Attributive Abstention in Retrieval-Augmented Generation via Multi-Agent Collaboration

Anonymous ACL submission

Abstract

Abstention is critical for Retrieval-Augmented Generation (RAG) systems to avoid generating erroneous or hallucinated answers, especially when retrieved documents are contradictory, irrelevant, or incomplete. Existing approaches primarily focus on model-centric confidence thresholds, neglecting the nuanced challenges posed by the quality of external data. To address this issue, we propose a novel framework, Attributive Abstention via Multi-Agent Collaboration (A2MAC), which employs large reasoning models to drive specialized agents in collaborative reasoning to assess document reliability and determine whether to abstain. By decomposing the decision-making process into three modular checks—contradiction detection, relevance evaluation, and reasoning sufficiency—our approach generates attributive abstention rationales grounded in the retrieved content. Unlike prior work, A2MAC explicitly addresses RAG-specific failures by simulating multi-perspective collaborations enabled by large reasoning models, enabling the system to identify data flaws and justify refusals transparently. Experimental results on benchmark datasets demonstrate that A2MAC outperforms state-of-the-art baselines in both abstention accuracy and explanatory correctness, establishing a robust foundation for trustworthy RAG systems that balance productivity with factual integrity.

1 Introduction

Large Language Models (LLMs) excel at knowledge-intensive tasks but suffer from two key limitations: hallucinations (fabricated outputs) and stale knowledge (inability to reflect real-time updates). Retrieval-Augmented Generation (RAG) addresses these issues by transforming the LLM into an information integrator—dynamically combining its parametric knowledge with documents retrieved from external databases. However, a new problem is that the quality of retrieved documents is difficult to ensure that they are all relevant, complete and non-contradictory to the question, which can lead to the model answering incorrectly as shown in Figure 1, so refusing to generate low-confidence outputs should be a basic capability of RAG and a requirement to ensure the factual capability of the model.

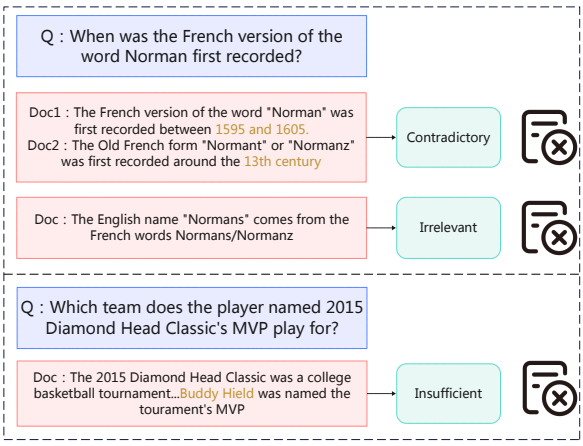


Figure 1: An example of three common issues in RAG documents: contradictory, irrelevant, and insufficient.

Most of the existing methods for rejecting answers are based on generative QA rather than RAG scenarios, and have achieved some results, which we summarize in the following categories: (1) Calibration-Based: The LLM scores its own output answers and sets a threshold; answers below the threshold are discarded. (2) Prompt-Based: Utilizes in-context examples and hints (e.g., "Answer the question only if answerable" or "Answer the below question if it is safe to answer") to determine whether a question can be answered. (3) Training-Based: Connects a shallow network to the LLM's hidden layers to train a binary classifier for deciding whether to answer or refuse. (4) Consistency-Based: Samples multiple answers and judges whether to respond based on their consis-

tency. (5) Hidden-States Steering Method in the Safety Domain: Uses the LLM's internal features through self-supervised methods (e.g., PCA) to identify dangerous queries and refuse to answer. These methods focus on whether a query is answerable; however, we emphasize evaluating issues in the retrieved documents, such as whether contradictory information, irrelevant documents, or incomplete information prevent answering the question.

Recall (Arditi et al., 2024) is the first work to study model robustness under counterfactual external knowledge, proposing a RAG dataset and testing refusal capabilities using prompt-based methods. Trust-Align (Bianchi et al., 2023) is the first to investigate hallucinations of LLMs in a RAG setup, where model responses should be exclusively grounded in retrieved documents rather than the model's parametric knowledge. They use SFT and DPO alignment methods to enable the model to learn both refusal and grounding capabilities. However, the process is complex, and generalization has not been validated (requiring re-fine-tuning on new datasets). Both works provide preliminary exploration of refusal in RAG settings but do not analyze the reasons for refusal from the perspective of the retrieved documents.

We have made two observations in current Abstention tasks. First, we observe that the retrieved documents often suffer from the following issues: (1) Contradictory Information: Some critical information contains multiple or contradictory details. For example, WikiContradict (Hou et al., 2024) contains real-world knowledge conflicts from Wikipedia. (2) Irrelevant Information: This may include completely unrelated content or content that shares some keywords but is not relevant to the question. (3) Incomplete Information: The answer to a question requires multiple elements, but the document provides only partial information, such as comparing the market value of companies A and B when only A's market value is available.

Another observation is that the rejection model typically only provides a rejection decision, while the reasoning behind it remains a black box and is difficult to interpret. However, in many fields, model attributability is essential. For example, in precision medicine, experts require far more information from the model than a simple binary prediction to support diagnosis. Similarly, in critical domains such as autonomous vehicles, transportation, security, and finance, AI algorithm decisions also need be attributive.

To address these two issues, we propose Attributive Abstention via Multi-Agent Collaboration(A2MAC), that can analyze whether the retrieved documents contain aforementioned issues. While providing a rejection decision, it also explains the reason for rejection. Our contributions are:

(1) We address three issues in the retrieved documents and designed a rejection model based on agent collaboration and reasoning.

(2) We achieve attributive abstention, providing transparent and verifiable rationales for refusal decisions, a critical step towards trustworthy AI.

(3) Experimental results show that A2MAC delivers superior abstention performance on RAG abstention datasets while providing reliable and attributable responses.

2 Related Work

2.1 Abstention of Large Language Models

There are three specific scenarios in which a large language model will need to reject an answer due to lack of capacity or security concerns: (1) The first scenario is where the model is capable of answering but has security implications that it cannot answer. Zhong et al. (2024); Bianchi et al. (2023); Lee et al. (2024b) learn abstention by fine-tuning LLMs using harmful and harmless inputs. Another way is to apply activation steering technique to alter model behaviors (Arditi et al., 2024; O'Brien et al., 2024; Lee et al., 2024a), which allows the model to refuse to answer by adding behavior vectors to the hidden layer. (2) The second case is when the answer needs to be rejected due to insufficient knowledge within the LLM; There has been a lot of work: Calibration-based methods set a threshold for (3) While sometimes we need external knowledge to help figure out the question, then verification of the external documents will lead to abstention.

2.2 Multi-Agent Collaboration

Multi-Agent Collaboration is a kind of agentic workflow, designed to enhance the accuracy and robustness of generated outputs by leveraging the collective reasoning and critical evaluation capabilities of multiple agents. This collaborative process fosters a dynamic environment where diverse perspectives converge, inconsistencies are identified, and more nuanced, well-supported solutions emerge through iterative discussion and refinement.

(Du et al., 2023) shows the effectiveness in improving factual correctness and reasoning accuracy by introducing a multi-agent process. PREDICT (Park et al., 2024) is a multi-agent debate framework leveraging LLMs to improve generalized hate speech detection by simulating debates between agents with diverse perspectives from different datasets. This approach builds consensus and overcomes overfitting caused by inconsistent labeling criteria, leading to better generalization and performance on various benchmarks. (Liu et al., 2024) designs a multi-agent framework that incorporates a voting mechanism to give a jailbroken score and reason, giving a promising method in jailbreak judge. These collectively demonstrate that LLMs, when integrated into multi-agent collaboration frameworks, excel in accuracy, robustness, generalization, and interpretability. By simulating human-like social decision-making processes (e.g., debate, negotiation, and voting), LLMs leverage their language understanding and generation capabilities to perform closer to human-level competence, addressing limitations of single-agent models such as bias, overfitting, or shallow reasoning.

3 A2MAC: The Proposed Approach

3.1 Problem Formulation

In a Retrieval-Augmented Generation (RAG) scenario, given a query q and a set of documents D, the task is to determine a judgment $J \in \{0, 1\}$. If the query is declined, $J = 0$, and the reason for abstention R, S is provided in the output. If the query is answered, $J = 1$, indicating the model's confidence in its answer A based on D and it will. R is a task-specific keyword corresponding to reasons such as contradictory, irrelevant, or insufficient, while S provides a thorough justification for abstention, typically including an analysis of both the query and the documents.Unlike prior studies that evaluate large models using general queries (Feng et al., 2024; Brahman et al., 2024), this approach tests the model's ability to provide answers grounded in the given documents, offering a more rigorous and insightful evaluation.

3.2 Approach Overview

To validate potential issues in RAG documents and improve the reliability of model responses, we introduce A2MAC, an acronym for **A**ttributive **A**bstention via **M**ulti-**A**gent **C**ollaboration. The A2MAC framework, depicted in Figure 2, comprises three core components: (1) **Fast Response**: Upon receiving a query and its corresponding documents, the model rapidly assesses whether to reject the query or identifies key aspects requiring further validation. (2) **Checker Collaboration**: Three agent-based checkers, guided by the Fast Response, individually undertake predefined or dynamically generated subtasks to conduct thorough validations. (3) **Fusion and Summary**: The results are integrated to deliver a final judgment, providing a clear abstention rationale and a concise summary.

3.3 Fast Response Module

First, the model generates an initial impression response to a given question and its corresponding document pair. Such rapid responses typically provide most of the useful information. To this end, we have developed three optional strategies: The first strategy involves a standard large language model directly generating a response to obtain a preliminary judgment on whether to refuse to answer $J_1 = LLM\{q, D\}$. The second strategy employs an inference-based large language model, which not only produces a refusal judgment but also provides a detailed reasoning process $J_1, S_1 = LLM\{q, D\}$. The third strategy extends the first two by not specifying which aspects of the document the model should use to determine whether to respond; instead, it allows the model to autonomously identify key points $P = \{p_1, p_2, p_3\}$ warranting further exploration, which are then evaluated by checkers in the second step. Notably, only one of these three strategies needs to be selected for use. In subsequent experiments, we will compare the effectiveness of these three strategies.

3.4 Checker Collaboration Module

This module employs a Fast Response approach, leveraging agent-based checkers to assess critical sub-issues. The checkers are categorized into two types based on their configuration: manually defined checkers and adaptive generative checkers. Each type produces three judgments, denoted as $J_2 = \{J_{21}, J_{22}, J_{23}\}$.

3.4.1 Manually Defined Checkers

To systematically detect the three major types of retrieval failures—*irrelevance*, *insufficiency*, and *contradiction*—we design three independent checkers, each specialized for one issue. Together, they constitute a complete and minimal taxonomy of potential document-level errors in RAG systems, as

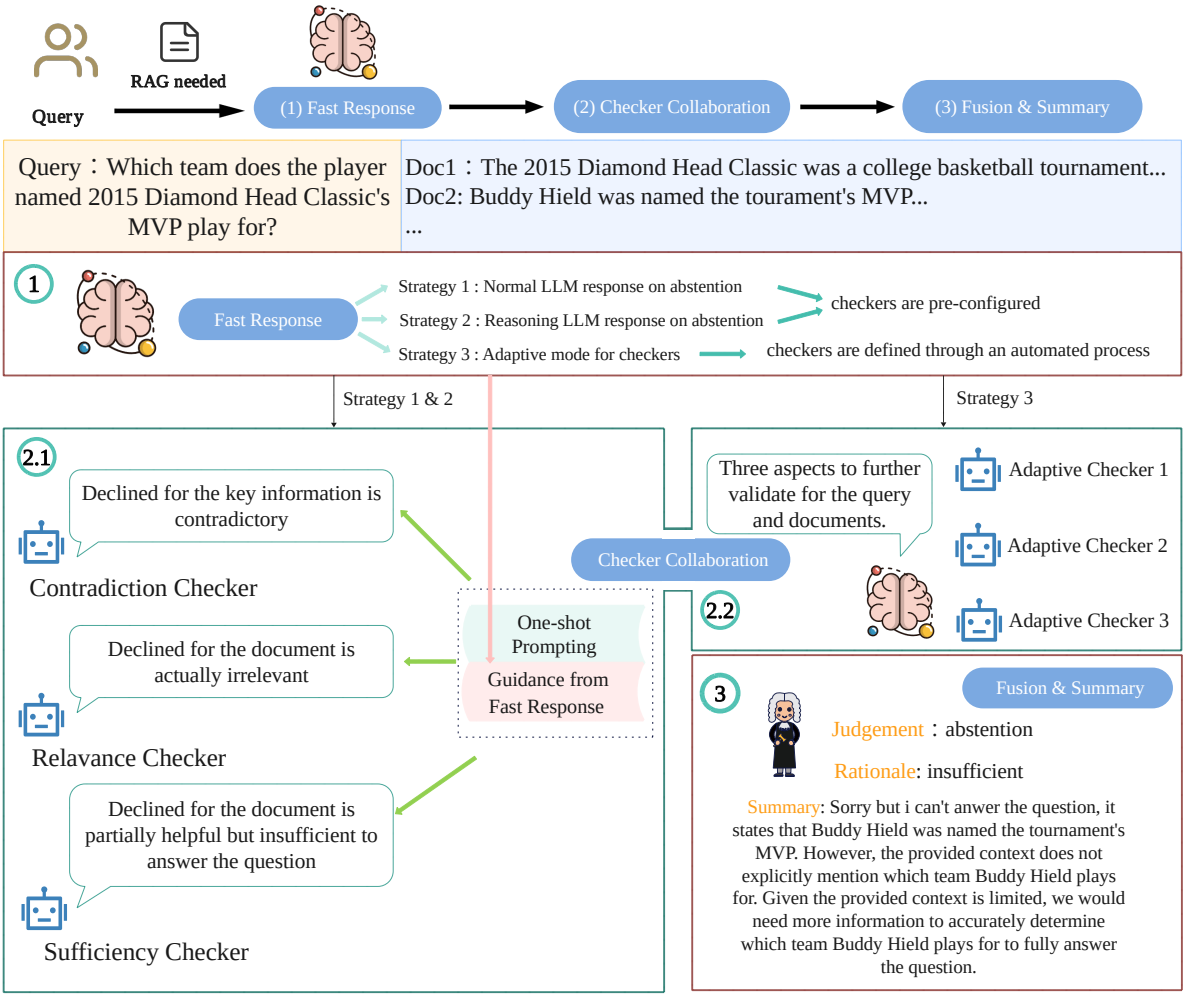


Figure 2: The illustration of A2MAC. A2MAC consists of three main components: Fast Response, Checker Collaboration and Fusion and Summary. These components are executed sequentially.

formally justified in Appendix E. This modular design also mitigates information overload that may occur when a single agent attempts to reason over multiple issues simultaneously. The implementation follows the procedures described in Strategy 1 and Strategy 2.

For Strategy 1, the formulation is defined as

$$J_{2i}, DE_i = \text{LLM}\{q, D, J_1, E\},$$

where $i \in \{1, 2, 3\}$ represents the three checkers, and E is a one-shot prompting defined as $E = \{(q_{\text{ex}}, a_{\text{ex}})\}$, where q_{ex} is an example query and a_{ex} is the corresponding expected output. The inclusion of E is motivated by the observation that the concepts of contradiction, relevance, and sufficiency are relatively abstract for LLMs, which may lead to deviations in task understanding. The one-shot prompting E facilitates a clearer understanding of the task requirements. Here, DE denotes a detailed explanation of the binary judgment J.

For Strategy 2, the formulation is extended to

$$J_{2i}, DE_i = \text{LLM}\{q, D, J_1, E, S_1\}, \quad i \in \{1, 2, 3\}.$$

Considering the efficiency of the agents, the checkers in this section employ smaller-scale, faster-responding models instead of reasoning models to optimize performance.

3.4.2 Adaptive Generative Checkers

In contrast to Manually Defined Checkers, Adaptive Generative Checkers does not explicitly instruct the checker on the specific perspectives from which to make further judgments. Instead, we aim to explore the model's capability to autonomously devise its own problem-solving strategies. The adaptive prompt can be found in appendix B.3:

In Strategy 3, we have identified three key factors, denoted as P, adaptively generated by the model, which influence the decision to refuse a query. Consequently, the checkers can indepen-

dently evaluate these three factors, formalized as follows:

For standard LLMs, the judgments and explanatory answers are expressed as

$$J_{2i}, DE_i = \text{LLM}\{p_i, q, D\}, \quad i \in \{1, 2, 3\}.$$

Similarly, for reasoning LLMs, the judgments and explanatory answers are given by

$$J_{2i}, DE_i = \text{LLM}\{p_i, q, D, S_1\}, \quad i \in \{1, 2, 3\}.$$

3.5 Abstention Decision Module

To integrate the judgments provided by the checkers into a final rejection decision and rationale, We consider the output from the checker as an ordered sequence of dialogue turns. Each turn consists of a prompt p_{2i}, an intermediate model output J_{2i}, and a corresponding explain DE_i. The dialogue history is represented as a sequence H, where each element is a dialogue tuple:

$$H = [(p_{21}, J_{21}, DE_{21}),\ (p_{22}, J_{22}, DE_{22}),\ (p_{23}, J_{23}, DE_{23})]\,.$$

Given the dialogue history H and a new prompt p_3, the final outputs J, S are generated by the language model as follows:

$$J, S = \text{LLM}(H,\ p_3).$$

If $J = 0$, an abstention reason R is given, otherwise answer the query with A. We opted for an LLM-based integration method over a standard majority vote to form a consensus. This is because majority voting can be brittle in complex scenarios, especially when a minority opinion is correct. Our approach can weigh the substance of each checker's argument, capturing valuable insights that a simple vote count would discard, thus leading to better results (Subsequent experimental results also demonstrate this point).

4 Experiments

In this section, we empirically verify the effectiveness of A2MAC by addressing the following research questions:

RQ1: Does A2MAC improve the abstention quality in general LLM and reasoning LLM?

RQ2: Which is more suitable for the Abstention Decision Module: majority voting or LLM-based integration?

RQ3: Does A2MAC improve the explanatory correctness of the reasons for response refusal?

RQ4: How is the qualitative illustrations of A2MAC?

RQ5: How does the number of checkers affect the abstention quality?

The source code and detailed prompts have been shared at: https://anonymous.4open.science/r/A2MAC-8E55.

4.1 Experiments Settings

4.1.1 Datasets.

To our knowledge, we could not find a suitable RAG abstention dataset focused on documentation issues. So we construct a new dataset: **Documentation-Induced RAG Abstention (DIRA)**, a RAG dataset featuring unanswerable questions due to contradictory, irrelevant, or insufficient documentation from SQuAD2 (Rajpurkar et al., 2018), Wikicontract (Hou et al., 2024), and Hotpotqa (Yang et al., 2018). Detailed descriptions of these source datasets are provided in Appendix C.

The construction of DIRA: Each data instance in DIRA can be defined as a 5-tuple(q, D, J, A, R).

The dataset incorporates the entirety of Wikicontradict data (Hou et al., 2024), where each instance inherently contains q and D, with $J = 0$, A left empty, and R labeled as contradictory. Additionally, 2,000 instances are sampled from SQuAD2 (Rajpurkar et al., 2018), half of which are unanswerable questions with R labeled as irrelevant, while the remaining half serve as answerable counterparts. Furthermore, 2,000 instances are selected from Hotpotqa (Yang et al., 2018), comprising 1,000 answerable original examples and 1,000 modified unanswerable instances where a golden document has been deliberately removed, resulting in R being labeled as insufficient. Each subset of 1,000 instances is balanced with 500 bridge-type and 500 comparison-type questions.

To enhance the robustness of the dataset, perplexity filtering (PPLF) is employed. Given that large language models (LLMs) are typically pretrained on extensive public corpora, they may occasionally rely on internal knowledge rather than the provided reference documents D when generating answers. Although explicit prompts instruct the model to strictly adhere to the given documents, an additional safeguard is implemented. Specifically, during the selection of SQuAD2.0 and HotpotQA data,

instances exhibiting the highest perplexity when D is withheld are prioritized. Formally, we first compute the perplexity for each candidate question q in the source data Q_s as:

$$PPL(q) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P_{LM}(w_i \mid w_{<i}, \varnothing_D)\right),$$

where N is the number of tokens, P_{LM} denotes the language model's probability distribution without document access ($\varnothing_D$), and w_i represents the i-th token. The final selected subset $Q^* \subset Q_s$ consists of the top-n questions with maximal perplexity:

$$Q^* = \text{argtopk}_{q \in Q_s}(PPL(q), k = n).$$

This selection criterion ensures the chosen questions are inherently difficult to answer without document support, thereby enforcing strict document dependence in the benchmark.

The dataset statistics is shown in Table 1

Table 1: Statistics of the dataset DIRA after PPL filtering and transformation from three datasets.

	Contradictory	Irrelevant	Insufficient
Unanswerable	253	1000	1000
Answerable	-	1000	1000

4.2 Metrics

Table 2: Categories of the Output, GT denotes the ground truth, R represents Response, and N_i indicates the number of classes.

GT \ R	Right Answer	Wrong Answer	Abstention
Unanswerable	-	N_2	N_1
Answerable	N_5	N_4	N_3

For the evaluation of abstention decisions J, numerous metrics have been developed and employed in existing literature. In our task, we primarily adopt accuracy (ACC) and the F1 score as evaluation metrics. Additionally, we report Precision and Recall, since an over-abstaining model or one that responds too confidently can result in one of these metrics being disproportionately high while the other remains significantly low (as demonstrated in later experiments). The F1 score, which balances Precision and Recall, thus serves as a more appropriate metric in this context:

$$\text{ACC} = \frac{N_1 + N_5}{N_1 + N_2 + N_3 + N_4 + N_5},$$

$$\text{Precision}_{\text{abs}} = \frac{N_1}{N_1 + N_2},$$

$$\text{Recall}_{\text{abs}} = \frac{N_1}{N_1 + N_3},$$

$$\text{F1}_{\text{abs}} = 2 \cdot \frac{\text{Precison}_{\text{abs}} \cdot \text{Recall}_{\text{abs}}}{\text{Precison}_{\text{abs}} + \text{Recall}_{\text{abs}}}.$$

For evaluating the generated answer A and the corresponding abstention rationale R, we employ the Exact Match (EM) metric.

To assess the quality of the summary explanation for abstention S, we adopt the LLM-as-a-judge(Li et al., 2024) approach. The specific prompt can be found in Appendix B.1:

To investigate whether the model explicitly indicates the reasons for refusing to answer certain questions, we propose a metric called the Right Reason Rate (R3). In the DIRA dataset, refusal rationales are categorized into three types. We evaluate the model's performance on refused questions using accuracy. It is formally defined as:

$$\text{R3} = \frac{\sum_{i \in TN} \mathbb{I}(R_{\text{pred},i} = R_{\text{gt},i})}{|TN|}$$

where $|TN|$ is the total number of true negative samples (corresponding to count N_1 in Table 2), $R_{\text{pred},i}$ and $R_{\text{gt},i}$ are the predicted and ground-truth reasons for the i-th sample, and $\mathbb{I}$ is the indicator function.

4.3 Baselines and and Implementaition Details

In order to verify the efficiency of our method, we consider these four baselines: **RC-RAG** (Chen et al., 2024) develop a counterfactual prompting framework that focuses on the problem the referred passage may be poor or the usage of the retrieved passages is improper, **Calibration-based** (Tian et al., 2023) methods refuse to answer by assessing model confidence or uncertainty, declining when probabilities fall below a threshold or entropy is high, **Priori** (Ren et al., 2023) will assess whether it can respond based on the available information before answering a question. We also established a **Base** method, which, building upon standard RAG, directly instructs the LLM to focus on the documents and assess whether the document contains contradictions, irrelevance, or insufficient information. Its prompt can be found in appendix **??**.

We adopt two representative backbones: GPT-4o-mini(Achiam et al., 2023) and DeepSeek-R1(Guo et al., 2025). GPT-4o-mini is a closed-source general-purpose LLM, while DeepSeek-R1 is an open-source reasoning large model. During the experiment, we directly invoked the API, setting the temperature of GPT-4o-mini to 1.0 and DeepSeek-R1 to 0.7. For efficiency considerations, the checker configuration utilized the lightweight GPT-4o-mini. Both the Fast Response Module and the Abstention Decision Module were tested using both backbone models.

4.4 RQ1: Performance Evaluation of A2MAC

Table 3 presents the experimental results of A2MAC and baseline methods on the DIRA dataset. A2MAC refers to the approach using manually defined checkers, while A2MAC-adaptive employs adaptive generative checkers. In the Fast Response and Abstention Decision modules, we utilized GPT-4o-mini and Deepseek-R1, respectively. In the checker collaboration module, which involves multiple calls to large language models, we both used the more lightweight GPT-4o-mini. For RQ1, we made the following observations.

Compared to the baselines, A2MAC improves the accuracy of both refusal and refusal rationale across general large models and reasoning large models,demonstrating A2MAC's superior capability in identifying unanswerable questions within RAG: (1)Compared to baselines, A2MAC consistently improves abstention accuracy. Its superiority over confidence-based methods like Calibration highlights the unreliability of model confidence scores, while its advantage over Priori suggests that fine-grained document analysis is more effective than high-level instructions.(2)A2MAC outperforms the RC-RAG method, indicating that examining issues within the document itself is more effective than scrutinizing how the LLM utilizes the document. RC-RAG evaluates both document quality and the appropriateness of the model's use of the document, achieving relatively high Recall scores of 0.6945 and 0.6516, respectively, which demonstrates its ability to identify questions that should be refused. However, it suffers from over-refusal, leading to lower Precision and F1 scores. (3)The Base method outperforms the previous three methods, as the prompt explicitly addresses potential issues in the document. However, its performance is not as good as A2MAC. This performance gain validates our hypothesis: decomposing the complex validation task into specialized, collaborative checks is more robust and reliable than using a single, monolithic prompt. Thus, the added complexity of our framework is justified by a clear and measurable improvement in performance.

Reasoning LLMs are more suitable for A2MAC than general large models. The Base method yields an F1 score of 0.7272 for GPT-4o-mini and 0.6225 for Deepseek-R1, reflecting GPT-4o-mini's superior instruction-following ability. However, the A2MAC-adaptive method with Deepseek-R1 achieves the highest F1 score, suggesting that reasoning LLMs excel in complex tasks and systems.

Adaptive checkers are more suitable for reasoning models. As shown in Table 3, for GPT-4o-mini, A2MAC-adaptive underperforms compared to Base and A2MAC. In contrast, for Deepseek-R1, A2MAC-adaptive yields the best results, suggesting that reasoning large models can autonomously identify fine-grained sub-issues and validate them through checker collaboration, whereas general large models rely more heavily on manual design.

4.5 RQ2: The Performance of Major Voting and LLM-based Integration in Abstention Decision Module

In this section, we explore the effectiveness of adopting majority voting and LLM-based integration when combining decisions from different checkers. We compare the ACC, Recall, Precision, and F1 scores of A2MAC and A2MAC-adaptive on two LLMs. As shown in Figure 3, we observe that the LLM-based integration method outperforms majority voting in most metrics, particularly in Precision. However, Recall may be higher with majority voting. Since Precision and Recall are typically difficult to improve simultaneously in rejection methods, F1 serves as a more balanced metric. This suggests that the majority voting method is more cautious in choosing to reject but struggles to make accurate judgments for questions that should be rejected. This is understandable, as the LLM-based integration method not only considers the checkers' brief decisions but also their detailed analyses, sometimes leading to the minority opinion yielding the correct answer. Overall, LLM-based integration appears to be the more suitable approach.

Table 3: Performance metrics for different methods on GPT-4o-mini and DeepSeek-R1 models.

	Method	ACC	Recall	Precision	F1	R3
gpt-4o-mini	Base	0.6828	0.6837	0.7759	0.7272	0.4523
	Priori	0.5575	0.3714	**0.8960**	0.5251	0.4491
	Calibration	0.4714	0.2708	0.8288	0.4082	0.3672
	RC-RAG	0.5805	0.6929	0.6320	0.6610	0.4566
	A2MAC	**0.6832**	**0.6945**	0.8835	**0.7777**	**0.5006**
	A2MAC-adaptive	0.5815	0.6062	0.7014	0.6503	-
deepseek-R1	Base	0.6170	0.4647	0.9424	0.6225	0.5750
	Priori	0.5617	0.3758	0.8943	0.5292	0.4586
	Calibration	0.5187	0.3959	0.8902	0.5481	0.4283
	RC-RAG	0.5368	0.6516	0.5837	0.6158	0.5338
	A2MAC	0.6967	0.6028	**0.9411**	0.7348	**0.5881**
	A2MAC-adaptive	**0.7115**	**0.6885**	0.9138	**0.7853**	**-**

4.6 RQ3: The Explanatory Correctness of Reasons

Our approach provides an analysis of the reasons for refusal decisions. To evaluate the quality of this analysis, we employed the LLM-as-a-judge method (as detailed in Section 4.2), assessing the scores (1–4) for the Base Method, A2MAC, and A2MAC-adaptive, using Deepseek-R1 as the LLM. As shown in Table 4, both A2MAC and A2MAC-adaptive demonstrate improvements over the Base Method, with the A2MAC method achieving the highest performance. This indicates that our approach enhances the ability to identify contradictions, irrelevance, and insufficiencies in documents within the RAG framework, significantly improving the interpretability of refusal decisions.

Table 4: Using LLM-as-a-Judge to evaluate the quality of refusal analysis.

	Score1	Score2	Score3	Score4	Average
Base method	8.251	7.804	8.731	7.214	8.000
A2MAC	9.212	9.373	9.540	8.778	9.226
A2MAC-adaptive	8.927	9.117	9.305	8.174	8.881

4.7 RQ4: Case Study

In table 5, we provide a qualitative illustrations of our method. In this example, the answer should be rejected based on the documents. The baseline methods all choose to answer, whereas our methods, A2MAC and A2MAC-adaptive, correctly opted to refuse to answer. In the A2MAC method, even though the second checker chose to answer, the final answer, determined through agent collaboration and discussion, remained correct, demonstrating that the final decision selectively adopts valid information. In the A2MAC-adaptive method, the Fast Response module assigned specific inspection tasks to the three checkers, which unanimously chose to refuse to answer, ultimately yielding the correct result.

For RQ5, we analyze the effect of the number of checkers on model performance in the appendix (see Appendix D for details).

5 Conclusion

Enhancing the trustworthiness of Retrieval-Augmented Generation (RAG) systems requires a direct focus on the quality and reliability of the underlying documentation. In this paper, we propose **A2MAC**, a novel framework for attributive abstention through multi-agent collaboration, designed to identify scenarios in Retrieval-Augmented Generation (RAG) systems where abstention is warranted due to contradictory, irrelevant, or insufficient documentation. A2MAC not only delivers abstention decisions but also effectively analyzes the underlying reasons for refusal. We validate the effectiveness of agent-based checkers under a manually designed configuration and demonstrate that, with adaptive generative settings, large reasoning models can successfully perform abstention tasks. The performance of adaptive checkers further underscores the framework's scalability, demonstrating its adaptability to cases where questions are unanswerable due to more complex document-related issues.

6 Limitations

We acknowledge several potential limitations that offer avenues for future research.

First, the multi-agent collaborative process, by its nature, introduces additional computational overhead and latency compared to single-pass generation models. Although this trade-off is justified by the gains in reliability and interpretability, future work could explore optimization strategies, such as parallelizing checker queries or utilizing smaller, task-specific models for the checker agents to enhance efficiency.

Second, our experiments are conducted on the newly constructed DIRA dataset, which is designed to cover three primary types of document-level issues. While this provides a controlled environment for evaluation, real-world scenarios may present more complex or subtle challenges. Future research could focus on expanding the dataset to include a wider variety of nuanced failure modes, further testing the framework's generalization capabilities.

Finally, the performance of A2MAC, especially the adaptive variant, is inherently linked to the reasoning capabilities of the underlying large language models. As these foundational models continue to evolve, the effectiveness and autonomy of the collaborative agents are expected to improve. Continuously evaluating the framework with newer and more powerful LLMs will be essential to track its long-term viability and performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. *Preprint*, arXiv:2407.12043.

Lu Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2024. Controlling risk of retrieval-augmented generation: A counterfactual prompting framework. *Preprint*, arXiv:2409.16146.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2024a. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*.

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and 1 others. 2024b. Learning diverse attacks on large language models for robust red-teaming and safety tuning. *arXiv preprint arXiv:2405.18540*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

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

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. Predict: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. *arXiv preprint arXiv:2402.11889*.

A Additional Experimental Results

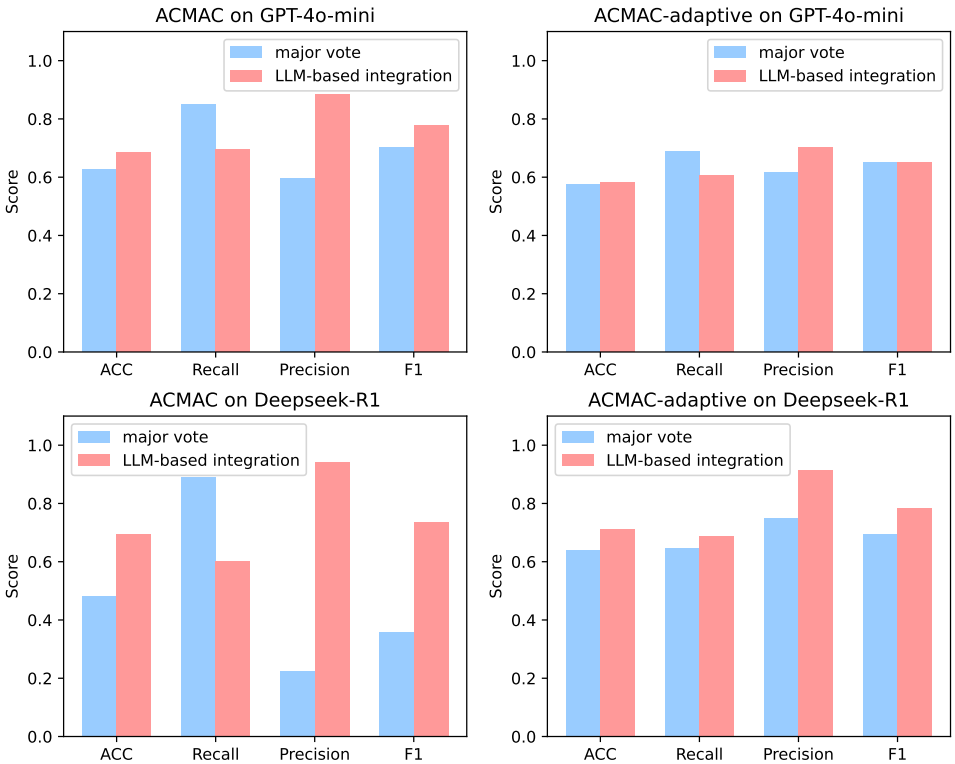


Figure 3: The comparison of major voting and LLM-based integration for the final decision, we list the results using the A2MAC and A2MAC-adaptive.

B prompts

B.1 Prompt for Quality Assessment

To evaluate the explanatory correctness of the refusal rationales (RQ3), we employed an LLM-as-a-judge approach. The following prompt was provided to the judge model to score the quality of a given explanation on a scale from 0 to 10 across four distinct criteria.

> Given a question and its corresponding reference document (answers must be based solely on the document, not the model's internal knowledge), the question has been deemed "unanswerable." Please evaluate the quality of the explanation provided for this judgment. For your evaluation, analyze the explanation based on the following four criteria, assigning a score from 0 to 10 for each:
> (1) Is the explanation clear and logically coherent? (10: perfectly clear and logical; 6: mostly clear with minor issues; 2: partially clear but confusing; 0: completely unclear.)
> (2) Does the explanation adequately justify why the question cannot be answered based on the document? (10: fully justified; 6: partially justified with gaps; 2: minimal justification; 0: no justification.)
> (3) Does the explanation accurately reference or cite the document's content? (10: fully accurate references; 6: mostly accurate with minor errors; 2: vague or incorrect references; 0: no references.)
> (4) Does the explanation include redundant or irrelevant information? (10: no redundancy; 6: minor redundancy; 2: significant redundancy; 0: entirely redundant or irrelevant.)
> Please List the scores for the four criteria in the format : Scores:[score1, score2, score3, score4], for example: Scores:[10, 9, 10, 7]

B.2 Prompt for Base Method

> Given the context: { } and the question: { }, please respond with 'Yes' if the answer should be declined due to the context being contradictory, irrelevant, or insufficient to form a complete reasoning chain; otherwise, respond with 'No'.

B.3 Prompt for Adaptive Generative Checkers

> Given the context: { } and the question: { }. You should find out three aspects that this question is related to in this format: I think to consider whether or not to reject the question, I still need to think about [aspect1,aspect2,aspect3].

C Details about Source Datasets

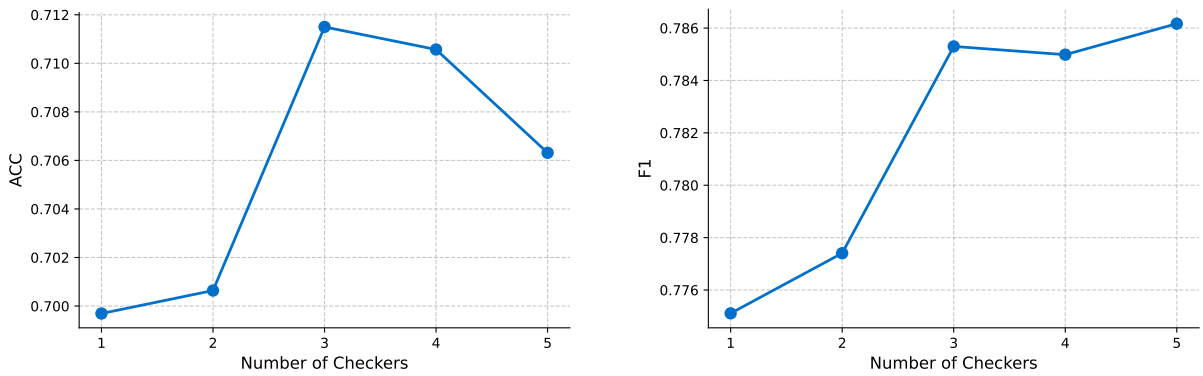
SQuAD2 (Rajpurkar et al., 2018): is a widely used benchmark for machine reading comprehension, extending SQuAD 1.0(Rajpurkar et al., 2016) by introducing unanswerable questions. Unanswerable questions are human-annotated and constructed to be semantically related but unresolvable.

Wikicontract (Hou et al., 2024): is a high-quality dataset extracted from Wikipedia, focused on evaluating the performance of large language models in handling real-world knowledge conflicts. It contains 253 labeled instances, each consisting of a question, contradictory paragraphs, and two corresponding answers.

Hotpotqa (Yang et al., 2018): is a dataset designed for multi-hop question answering. Each question is accompanied by 10 candidate documents, of which 2 contain the information required to answer the question, while the remaining 8 serve as distractors, challenging models' retrieval and information integration capabilities.

D The Effect of the Number of the Checkers

We analyze the impact of the number of checkers in the Checker Collaboration Module when adopting Strategy 3 in the Fast Response. The Fast Response Module and the Abstention Decision Module employ Deepseek-R1, and the Checker Collaboration Module utilizes GPT-4o-mini. The number of checkers is set to [1, 2, 3, 4, 5], and performance is evaluated using F1 and ACC metrics. As shown in Figure 4, the F1 score initially increases with the number of checkers but then remains relatively stable. In contrast, the ACC score exhibits a slight decline when the number of checkers exceeds three. This suggests that, under the Adaptive Generative Checkers setting, an appropriately higher number of checkers can enhance model reasoning. However, an excessive number of checkers may lead to information overload, causing critical information to be obscured and ultimately degrading model performance.



(a) ACC across different numbers of checkers

(b) F1 across different numbers of checkers

Figure 4: The effect of the number of the checkers for A2MAC-adaptive.

E Theoretical Completeness of the Three Aspects

The issues of *irrelevance*, *insufficiency*, and *contradiction* together provide a complete characterization of document-level errors in RAG retrieval.

Let q denote the query, a the correct answer, and $D = \{d_1, \ldots, d_n\}$ the retrieved documents. Let $\mathrm{Cl}(D)$ be the logical closure of D, and M_a the minimal set of premises sufficient to infer a. Then, for any retrieved set D, three mutually exclusive failure cases arise:

1. **Irrelevant:** $\mathrm{Cl}(D) \cap M_a = \varnothing$. The documents contain no premises useful for deriving the answer.
2. **Insufficient:** $\mathrm{Cl}(D) \cap M_a \neq \varnothing$ but $\mathrm{Cl}(D) \nvdash a$. Some relevant evidence exists but key premises are missing for a complete inference.
3. **Contradictory:** $\mathrm{Cl}(D) \vdash \neg a$ or $\mathrm{Cl}(D)$ is inconsistent. The retrieved content provides conflicting evidence that prevents a reliable answer.

These three cases exhaust all possible deviations from the ideal retrieval set D^* that would entail a. Other observable problems (e.g., redundancy, noise, or ambiguity) can be reduced to one of these categories— either providing no useful evidence (irrelevant), incomplete support (insufficient), or conflicting information (contradictory). Hence, the triad forms a minimal and exhaustive taxonomy of document-level failure modes in RAG.

F Case Study Example

To provide a qualitative illustration of our method's performance (RQ4), this section presents a detailed case study from the DIRA dataset. The example chosen is one where the provided document is insufficient to definitively answer the question. The table below compares the direct (and incorrect) answers from baseline methods with the detailed, step-by-step reasoning processes of our A2MAC and A2MAC-adaptive frameworks, demonstrating their ability to correctly identify the insufficiency and abstain with a clear rationale.

Table 5: An example (No.351) from dataset DIRA, to analyze the responses of different methods when a refusal is required due to documents retrieved by RAG, the large model used here is DeepSeek-R1.

Questions: What area has the most diversity of anywhere in the country? **Documents:** Southern California consists of one of the more varied collections of geologic, topographic, and natural ecosystem landscapes in a diversity outnumbering other major regions in the state and country. The region spans from Pacific Ocean islands, shorelines, beaches, and coastal plains, through the Transverse and Peninsular Ranges with their peaks, into the large and small interior valleys, to the vast deserts of California. **Does the question should be answered or rejected:** rejected.
Base answer: Southern California.(×)
Priori: Southern California.(×)
Calibration: Probability: 0.85 (answered).(×)
RC-RAG: The answer is directly supported by the explicit statement in the passage: *""Southern California has a diversity outnumbering other major regions in the state and country.""* This evidence clearly identifies Southern California as the most diverse area, aligning with the response. The prior uncertainty likely stems from misinterpreting the feedback, but the passage unequivocally validates the answer." (answered).(×)
A2MAC: contradicition checker: rejected. relevance checker: answered. sufficiency checker: rejected.

Looking at the context, it states that Southern California has a varied collection of geologic, topographic, and ecosystems, ""outnumbering other major regions in the state and country."" It lists various features from Pacific islands to deserts. The key here is whether the context confirms that Southern California is the *most* diverse in the country. The context says ""one of the more varied"" and ""diversity outnumbering other major regions,"" but does it explicitly state it's the *most* diverse in the entire country? It might imply that, but the wording is a bit vague. Also, there's no direct comparison to other specific regions outside the state. The user's question is about the entire country, so even if Southern California is more diverse than other regions in the state, does that translate to the whole country? The context doesn't provide data comparing it to all other U.S. regions.
Insufficient context would mean there's not enough info to form a complete reasoning chain. Here, the context does suggest Southern California is highly diverse, but without explicit confirmation that it's the most diverse nationwide, the answer might need more data. However, the context does say ""outnumbering other major regions in the state and country,"" which could be interpreted as it being the most diverse. So maybe it's sufficient. But the example given by the user in the previous step had a ""Yes"" because the context didn't compare to other areas. Wait, in the previous similar question, the assistant answered ""Yes"" because the context didn't compare. But in this case, the context does mention ""outnumbering other major regions in the state and country,"" which might be enough. Hmm. But the exact phrasing is ""one of the more varied collections"" which might not definitively claim it's the absolute most diverse. So there's ambiguity. Therefore, the context might be insufficient to conclusively state Southern California is the most diverse in the country. Hence, the answer should be declined due to insufficient context.(rejected).(✓)

A2MAC-adaptive:
checker1(scope of comparison (regional vs. national)): rejected.
checker2(criteria for "diversity" (geologic, topographic, ecosystem)): rejected.
checker3(context's assertion of Southern California ranking): rejected.
the context asserts Southern California has ""one of the more varied collections"" but does not explicitly state it is *the most* diverse area in the country, leaving room for comparison uncertainty.","Okay, let me break this down step by step. The user is asking whether Southern California has the most diversity of anywhere in the country based on the given context.
The context says Southern California is ""one of the more varied collections"" and mentions it outnumbers other major regions in the state and country. It lists various landscapes like Pacific Ocean islands, coastal plains, mountain ranges, valleys, and deserts.
The key phrase here is ""one of the more varied,"" which implies it's among the top but not necessarily the absolute number one. The question is asking for the area with the ""most diversity,"" which requires the top rank. Since the context doesn't explicitly state Southern California is the most diverse, just that it's highly varied compared to other regions, it can't confirm the highest ranking.
Additionally, the context compares it to ""other major regions,"" but doesn't specify that all possible areas in the country have been considered. There might be other regions not mentioned that could have equal or greater diversity.
So, because the context doesn't definitively rank Southern California as the most diverse nationwide, the answer should be declined.(rejected).(✓)

G Discussion on the Effectiveness-Efficiency Trade-off

We acknowledge that our proposed method requires a four-fold increase in computational cost compared to a single-LLM approach. However, this increase is not a simple repetition, but rather a deliberate investment in a fundamentally more robust and trustworthy decision-making process.

A single-LLM query is analogous to consulting a single, brilliant general practitioner: fast and often correct, but susceptible to single-point failures and perspective limitations. In contrast, our framework models a specialist consultation process. The initial call acts as a triage. The subsequent three calls

function as parallel, independent specialists (e.g., for contradiction, relevance, and sufficiency), each performing a deep, focused analysis in their domain. The final call acts as a consensus panel, synthesizing the specialist reports to form a cross-validated and highly reliable final judgment.

Therefore, the cost is a direct trade-off for qualitative gains that are critical in high-stakes scenarios and strictly grounded tasks: depth of analysis, robustness against error, and transparent attributability. For applications where accuracy and reliability are paramount, this structured, multi-perspective validation process provides a value that a single, monolithic query cannot.