requirements] ≤L_sum characters; cover key contention points & hallucination risk cues.
                        If converged, append "Final verdict: hallucination present/absent".\n\n"
"[Sole valid output format]\n"
"{\n"
'  "evaluation": {\n'
'    "S1": 0.00,\n'
'    "S2": 0.00,\n'
'    "S3": 0.00,\n'
'    "S4": 0.00,\n'
'    "confidence": 0.00,\n'
'    "J_r": 0.00\n'
'  },\n'
'  "summary": "",\n'
'  "decision": "continue"\n'
"}"
```

Figure 10: Prompt Design Workflow for the Judge Agent in IMPD-MACD

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.

Pro/Con stance-specific prompting. For each specialized role, we instantiate two stance-specific versions: a Pro-side agent and a Con-side agent. Although both agents share the same functional role, their objectives are different. The Pro-side agent is instructed to identify supported claims, coherent reasoning paths, and evidence that justifies the original answer. In contrast, the Con-side agent is instructed to search for unsupported claims, factual conflicts, reasoning gaps, and contextual inconsistencies. This design prevents the debate from degenerating into repeated self-confirmation and encourages agents to examine the answer from opposing perspectives. Figures 8 and 9 illustrate the Pro-team and Con-team prompt designs using the Logic Agent as an example.

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 optimize 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 Agent Inputs and Outputs

At each debate round, every specialized agent receives the original question, the initial answer, the retrieved external knowledge, its own team stance, the previous-round Judge summary, and the visible opponent information determined by the hierarchical visibility mechanism. Each agent outputs a structured analysis containing: (1) key claims or

reasoning points under its assigned role, (2) supporting or conflicting evidence, and (3) a concise role-level conclusion.

The Comprehensive Agent receives the outputs of the three specialized agents within the same team and produces a team-level conclusion. The Judge Agent receives the Pro conclusion, the Con conclusion, the retrieved knowledge, the previous debate history summary, and the interaction-quality signal. It outputs a structured judgment containing the current-round evaluation, a summary for the next round, and a decision indicating whether the debate should continue or converge.

D Detailed Method Derivations

D.1 Dual-Mechanism Poisson Process Modeling for Hallucination Probability

This section presents a mathematical modeling framework for multi-agent debate to quantify the probability of hallucinations occurring at different rounds. The model is grounded in the computational complexity of Transformer-based language models and the Gibbs-Boltzmann sampling mechanism, combined with the independent increment assumption of multi-agent reasoning, providing a theoretical foundation for optimizing subsequent multi-agent debates.

First, let the debate round be $r \in \{0, 1, \dots\}$, and denote the total context length visible to each agent at round r as n_r. Since the reasoning paths of individual agents in the same round are mutually independent conditioned on the current context, the "new hallucination triggering" can be modeled as an incremental Poisson process with instantaneous intensity ("unit time triggering rate") denoted as λ_r. From the perspective of Gibbs sampling at fixed temperature T, erroneous events (i.e., hallucinations) in LLM output sequences can be abstracted as an arrival process with exponentially distributed waiting times; when $\lambda_r \ll 1$, single-round hallucination occurrence can be further approximated as a Bernoulli trial with occurrence probability λ_r.

Considering two opposing driving forces that jointly influence the hallucination rate: first, the suppression effect brought by complementary knowledge injection in the early stages of debate; second, the accumulation of long-range interaction noise caused by continuous context expansion. Combined with the $\mathcal{O}(n^2)$ complexity of Transformer attention, we decompose λ_r into a superposition of exponential decay and quadratic growth terms:

$$\lambda_r = \underbrace{\lambda_0\, e^{-\beta r}}_{\text{Knowledge Suppression}} + \underbrace{\kappa\, n_r^2}_{\text{Complexity Noise}} \tag{5}$$

where $\lambda_0 > 0$ is the initial hallucination propensity, $\beta > 0$ measures the decay rate of effective knowledge; $\kappa > 0$ is the context noise coefficient, reflecting the mismatch between channel numbers and attention head combinations in long sequences. The context length adopts a linear approximation $n_r = n_0 + \mu r\ (\mu > 0)$. The cumulative intensity of the Poisson process $\Lambda_r = \sum_{i=1}^{r} \lambda_i$ determines the cumulative probability of "at least one hallucination occurring by round r":

$$H(r) = 1 - \exp\!\big(-\Lambda_r\big) \tag{6}$$

When λ_r approximates the marginal hallucination probability $\Delta H(r)$, taking the derivative of Eq. (5) yields the extremal condition:

$$\frac{d\lambda_r}{dr} = 0 \implies -\beta\lambda_0 e^{-\beta r} + 2\kappa\mu n_r = 0 \tag{7}$$

which gives the inflection point round:

$$r^\star = \frac{1}{2\mu}\big(\beta n_0 - \mu\big) \tag{8}$$

This value is positive only when $\beta n_0 > \mu$, indicating that within the experimental interval, there will first be a hallucination decline phase dominated by knowledge advantage, followed by a turn to increase when context noise overtakes. If the phenomenon of "accuracy first rising then falling" can be observed empirically, then this round is $r^\star$, which further substantiates the effectiveness of using this as a multi-agent interaction optimization criterion.

D.2 Multi-Round Interactions Among Agents

To formally describe the interaction process among agents in multi-round debates, we first define the agent set participating in the debate system:

$$\mathcal{A} = \{a_1, a_2, \dots, a_A\} \tag{9}$$

where $A = 9$. These agents are systematically divided into two opposing teams with no overlap:

$$\mathcal{A}_{\text{Pro}} \cup \mathcal{A}_{\text{Con}} = \mathcal{A}, \quad \mathcal{A}_{\text{Pro}} \cap \mathcal{A}_{\text{Con}} = \varnothing \tag{10}$$

Based on the dual-mechanism Poisson model analysis in Section 3.3, we can derive that the hallucination probability in single-round interactions is directly related to the irrelevant information quantity I_{irr}:

$$P(\text{Hallucination}) = g(I_{\text{irr}}) \tag{11}$$

This relationship indicates that controlling irrelevant information quantity is key to improving debate quality. Under the theoretical framework derived from the dual-mechanism Poisson model, the optimal point $r^\star$ of debate rounds shifts rapidly leftward with the quadratic growth of context length n_r. If all agents are allowed to fully accumulate history in each round, the corresponding $n_r \to \infty$ will cause $r^\star$ to collapse to 1 in practice, completely eliminating the performance window available between early suppression-late noise". More importantly, due to the inherent context window limitations of LLMs (MaxLen), the context $\mathcal{C}_r(a)$ received by each agent a at round r must satisfy the constraint:

$$\|\mathcal{C}_r(a)\| \leq \mathsf{MaxLen} \tag{12}$$

To simultaneously address the theoretical inflection point leftward shift and practical context limitation problems, we design a **hierarchical visibility constraint and sliding summarization** mechanism. Specifically, both Pro and Con teams deploy four functional agents {Factual, Logic, Context Consistency, Comprehensive}, and introduce a Judge Agent independent of both sides. The Judge Agent undertakes dual responsibilities in the system: first, serving as an information filter to implement cross-agent summary transmission; second, evaluating debate quality and controlling the debate convergence process (detailed mechanisms will be elaborated in the next section).

Let $\mathbf{y}_{r,a}$ denote the token sequence output by agent a at round r. We define cross-agent context channels using a sparse visibility matrix $G \in \{0,1\}^{A\times A}$ (where $A = 9$, including 8 sub-agents from both sides and 1 Judge Agent). Specifically, the visibility rules are designed as follows:

- **Same-role cross-stance visibility**: If agents a and b belong to different stances but serve the same role (e.g., Pro Factual Agent and Con Factual Agent), then $G_{ab} = 1$, and agent a can read the complete output $\mathbf{y}_{r,b}$ of agent b.
- **Different-role cross-stance summarization**: If agents a and b belong to different stances and serve different roles (e.g., Pro Factual Agent and Con Logic Agent), then $G_{ab} = 0$, and agent a can only obtain a Judge-provided summary $\text{sum}(\mathbf{y}_{r,b})$ with length $\leq L_{\text{sum}}$ characters.

This design enables each agent to obtain: (1) the complete output from the previous round of the opposing same-role agent, for in-depth professional confrontation; (2) Judge summaries of opposing different-role agents, maintaining overall debate situation awareness while controlling information redundancy.

The Judge Agent's summarization function can be expressed as:

$$\text{JudgeSummarize}(\mathbf{y}_{r,\cdot}, L_{\text{sum}}) \to \mathbf{s}_r \tag{13}$$

where $\mathbf{s}_r$ is the refined summary for round r, satisfying $|\mathbf{s}_r| \leq L_{\text{sum}}$. The Judge Agent only retains the complete debate traces of the most recent two rounds $\{\mathbf{y}_{r-1,\cdot}, \mathbf{y}_{r,\cdot}\}$ as the basis for judgment, and broadcasts the refined summary $\mathbf{s}_r$ at the end of each round.

Under this mechanism, the effective context length for a single agent at round $r+1$ can be written as:

$$n_{r+1,a} = n_0 + \sum_b G_{ab}\,|\mathbf{y}_{r,b}| + \sum_b (1-G_{ab})\,L_{\text{sum}}, \tag{14}$$

and the global average context length becomes:

$$\bar{n}_{r+1} = n_0 + \rho\,\bar{y}_r + (1-\rho)\,L_{\text{sum}}, \quad \rho = \frac{1}{A^2}\sum_{a,b} G_{ab}, \tag{15}$$

where $\bar{y}_r$ is the average output length of the previous round. It can be seen that $\rho < 1$ will convert the original quadratic growth term κn_r^2 to $\kappa(\bar{n}_r)^2$ with the growth coefficient adjusted by ρ. When $\rho \ll 1$ and L_{sum} is controlled, the theoretical inflection point shifts to:

$$\begin{aligned} r^\star_{\text{eff}} &= \frac{1}{2\mu_{\text{eff}}}\big(\beta n_0 - \mu_{\text{eff}}\big), \\ \mu_{\text{eff}} &= \rho\mu + \frac{(1-\rho)L_{\text{sum}}}{\Delta r}. \end{aligned} \tag{16}$$

Where Δr denotes the slope of the single-round average output length, lowering the mutual-visibility ratio ρ or shortening Judge summaries

L_{sum} can markedly postpone the inflection point at which the hallucination rate begins to rise again. We formalize this inflection point as $r^{\star}_{\text{eff}}$, defined as the maximum debate round the system should permit and, consequently, the upper bound of the hallucination window. By delaying $r^{\star}_{\text{eff}}$, the framework allows knowledge-injection debates to proceed sufficiently without succumbing to the rapid degradation caused by long-context noise.

In this architecture, each team's Comprehensive Agent aggregates the viewpoints of its four specialist agents and produces a structured summary for the Judge. The Judge Agent then decides, on a per-round basis, whether to distribute full content or compressed summaries to the other agents, using the two teams' outputs plus their Comprehensive summaries as input. This centralized information-distribution mechanism ensures that every agent's context $\mathcal{C}_r(a)$ respects the length constraint in Eq. (12) and that the hallucination probability $P(\text{Hallucination})_r$ is minimized by controlling the quantity of irrelevant information I_{irr}. Taken together, these controls uphold high-quality multi-round interactions while guaranteeing that the entire debate converges by—or, if possible, before—the boundary $r^{\star}_{\text{eff}}$.

D.3 Multi-Agent Debate Process

In the formal debate phase, we mathematically formulate the multi-round interaction process. Define the debate state at round r as:

$$S_r = \{Q, A_0, \mathcal{K}_{\text{ext}}, \mathcal{K}_{\text{param}}, \mathcal{H}_{r-1}\} \tag{17}$$

where Q represents the input question, A_0 is the initial answer, $\mathcal{K}_{\text{ext}}$ and $\mathcal{K}_{\text{param}}$ represent external knowledge and parameterized knowledge respectively, and $\mathcal{H}_{r-1}$ is the debate history from the previous round.

Each agent's perspective at round r can be formalized through a perspective function:

$$V_r^a = f_{\text{perspective}}(S_r, \mathcal{K}_{\text{team},a}) \tag{18}$$

where $\mathcal{K}_{\text{team},a}$ represents the specific knowledge of the team to which agent a belongs.

The conclusions of both Pro and Con teams are obtained through the integration function of the Comprehensive Agent:

$$C_r^{\text{team}} = \text{Consolidate}(\{V_r^i\}_{i=1}^4, \mathbf{s}_{r-1}) \tag{19}$$

where $\{V_r^i\}_{i=1}^4$ are the viewpoints of the four professional agents on the same side, and $\mathbf{s}_{r-1}$ is the summary provided by the Judge Agent from the previous round.

The convergence judgment of the debate process is determined by the Judge Agent's scoring. The debate convergence condition is:

$$J_r \leq \epsilon \quad \text{and} \quad \text{Var}\big(\{J_{r'}\}_{r'=\max(1,r-\ell)}^{r}\big) \leq \sigma_{\text{thresh}} \tag{20}$$

where ϵ is the preset convergence threshold, J_r is the Judge Agent's score at round r (detailed scoring mechanism described in the next section), ℓ is the historical scoring window length, and σ_{thresh} is the variance threshold. The first condition ensures that the score meets quality requirements, and the second condition ensures scoring stability, avoiding erroneous convergence due to occasional low scores. If convergence conditions are not met, the debate continues to round $r+1$:

$$J_r > \epsilon \quad \text{or} \quad \text{Var}\big(\{J_{r'}\}_{r'=\max(1,r-\ell)}^{r}\big) > \sigma_{\text{thresh}} \tag{21}$$

The debate process continues until reaching the maximum number of rounds $r_{\text{max}} = r^{\star}_{\text{eff}}$ or satisfying convergence conditions. At the end of the debate (round $r = R$), a new Refiner Agent is introduced to generate the final report:

$$\mathcal{R} = f_{\text{refine}}\big(\{C_r^{\text{Pro}}, C_r^{\text{Con}}, J_r, Q_r\}_{r=1}^{R}\big) \tag{22}$$

where Q_r is the interaction quality assessment result for round r.

D.4 Unified Quality Assessment and Convergence Judgment

As the core supporting mechanism for the debate process in Section 3.5, the quality assessment system needs to provide the Judge Agent with J_r scores after each round of debate to decide whether the debate should continue ($J_r > \epsilon$) or converge ($J_r \leq \epsilon$). Due to the use of black-box LLM models, the Judge Agent cannot access internal model states, so the scoring system is designed entirely based on agents' text outputs.

The interaction quality Q_r among agents serves as an important foundation for Judge scoring, measured by aggregating the interaction performance of individual agents:

$$Q_r = \sum_{i=1}^{4} w_i \cdot q_r^i(a_i, \text{Opp}(a_i)) \tag{23}$$

where w_i is the agent weight, and q_r^i measures the interaction quality between agent a_i and its opponent $\text{Opp}(a_i)$. For agent a_i, based on its visible historical context $\mathcal{C}_r(a_i)$ (as defined in Eq. (12)), interaction quality assessment includes three dimensions q_{sem}^i, q_{info}^i, and q_{time}^i:

Semantic Relevance measures the degree of viewpoint opposition through cosine similarity:

$$q_{\text{sem}}^i = \text{softmax}\big(\text{Cos}(\mathbf{y}_{r,a_i}, \mathbf{y}_{r,\text{Opp}(a_i)})\big) \tag{24}$$

where $\mathbf{y}_{r,a_i}$ and $\mathbf{y}_{r,\text{Opp}(a_i)}$ represent the output token sequences of agent a_i and its opponent at round r, respectively.

Information Value is assessed based on conditional entropy evaluation of new content, incorporating the irrelevant information theory from Section 3.4:

$$q_{\text{info}}^i = \frac{\mathcal{H}_e(\mathbf{y}_{r,a_i} \mid S_r)}{\mathcal{H}_{\text{e_max}}^r} \tag{25}$$

where S_r is the debate state at round r as defined in Eq. (17), and $\mathcal{H}_{\text{e_max}}^r$ is the maximum information entropy for round r. The conditional entropy is calculated as:

$$\mathcal{H}_e(\mathbf{y}_{r,a_i} \mid S_r) = -\sum_{t_j} P(t_j \mid S_r) \log_2 P(t_j \mid S_r) \tag{26}$$

where t_j represents tokens in the agent's output, and the conditional probability calculation includes Laplace smoothing (γ is the smoothing parameter):

$$P(t_j \mid S_r) = \frac{\text{count}(t_j, \mathcal{Y}_{r,i}) + \gamma}{\sum_{t_k \in V} \text{count}(t_k, \mathcal{Y}_{r,i}) + |V|\gamma}, \quad \mathcal{Y}_{r,i} = \left\{\mathbf{y}_{r,a_i}, \mathbf{y}_{r,\text{Opp}(a_i)}\right\}. \tag{27}$$

Temporal Efficiency is calculated using an exponential decay function:

$$q_{\text{time}}^i = e^{-\alpha(r - r_{\text{start}}^i)}, \qquad \alpha > 0 \tag{28}$$

where α is the temporal decay coefficient, and r_{start}^i is the round at which agent a_i started participating in the debate.

The comprehensive interaction quality for agent a_i at round r is:

$$q_r^i = q_{\text{sem}}^i \cdot q_{\text{info}}^i \cdot q_{\text{time}}^i \tag{29}$$

This assessment mechanism possesses the following theoretical properties: consistency ($\lim_{|\mathcal{H}_{r-1}| \to \infty} \text{Var}(q_r^i) \to 0$), boundedness ($q_r^i \in [0, 1]$), and temporal monotonicity (when $r < r^\star$, q_{time}^i decreases with rounds). Through integrating the Poisson process model from Section 3.3 and the hierarchical visibility mechanism from Section 3.4, precise quantification of debate interaction quality consistent with the overall theoretical framework is achieved.

The Judge Agent conducts four-dimensional evaluation based on structured prompts, with inputs including conclusions from both Pro and Con teams C_r^{Pro}, C_r^{Con}, and interaction quality Q_r. The scoring dimensions include: interaction quality dimension $stage_1$ directly standardizes agent interaction evaluation results; factual accuracy dimension $stage_2$ enables the Judge Agent to verify consistency between key factual claims and external knowledge $\mathcal{K}_{\text{ext}}$ through prompt instructions; logical coherence dimension $stage_3$ combines theories from Section 3.4, guiding the Judge Agent through prompts to identify and penalize irrelevant information I_{irr} while evaluating argument structure completeness; convergence trend dimension $stage_4$ judges debate convergence trends based on historical scoring sequences $\{J_{r'}\}_{r'=1}^{r-1}$, balancing current round quality with historical stability. The Judge Agent calculates confidence ξ_r through built-in consistency checking mechanisms, and the final score is synthesized as:

$$J_r = \left(\sum_{i=1}^{4} w_i s_i\right) \xi_r \tag{30}$$

where $\xi_r \in [0, 1]$ is the Judge Agent's confidence assessment at round r, obtained by analyzing internal consistency among dimensional scores. To ensure no bias exists among agents, the agent weights w_i and Judge Agent scoring weights adopt the same weight configuration, satisfying $\sum_{i=1}^{4} w_i = 1$.

Convergence judgment directly serves the debate process control in Section 3.5, adopting dual conditions to ensure debate quality and stability:

$$J_r \leq \epsilon \quad \text{and} \quad \text{Var}(\{J_{r'}\}_{r'=\max(1,r-\ell)}^{r}) \leq \sigma_{\text{thresh}} \tag{31}$$

where ϵ is the preset convergence threshold, ℓ is the historical scoring window length, and σ_{thresh} is the variance threshold. When convergence conditions are satisfied, the debate process transitions to the report generation phase described in Section 3.5; otherwise, it continues to the next round of debate while utilizing current scores J_r and interaction quality Q_r to adjust parameters ρ and L_{sum} in Section 3.4, dynamically optimizing information visibility to maintain the optimal debate round $r^{\star}_{\text{eff}}$ under the theoretical framework of Section 3.4. As a result, the entire debate system converges before the hallucination window $r^{\star}_{\text{eff}}$ arrives, ensuring output accuracy while saving computational resources through early termination.

The Judge Agent's prompt design fully utilizes external knowledge retrieval results from Section 3.1, implementing multi-dimensional evaluation through structured instruction templates: factual verification is completed through text matching with retrieved knowledge, logical evaluation is executed through argument completeness checklists, interaction quality directly references Q_r values, and convergence judgment is based on statistical analysis of scoring history. This prompt-based implementation approach avoids complex internal state calculations, ensuring operability in black-box model environments while maintaining consistency with the overall theoretical framework.

E Refiner Agent's Hallucination Correction Mechanism

As a supplementary component of the IMPD-MACD system, the Refiner Agent undertakes the crucial role of final review and refinement of hallucination detection content. Although the primary objective of this paper is hallucination detection, considering the demand for high-quality answers in practical application scenarios, we design a two-round iterative correction mechanism for the Refiner Agent to modify and improve original answers after detecting hallucinations.

Based on the debate convergence judgment from Sections 3.5 and 3.6, when the system determines the existence of hallucinations, the Refiner Agent initiates a two-round consecutive correction process. In the first round of correction, the Refiner Agent receives the final report $\mathcal{R}$ defined in Eq. (22) and the external knowledge $\mathcal{K}_{\text{ext}}$ from Section 3.1, generating a preliminary corrected version:

$$A^{(1)}_{\text{corr}} = f^{(1)}_{\text{refine}}(\mathcal{R}, \mathcal{K}_{\text{ext}}, A_0) \tag{32}$$

where A_0 is the original answer defined in Eq. (17), and $f^{(1)}_{\text{refine}}$ is the first-round correction function. The second round of correction further optimizes based on the first-round results and the system's internal parameterized knowledge $\mathcal{K}_{\text{param}}$:

$$A^{(2)}_{\text{corr}} = f^{(2)}_{\text{refine}}(A^{(1)}_{\text{corr}}, \mathcal{K}_{\text{param}}, \mathcal{K}_{\text{ext}}) \tag{33}$$

To quantify correction effectiveness, we design a dual knowledge source-based scoring system:

$$\begin{aligned} E_{\text{corr}} = \eta \cdot E_{\text{ext}} \left(A^{(2)}_{\text{corr}}, \mathcal{K}_{\text{ext}} \right) \\ + (1-\eta) \cdot E_{\text{param}} \left(A^{(2)}_{\text{corr}}, \mathcal{K}_{\text{param}} \right). \end{aligned} \tag{34}$$

where $E_{\text{ext}}(\cdot,\cdot)$ and $E_{\text{param}}(\cdot,\cdot)$ are scoring functions based on external knowledge and internal parameterized knowledge, respectively, $\eta \in [0,1]$ is the weight balancing parameter, and $E_{\text{corr}} \in [0,10]$, with higher scores indicating better correction quality.

This correction mechanism forms organic integration with the aforementioned hallucination detection process: it relies on quality assessment results from Eq. (23) and debate reports from Eq. (13) as inputs, fully utilizes external knowledge retrieval results from Information-Gathering Agent, and automatically triggers after the detection process confirms the existence of hallucinations, forming a complete closed loop of "detection→correction→verification". Detailed effectiveness evaluation results of the correction mechanism are shown in Table 3. It should be emphasized that this correction function serves as a supplementary component of the system and does not affect the achievement of this paper's core objective of hallucination detection.

F 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.

F.1 Dataset Sampling and Evaluation Split

Semi-Open-HaluQA contains 1,800 semi-open Q&A entries with binary hallucination labels. For the main evaluation, we first select 600 high-quality samples from the full dataset and then randomly sample 200 instances using random seed 42. Each instance contains a question, an initial answer, supporting knowledge, and a binary label indicating whether the answer is hallucination-free.

For HaluEval (Li et al., 2023), we follow its original structure, where each question is paired with both correct and hallucinatory answers. To construct a binary hallucination detection setting comparable to Semi-Open-HaluQA, we randomly select either the correct answer or the hallucinatory answer for each question using random seed 2025, and then assign the corresponding binary label. We sample 200 HaluEval instances for evaluation. In all experiments, label 0 denotes that the answer contains hallucination, while label 1 denotes that the answer is hallucination-free.

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

G Detailed Ablation Analysis

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 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 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.

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.

H Hyper-Parameter Sensitivity

In this study, multiple agents are introduced within each faction, requiring an evaluation of their relative contributions to the final decision. We assign each agent a weight w_i and vary the variance among these weights to assess how different weighting strategies affect performance:

$$\mathrm{Var}^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \bar{w})^2. \tag{35}$$

Figures 11 and 12 show the results under different variance settings. When the variance is relatively small, IMPD-MACD generally achieves better Accuracy, Recall, and F1 scores, indicating that balanced agent weights are preferable when each team contains only a limited number of agents.

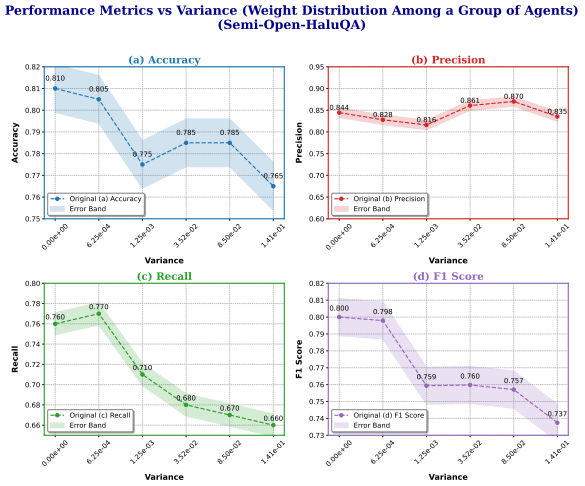

Figure 11: Influence of agent weight distribution on Semi-Open-HaluQA.

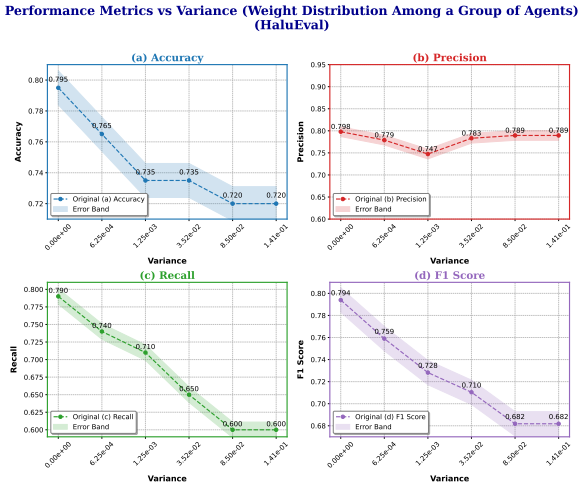

Figure 12: Influence of agent weight distribution on HaluEval.

I 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 ini-

tiate 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.