2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Enhancing Hallucination Detection in Large Language Models through a Dual-Position Debate Multi-Agent Framework

Qile He1 , Siting Le2

1 Guangdong Polytechnic Normal University, Guangdong 510450, China
heqile927@gmail.com
2 University of Wisconsin-Madison, Wisconsin 53706, America
alicia.st.le@gmail.com

Abstract. Large language models (LLMs) have demonstrated remarkable capabilities but are prone to generating factual inconsistencies, or "hallucinations." Addressing this challenge is crucial for the reliable deployment of LLMs. This paper introduces a novel Dual-Position Debate (DPD) framework designed to enhance the veracity of LLM-generated content and mitigate hallucinations. DPD simulates a human debate by organizing agents into affirmative and negative teams, each comprising information gatherers, rebutters, analysts, and a summarizer. These agents collaboratively construct arguments, critique opposing viewpoints, and synthesize their findings. Furthermore, multiple independent LLMs act as referees, evaluating the debate and rendering judgments to ensure fairness and encourage rigorous information scrutiny. Extensive experiments across question answering, summarization, and dialogue tasks demonstrate the efficacy of the DPD framework, outperforming existing baseline methods in reducing hallucinations.

Keywords: Large language models · Hallucination Detection · Multi-Agent Debate

1 Introduction

The rapid advancement of large language models (LLMs) has expanded cross-domain natural language processing applications31, but it also brings challenges such as high costs for parameter updates and reasoning limitations1339, leading to increased hallucinations in models like ChatGPT and GPT-42026. Hallucination detection14 is crucial. Research shows that evaluating output consistency and using post-processing methods11026 can help detect and correct hallucinations. Various strategies have been proposed across different model stages (pre-processing, alignment, and decoding). Nonetheless, existing methods still face the following issues: **(1) Intervention challenges:** Difficult to apply to parameter-agnostic, black-box large language models12. **(2) Misplaced focus:** Neglecting output accuracy verification.

To address these challenges, we propose a Dual-Position Debate (DPD) framework that simulates human debates. It divides agents into affirmative and negative sides, each with information gatherers, rebutters, and analysts, aiming to eliminate weak points that cause hallucinations. Powerful LLMs serve as referees to evaluate arguments, ensuring fairness and rigorous information scrutiny. This enhances reasoning and reduces hallucinations.

We conducted extensive experiments on three generation tasks: question answering, summarization, and dialogue. The results robustly support the efficacy of our proposed approach. Through thorough analysis and comparisons with existing methodologies, we demonstrate the superiority of our method. Our contributions can be summarized as follows:

We introduced a Dual-Position Debate (DPD) framework that simulates human debate.
This framework significantly enhances the capability to identify and reduce hallucinations in generated content.

Experimental results across the three generation tasks indicate that our proposed framework outperforms baseline methods.

2 Related Work

2.1 Hallucination Detection

Before the advent of large language models, hallucination detection had already become an important research topic in the field of natural language processing[20]. Previous studies primarily focused on identifying illusion phenomena across various tasks, such as summarization[12,9,22], dialogue systems[3], question answering[19], and machine translation[33]. The main objective of these methods was to uncover discrepancies between the generated content and the input data, as well as inconsistencies within the generated content itself. However, these methods were typically tailored to specific task models, lacking broad applicability. With the emergence of large language models, some research[8,15] began to approach the illusion detection task by directly prompting these models. In addition to task-specific methods, there are also some illusion detection mechanisms specifically designed for large language models. For instance, some approaches evaluate illusion detection by assessing the consistency among generated samples[21,36].

2.2 Multi-Agent Debate

Multi-Agent Debate (MAD) has emerged as a vital approach for enhancing logical reasoning within multi-agent collaboration. This framework, introduced by 19 fosters divergent thinking among Large Language Models (LLMs), guided by a mediator who oversees the discussion to direct agents toward a solution. Current research

emphasizes common-sense reasoning in real-world contexts, demonstrated by34, where debate methodologies effectively enhance LLM reasoning capabilities. Additionally, 7 examines how the number of agents and debate rounds impacts the accuracy of outcomes. 36simulated the academic peer review process, allowing agents to correct each other. Multiple agents can also improve decision-making and reasoning abilities through cooperative and adversarial debates[5,17,28].

3 Methodology

3.1 Grouping Strategy

In our framework, we establish pro and con camps that approach the issue from opposing perspectives to enhance hallucination detection. As shown in Figure 1, each camp comprises information gatherers, rebuttals, analysts, and summarizers. Information gatherers collect relevant data, rebuttals challenge and counterargue, analysts conduct in-depth examinations, and summarizers integrate views into a final judgment. Through multiple rounds of vigorous debate, diverse perspectives are thoroughly contested, enhancing the evaluation's comprehensiveness and depth. Ultimately, referees vote to determine the winning side. This structured multi-agent debate enables accurate and credible judgments, effectively identifying and addressing hallucinations in the text.

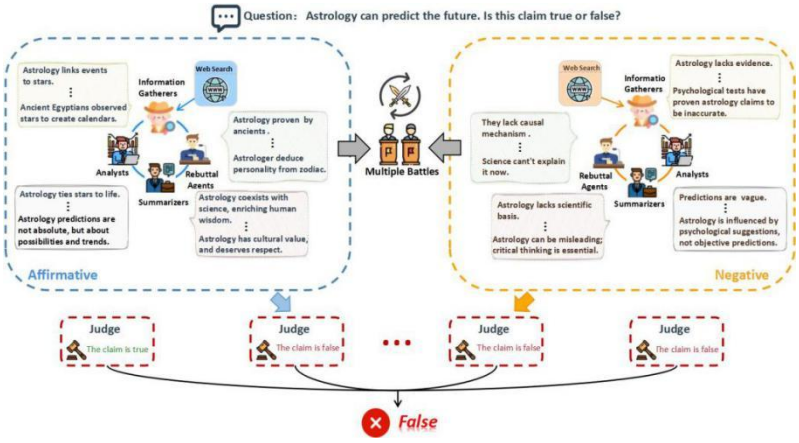

Fig. 1: An overview of DPD.

3.2 Roles of Various Agents

Information Gatherers Search for information related to one's own viewpoint and provide support to teammates.

Each camp will search for information that aligns with their own viewpoint at the beginning, and this search will occur only once at the outset (since, in a real debate scenario, it is not feasible to search for information while simultaneously debating).

The search results will be stored as document records[2] (in a database), and during each subsequent round of debate, the most relevant content will be selected from the documents based on the current context, in order to provide more information to other agents.

The content of the search includes: information that supports one's own viewpoint, and information that refutes the opponent's viewpoint.

The search method involves using the Google Search API and employing a knowledge base for Retrieval-Augmented Generation (RAG)[14].

Analysts This agent analyzes the original input information from the following aspects:

Phrase Analysis: Analyze the title and content for phrases that are "sensational" or "incendiary." If such phrases are present, the content is more likely to be false.

Common Sense Analysis: Analyze whether the title and content align with common sense. If they do not, it is highly likely to be false.

Bias Analysis: Analyze whether the news article has a biased political stance. If it does, it may be deliberately spreading false information.

Language Analysis: Analyze for spelling mistakes, grammatical errors, etc. If such errors are present, it suggests a higher likelihood of falsehood.

Finally, based on these different dimensions, summarize the issues identified in the input information. This part, like the information gatherers, is only called once at the beginning, and the results are then reused by other agents.

Rebuttals Attack the opponent's statements from the previous round.

Initialization: Use the opponent's viewpoint as the initialization. For example, if you are on the affirmative side, use the negative side's viewpoint as the initialization.

Process: Each time, refer to the previous round's statements from the opposing camp, summarize the information gathered by your teammates (information gatherers and analysts), and refute the opponent's viewpoints, extracting the refutation content.

Content of Analysis: The specific angles for refutation include the authenticity of the evidence provided by the opponent, the authority of the sources, the consistency of the context, any contradictions in language, and whether the generated explanations contain repetitive statements.

Summarizers Use uncertainty methods to summarize the information provided by different group members, with the specific aggregation method[35] as follows:

Uncertainty Measurement: The model utilizes its own generated uncertainty measures by producing outputs through multiple samplings.

Uncertainty Propagation Method: Incorporate uncertainty into subsequent processing, and adjust the attention mechanism of large language models to scale the attention weights of tokens based on the confidence levels of agents.

3.3 Judge

Multidimensional Scoring System Each judge j rates both sides of the debate (affirmative A, negative N) based on the following dimensions d, and the scoring results are used for subsequent voting and final judgment. Let $S_{j,d,A}$ and $S_{j,d,N}$ represent the score assigned by judge j to side A and side N respectively for dimension d, where $d \in \{E, C, N, P\}$. The scores are normalized to the range [0, 1].

Credibility of Evidence $S_{j,E,A} \in [0, 1]$: The authenticity and authority of the evidence provided by both sides.

Logical Consistency $S_{j,C,A} \in [0, 1]$: The logical rigor between the arguments and evidence presented by both sides.

Language Neutrality $S_{j,N,A} \in [0, 1]$: The presence of emotional or biased language tendencies.

Overall Persuasiveness $S_{j,P,A} \in [0, 1]$: The clarity, completeness, and persuasiveness of the overall argumentation.

Dynamic Confidence-Weighted Consensus Voting: The core of this mechanism lies in the fact that the voting weight of each judge is not fixed, but is dynamically adjusted based on their confidence levels regarding specific aspects[4].

Initial Voting: Each judge j casts a vote V_j on the debate outcome (e.g., for the affirmative side, for the negative side, or a tie) according to their professional role. The vote can be represented as:

$$V_j = \begin{cases} +1, & \text{if judge j votes for the affirmative side} \\ -1, & \text{if judge j votes for the negative side} \\ 0, & \text{if judge j votes for a tie} \end{cases}$$

Confidence Weighting: Each judge's voting result is multiplied by their respective confidence rating to obtain a weighted vote Wj: Wj = Vj × Cj For example: A factual judge believes the affirmative side wins (Vfactual = 1) with a confidence of 90% (Cfactual = 0.9), so the affirmative side receives a weight of Wfactual = 1×0.9 = 0.9. A logical judge believes the negative side wins (Vlogical = −1) with a confidence of 70% (Clogical = 0.7), so the negative side receives a weight of Wlogical = −1 × 0.7 =−0.7. A rhetorical judge believes the result is a tie (Vrhetorical = 0)

with a confidence of 60% (Crhetorical = 0), so the tie receives a weight of Wrhetorical = 0 × 0.6 = 0.

（3）**Consensus Assessment:** A "consensus" metric,

denoted as Consensus, is introduced to measure the level of agreement among the judges' judgments. We use the following formula metric for measurement:

$$Consensus = \frac{\sum_{j=1}^{N} C_j \cdot I(V_j = Vote_{win})}{\sum_{j=1}^{N} C_j}$$

where the numerator sums the confidence scores Cj of all judges whose vote Vj matches the vote corresponding to the winning outcome (V otewin, determined by the highest weighted total), and the denominator normalizes by the total confidence of all judges; a high consensus indicates strong agreement, particularly among highly confident judges, supporting the winning outcome, while a low consensus suggests weaker agreement or significant confidence in opposing viewpoints, potentially prompting further analysis or the consideration of additional perspectives.

Result Aggregation: The weighted voting results are aggregated. The overall weight for each possible outcome (affirmative, negative, tie) is calculated by summing the weighted votes for that outcome.

Total weight for Affirmative (W_A): Sum of W_j for all judges who voted for the affirmative side ($V_j = 1$). $W_A = \sum_{j=1}^{N} W_j \cdot I(V_j = 1)$.

Total weight for Negative (W_N): Sum of W_j for all judges who voted for the negative side ($V_j = -1$). $\mathrm{W_N} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{W_j} \cdot \mathrm{I}(\mathrm{V_j} = -1)$.

Total weight for Tie (W_T): Sum of W_j for all judges who voted for a tie (V_j = 0). $\mathrm{W_T} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{W_j} \cdot \mathrm{I}(\mathrm{V_j} = 0)$. The option with the highest total weight is the final outcome of the debate. $FinalOutcome = \arg\max\{W_A, W_N, W_T\}$.

4 Experiments

We performed experiments involving three generative tasks: Knowledge-Based Question Answering (KB-QA), Dialogue, and Summarization.

4.1 Experimental Setup

For all three tasks, we instruct ChatGPT to perform claim extraction, generate queries, and conduct multi-agent debate verification. The verification process is carried out for a minimum of two rounds, during which 10 snippets of evidence are gathered. The selected transition method entails switching to the skeptic agent when the response is assessed as True.

Table 1: The Number of positive and negative samples in different datasets.

Datasets	Positive	Negative
Factool QA	23	27
HaluEval QA	75	75
HaluEval Summarization	25	25
HaluEval Dialogue	80	70

Datasets and Baselines In this paper, experimental evaluations were conducted across three distinct tasks: Question-Answer (QA), Summarization, and Dialogue. Experimental datasets were sourced from the following two canonical databases:

Factool[1]: The Factprompts dataset comprises real-world questions and ChatGPT-generated responses, accompanied by Factool-annotated claims extracted therefrom.

HaluEval[15]: HaluEval constitutes an extensive corpus of model-generated and human-annotated hallucinated instances, obtained via a sample-then-filter methodology, serving as an evaluative benchmark for language model performance in hallucination recognition.

For evaluation, we randomly selected 150 QA, 50 Summarization (reflecting higher output complexity and claim density), and 150 Dialogue samples from HaluEval, ensuring a balanced binary distribution ($p = 0.5$) of positive and negative instances. We compared our method against Factool, HaluEval's few-shot prompting, and the self-check method[1].

Implementaion Details KB-QA: For complex and information-dense Question-Answer (QA) datasets, such as those found in Factool[1], we deconstructed responses into multiple atomic claims and performed dual-position debate verification on each individual claim. If one of the claims was determined to be hallucinated, the original response was classified as non-factual. Since the Factool dataset did not contain corresponding evidence, Google Search was utilized to obtain necessary evidence for this verification.

Conversely, for simpler QA data, as exemplified in HaluEval[15], where answers may consist of singular entities (e.g., "What is the name of the American quarterly fashion magazine published by Condé Nast? Vogue."), we concatenated the questions and answers to create QA pairs. Subsequently, dual-position debate verification was directly applied to these pairs, utilizing the evidence provided within the dataset.

Summarization: In the case of model-generated summaries, these were viewed as responses that were decomposed into multiple claims, each of which was verified independently. The document corresponding to the summary served as the evidence base. To prevent issues with excessively long input queries, each sentence from the document was encoded separately, along with the query. We then selected the top 10 most relevant sentences to serve as evidence for the current claim.

Dialogue: A key challenge in dialogue tasks is extracting factual claims from responses with subjective content. To tackle this, we added a preprocessing step in the information gatherers, directing ChatGPT to eliminate subjective elements before extracting claims, retaining only informative sections for verification. Additionally, we integrated dialogue history and external knowledge as supplementary evidence during claim extraction and verification.

Table 2: The Factool Dataset[1] evaluates four methods based on Accuracy(%), Recall(%), Precision(%), and F1 score(%). Claim-Level indicates the assessment of all annotated claims, while Response-Level refers to the evaluation of the initial responses. Superior scores are emphasized in **bold**.

Method	**Claim-Level**				**Response-Level**			
	Acc.	R	P	F1	Acc.	R	P	F1
Self-Check (0)	75.54	90.40	80.00	84.88	54.00	60.87	50.00	54.90
Self-Check (3)	69.53	81.36	79.12	80.23	54.00	47.83	50.00	48.89
FACTOOL	74.25	73.45	90.91	81.25	64.00	43.48	66.67	52.63
Our Method	78.30	82.42	89.24	85.69	73.00	54.27	81.20	65.0

Table 3: Our method along with the baseline was evaluated on the HaluEval dataset[15]. The dataset was used to conduct experiments on three key tasks: QA, Summarization, and Dialogue. The highest achieving scores are presented in **bold**.

Method	**QA**				**Summarization**				**Dialogue**			
	Acc.	**R**	**P**	**F1**	**Acc.**	**R**	**P**	**F1**	**Acc.**	**R**	**P**	**F1**
HaluEval	56.00	77.33	54.21	63.74	58.00	**100.0**	54.35	**70.42**	68.00	**75.71**	63.10	68.83
FACTOOL	67.33	**86.67**	62.50	72.63	64.00	48.00	70.59	57.14	74.67	70.00	74.24	72.06
Ours	**75.00**	84.00	**70.59**	**76.71**	**72.00**	66.41	**74.19**	70.08	**78.00**	71.43	**80.00**	**75.47**

4.2 Performance Analysis

The empirical results of our experiments are detailed in Tables 2 and 3. Table 2 presents a comparative performance analysis of our proposed method on the Factool dataset[1], outlining the results at the claim and response levels. Compared to various existing methods, our proposed approach consistently achieves the highest accuracy. This indicates that our method has a significant advantage in handling complex fact-checking tasks, effectively identifying true claims and responses.

Table 3 presents the evaluation results on the HaluEval dataset[15]. By comparing the performance of different methods, we observe that our approach demonstrates the best accuracy across most metrics in all three tasks. This further confirms the effectiveness and robustness of our method. It is worth noting that our method exhibits relatively low recall in the three tasks of this dataset. This is because we adopt a core strategy of questioning claims that are verified as factual, ensuring precise detection of errors when

claims are misclassified. However, this approach also leads to some statements that do not contain hallucinations being incorrectly judged as non-factual. The delicate trade-off between precision and recall resulting from our conservative verification method will be further elaborated in the next subsection.

Table 4: **Analysis of Various Transition Methods.** We assessed the impact of different transition methods by evaluating their performance on 80 QA samples, with a minimum of two debate rounds. The highest scores are indic ated in **bold**.

Method	Acc.	R	P	F1
Always **Skeptic**	69.00	85.01	63.10	72.43
Always **Trust**	73.20	85.18	66.62	74.77
True → **Trust**	71.92	**90.42**	64.41	75.23
True → **Skeptic**	**74.35**	87.81	**67.32**	**76.21**

Table 5: Comparison with Non-GPT Method. We benchmarked our method against the Non-GPT Method Wecheck on Factool data, with top scores bolded.

Method	Acc.	R	P	F1
Wecheck	65.82	66.02	86.51	74.89
Our method	78.30	82.42	89.24	85.69

4.3 Ablation Study

Transition Methods We investigated the effects of various transition methods. Utilizing the QA section of the HaluEval dataset [15], we extracted 80 samples to assess the impact of four distinct transition strategies: transitioning to S2 when the prior state evaluated the current claim as lacking hallucination (True → **Skeptic**); transitioning to S1 under the same condition (True → **Trust**); and consistently transitioning to S1 or S2 regardless of the preceding state's assessment of the claim. The findings, as illustrated in Table 4, indicate that the True → **Skeptic** transition method yielded the highest performance across three metrics. This can be attributed to the method's focus on challenging claims that were classified as factual in the previous state, leading to a thorough examination for any possible oversights. As discussed in the performance analysis section, this approach results in lower recall scores compared to the True → **Trust** method, while demonstrating higher precision values.

Minimum Rounds of Debate We examined the impact of different minimum numbers of debate rounds on outcomes. Using the previously compiled HaluEval dataset[15], we analyzed three distinct tasks by varying the minimum debate rounds from 0 to 3. The results, as shown in the figure2, were obtained using the "True →

Skeptic" transition method. We found that performance generally improves when the minimum rounds are set to 1 or 2. However, there is a significant decrease in efficacy when the minimum rounds are set to 3.

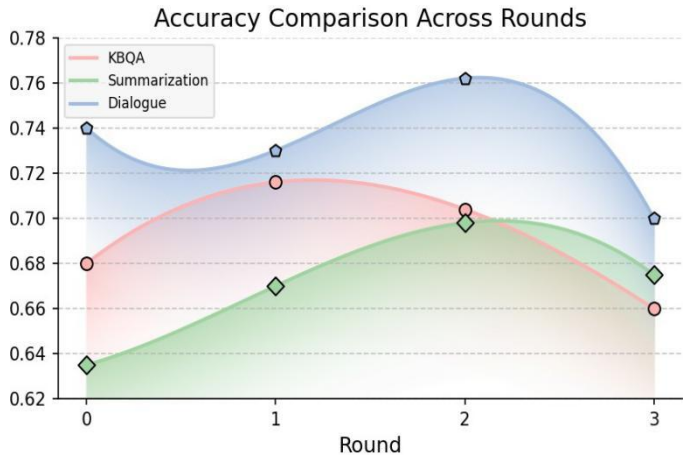

Fig. 2: **Comparison of Different Minimum Debate Rounds.** The debate transition method is set to True → **Skeptic**. The **x**-axis displays the varying minimum debate rounds, while the **y**-axis shows the corresponding detection accuracy.

Comparison with Non-GPT Method In the multi-agent verification phase of the experiment using the Factool dataset, we implemented the WeCheck[30] methodology to perform an ablation study, highlighting the advantages of our approach. We maintained the initial two steps constant, employing the Factool method to extract claims and gather evidence. The claims were treated as hypotheses, while the evidence served as the premises. Instances with WeCheck scores of 0.5 or higher were classified as factual. As indicated by the experimental results presented in Table 5, our method demonstrates substantial benefits during the verification stage compared to the non-GPT method.

Table 6 : Impact of Module Exclusion on Model Performance

Method	Acc.	R	P	F1
Our method	78.30	82.42	89.24	85.69
(w/o) Information Gatherers	72.50	76.00	83.50	79.57
(w/o) Analysts	74.00	78.44	85.03	81.60
(w/o) Summarizers	75.20	79.74	86.23	82.86
(w/o) the three modules	65.75	67.22	70.50	68.82

Effect of Modules To evaluate the contribution of each component within the DPD framework, we conducted ablation studies by removing individual modules, as detailed in Table 6.

Information Gatherers Module Removal: Without this module, dialogue responses tend towards subjective opinions, lacking objective factual grounding. This absence of verifiable information impedes fact-checking efficacy, consequently compromising judgment and prediction accuracy.

Analyzer Module Removal: Eliminating this module diminishes the model's core decision-making and pattern recognition capabilities. This impairs its ability to accurately detect errors and fabricated information within the generated content, thus degrading overall detection performance. Concurrently, the model's reasoning and logical assessment faculties are weakened, hindering its capacity to evaluate the logical coherence of generated content and further reducing the accuracy of hallucination detection.

Summarizer Module Removal: The absence of this module results in verbose and unwieldy outputs, impeding users' ability to rapidly extract key information and diminishing the usability of hallucination detection results. Moreover, information redundancy due to ineffective summarization obscures the core meaning and conclusions of the detection process, thereby affecting user comprehension and application of these findings.

Simultaneous Removal of All Three Modules: Completely ablating these modules renders the model incapable of identifying and analyzing hallucinated information, preventing the generation of meaningful detection results. The model's operation devolves to near-random behavior, with detection performance approaching chance levels.

In conclusion, each module is integral to the DPD framework's functionality. Removing any single module significantly degrades the system's overall performance, underscoring the critical role of the information gatherer, analyzer, and summarizer in the model architecture.

5 Conclusion

In this paper, we aim to improve hallucination detection accuracy in content generated by large language models and ensure its broad applicability beyond specific tasks. To achieve this, we propose a novel Dual-Position Debate (DPD) framework that integrates affirmative and negative debates to enhance detection accuracy. The framework divides agents into two teams—affirmative and negative—each comprising roles such as information gatherers, rebutters, analysts, and summarizers. These agents collaborate within their teams and engage in cross-team debates. Evaluations on the KBQA and HaluEval datasets demonstrate the effectiveness of DPD. Our approach is generalizable and can be combined with other post-processing methods for superior hallucination detection and mitigation performance.

References

1.
2. Chern, I., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., Liu, P., et al.: Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528 (2023)
3. Coffey, A.: Analysing documents. The SAGE handbook of qualitative data analysis pp. 367–379 (2014)
4. Das, S., Saha, S., Srihari, R.K.: Diving deep into modes of fact hallucinations in dialogue systems. arXiv preprint arXiv:2301.04449 (2023)
5. Dogan, A., Birant, D.: A weighted majority voting ensemble approach for classification. In: 2019 4th International Conference on Computer Science and Engineering (UBMK). pp. 1–6. IEEE (2019)
6.
7. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
8. Elaraby, M., Lu, M., Dunn, J., Zhang, X., Wang, Y., Liu, S., Tian, P., Wang, Y., Wang, Y.: Halo: Estimation and reduction of hallucinations in open-source weak large language models. arXiv preprint arXiv:2308.11764 (2023)
9. Fu, Y., Peng, H., Khot, T., Lapata, M.: Improving language model negotiation with self-play and in-context learning from ai feedback. arXiv preprint arXiv:2305.10142 (2023)
10. Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A.T., Fan, Y., Zhao, V.Y., Lao, N., Lee, H., Juan, D.C., et al.: Rarr: Researching and revising what language models say, using language models. arXiv preprint arXiv:2210.08726 (2022)
11. Goyal, T., Durrett, G.: Annotating and modeling fine-grained factuality in summarization. arXiv preprint arXiv:2104.04302 (2021)
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 1–42 (2018)
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys 55(12), 1–38 (2023)
14. Krys´cin´ski, W., McCann, B., Xiong, C., Socher, R.: Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840 (2019)
15. Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P.N., Shoeybi, M., Catanzaro, B.: Factuality enhanced language models for open-ended text generation. Advances in Neural Information Processing Systems 35, 34586–34599 (2022)
16. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33, 9459–9474 (2020)
17. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y., Wen, J.R.: Halueval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:2305.11747 (2023)
18. Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: Eliciting truthful answers from a language model. Advances in Neural Information Processing Systems **36** (2024)

19. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118 (2023)
20. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI Open 5, 208–215 (2024)
21. Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., Singh, S.: Entity-based knowledge conflicts in question answering. arXiv preprint arXiv:2109.05052 (2021)
22. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G.: Hallucination detection and hallucination mitigation: An investigation. arXiv preprint arXiv:2401.08358 (2024)
23. Manakul, P., Liusie, A., Gales, M.J.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896 (2023)
24. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2005.00661 (2020)
25. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article 2(5) (2023)
26. Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al.: Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813 (2023)
27. Sun, Q., Yin, Z., Li, X., Wu, Z., Qiu, X., Kong, L.: Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. arXiv preprint arXiv:2310.00280 (2023)
28. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
29. Wang, H., Du, X., Yu, W., Chen, Q., Zhu, K., Chu, Z., Yan, L., Guan, Y.: Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. arXiv preprint arXiv:2312.04854 (2023)
30. Wang, Q., Wang, Z., Su, Y., Tong, H., Song, Y.: Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? arXiv preprint arXiv:2402.18272 (2024)
31. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
32. Wu, W., Li, W., Xiao, X., Liu, J., Li, S., Lv, Y.: Wecheck: Strong factual consistency checker via weakly supervised learning. arXiv preprint arXiv:2212.10057 (2022)
33. Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N.A., Ostendorf, M., Hajishirzi, H.: Fine-grained human feedback gives better rewards for language model training. Advances in Neural Information Processing Systems 36, 59008–59033 (2023)
34. Xiong, K., Ding, X., Cao, Y., Liu, T., Qin, B.: Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. arXiv preprint arXiv:2305.11595 (2023)
35. Xu, W., Agrawal, S., Briakou, E., Martindale, M.J., Carpuat, M.: Understanding and detecting hallucinations in neural machine translation via model introspection. Transactions of the Association for Computational Linguistics 11, 546–564 (2023)
36. Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., Wu, Y.: Towards reasoning in large language models via multi-agent peer review collaboration. arXiv preprint arXiv:2311.08152 (2023)
37. Yoffe, L., Amayuelas, A., Wang, W.Y.: Debunc: mitigating hallucinations in large language model agent communication with uncertainty estimations. arXiv preprint arXiv:2407.06426 (2024)

38. Zhang, J., Li, Z., Das, K., Malin, B.A., Kumar, S.: SacΘ3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. arXiv preprint arXiv:2311.01740 (2023)
39. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al.: Siren's song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219