When to Trust Context: Self-Reflective Debates for Context Reliability

Anonymous ACL submission

Abstract

Large language models frequently encounter conflicts between their parametric knowledge and contextual input, often resulting in factual inconsistencies or hallucinations. We propose Self-Reflective Debate for Contextual Reliability (SR-DCR), a lightweight framework that integrates token-level self-confidence with an asymmetric multi-agent debate to adjudicate such conflicts. A critic, deprived of context, challenges a defender who argues from the given passage; a judge model evaluates the debate and determines the context's reliability. The final answer is selected by combining the verdict with model confidence. Experiments on the ClashEval benchmark show that SR-DCR consistently improves robustness to misleading context while recovering accuracy on trustworthy inputs, outperforming both classical debate and confidence-only baselines with minimal computational overhead.

1 Introduction

Large language models (LLMs) (Wang et al., 2024; Guo et al., 2024) reason using two distinct sources of knowledge: *parametric knowledge*, stored in model weights, and *contextual input*, provided at inference time. This dual-source capability grants LLMs remarkable flexibility, but also exposes a critical vulnerability: when internal priors and external context conflict, models frequently generate hallucinated or factually inconsistent outputs. These failures range from parroting erroneous passages (e.g., "people eat rocks") to relying on outdated beliefs (e.g., mislocating the Eiffel Tower in Rome).

Such errors highlight the absence of a principled mechanism for reconciling contradictions between knowledge sources. Empirical studies show that LLMs often defer to memorized facts even when these subtly contradict the context (Brown et al., 2020; Longpre et al., 2021). Retrieval-augmented generation (RAG) systems compound the problem by amplifying misleading context when it appears superficially fluent or plausible (Niu et al., 2023). Reliance on either source in isolation leads to brittle and unpredictable behavior.

Recent work has explored token-level confidence as a signal for epistemic reliability. For example, Wu et al. (2024) compare token probabilities with and without context to select the more trustworthy response. More broadly, output confidence—estimated from log-probabilities—correlates with familiarity: confident predictions tend to reflect well-known facts, while uncertain answers often correspond to ambiguous or unfamiliar queries (Wang et al., 2022).

In parallel, *multi-agent debate* (MAD) (Du et al., 2023; Chan et al., 2023; Liang et al., 2023; Kenton et al., 2024; Lang et al., 2025; Agarwal and Khanna, 2025) has emerged as a promising oversight mechanism. By prompting agents to critique each other, it enables deliberation, mitigates hallucinations, and often allows even weak judges to identify correct answers (Irving et al., 2018; Michael et al., 2023). However, existing MAD frameworks typically assume consistent inputs across agents, and rarely examine how debates unfold when internal priors and external context offer conflicting evidence.

Our Contributions. This work addresses the open challenge of resolving internal–external knowledge conflict in LLMs. We contribute:

1. A systematic analysis of model behavior under graded contextual perturbations, revealing how token-level confidence and factual familiarity shape LLM preferences for internal versus external knowledge.

2. An evaluation of standard MAD setups under adversarial conflict, showing that symmetric debate alone often fails to arbitrate between incompatible sources.

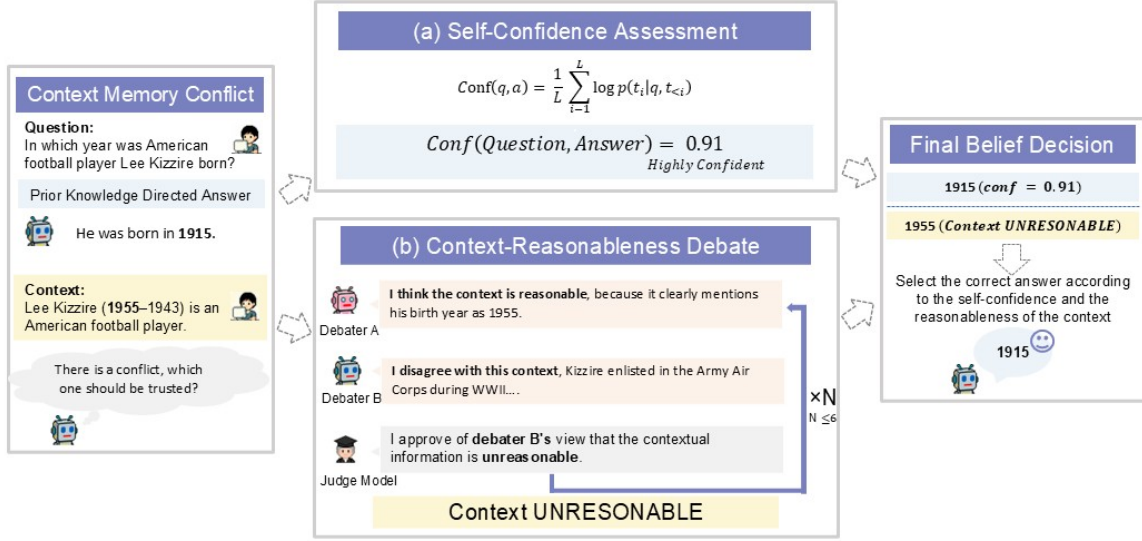3. A novel framework, **Self-Reflective Debate for**

Figure 1: Overview of **SR-DCR** (Self-Reflective Debate for Contextual Reliability), a debate-driven framework for resolving conflicts between parametric priors and contextual evidence. When the model's prior belief conflicts with the external passage (left), SR-DCR evaluates two key signals: (a) the self-confidence of the zero-context answer $\hat{a}_{\text{PRIOR}}$; and (b) the context's reasonableness, determined via an asymmetric debate between two LLM agents. A judge model monitors the debate over N rounds and issues a verdict on the trustworthiness of the context. The final answer is selected by jointly considering both the model's self-confidence and the judged context reliability (right).

Contextual Reliability (SR-DCR), illustrated in Fig. 1. SR-DCR combines token-level self-confidence with an **Asymmetric Context Verification Debate (ACVD)**, in which one agent defends the context and another, deprived of it, argues from prior knowledge. A judge resolves the debate over multiple rounds, and a final decision rule integrates both context reliability and confidence to select the answer or abstain. This framework improves robustness and factual accuracy across multiple QA benchmarks.

2 Preliminaries

2.1 Background of Knowledge Conflict

Problem Statement. We investigate how LLMs reconcile conflicting signals between their internal factual priors and externally provided context. Specifically, each instance is defined as a triple (q, a, c), where:

- q is a question generated from a factual *(subject, relation, object)* tuple.

- a is the correct answer derived from the object.

- c is an evidence passage intended to support or contradict a.

The model is given the pair (q, c) and tasked with answering the question q. Our goal is to evaluate whether it can correctly predict a, even when the passage c contains misleading or conflicting information.

To probe the model's robustness, we introduce targeted perturbations to the answer-bearing entity in c while keeping both q and a fixed. These perturbations are constructed at four increasing levels of contradiction—*subtle*, *mild*, *moderate*, and *blatant*—to test how the model balances contextual evidence against its parametric knowledge.

Illustrative Example. In the WIKIPEDIA YEARS domain, consider the question: *"In which year was the census conducted that reported the population of Lukhi village in Iran as 35?"* The correct answer a is 2006, based on the underlying knowledge tuple. We then modify the corresponding passage c to simulate contradiction:

- A **moderate** perturbation replaces the year with 1966, creating a subtle conflict.

- A **blatant** perturbation changes it to 2106, introducing an implausible inconsistency.

The remainder of the passage remains unchanged. These graded interventions enable controlled analysis of how LLMs resolve factual conflicts under varying degrees of contextual reliability. More examples are put in Appendix B.

2

2.2 Assessing Knowledge Recall

To estimate whether a model *knows* the correct answer a to a question q independently of any supporting passage, we employ a sampling-based probing method inspired by SLICK (Gekhman et al., 2024). The core idea is to gauge how frequently a model recalls the correct answer when presented with the question alone, using no external context.

Sampling Procedure. For each QA pair (q, a), we generate $N = 32$ completions from the base model at a fixed temperature $T = 0.5$, using a 4-shot prompt for few-shot conditioning. Then a group of predictions is obtained as $\{\hat{a}_1, ..., \hat{a}_N\}$, and the *sampled accuracy* is defined as:

$$\text{Acc}\left(q, a, \{\hat{a}_i\}_{i=1}^N\right) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{a}_i = a\right],$$

where $\mathbb{1}[\cdot]$ is the indicator function. This provides a robust estimate of the model's prior knowledge of a. This is distinct from *self-confidence* (discussed later), which reflects belief in a single prediction.

Knowledge Categorization. Based on sampled accuracy, we categorize each (q, a) pair into one of four *knowledge tiers* (see Tab. 3):

- **Highly Known:** $\text{Acc}\left(q, a, \{\hat{a}_i\}_{i=1}^N\right) \geq 0.85$ — strong and consistent recall,

- **Maybe Known:** $0.30 \leq \text{Acc}\left(q, a, \{\hat{a}_i\}_{i=1}^N\right) < 0.85$ — partial or inconsistent recall,

- **Weakly Known:** $0 < \text{Acc}\left(q, a, \{\hat{a}_i\}_{i=1}^N\right) < 0.30$ — sporadic or weak recall,

- **Unknown:** $\text{Acc}\left(q, a, \{\hat{a}_i\}_{i=1}^N\right) = 0$ — no evidence of prior recall.

This taxonomy allows us to stratify the model's factual knowledge without relying on any auxiliary retrieval or context conditioning.

2.3 Quantifying Self-Confidence

Definition. In addition to knowledge qualification, which reflects whether a model can recall a fact, we also measure how strongly the model *believes* in its answer when generating a response. We define a model's **self-confidence** for a prediction $\hat{a} = \langle t_1, \ldots, t_L \rangle$ to question q as the average log-probability assigned to the answer tokens (Wang et al., 2022):

$$p_\theta(\hat{a} \mid q) = \frac{1}{L} \sum_{i=1}^{L} \log p(t_i \mid q, t_{<i}).$$

This score is derived from a single greedy prediction conditioned on a fixed few-shot prompt without temperature sampling. A higher confidence value implies greater internal belief in the generated answer.

Distinction from Knowledge Recall. Unlike sampled accuracy, which captures the empirical frequency of correct recall across many completions, self-confidence reflects the strength of belief in a single response. These two axes—*knowability* and *belief strength*—may diverge: a model might confidently predict an incorrect answer (high confidence, low recall), or may inconsistently predict a correct one (low confidence, high recall).

3 Method

We introduce *Self-Reflective Debate for Contextual Reliability (SR-DCR)*, a framework that combines asymmetric MAD with self-confidence–aware reasoning to evaluate context trustworthiness and determine final answers under conflicting information. SR-DCR unfolds in three stages:

1. **Asymmetric Context Verification Debate (ACVD)**: Agents debate the reliability of the context from asymmetric viewpoints.

2. **Self-Confidence Estimation**: The model predicts a zero-context answer $\hat{a}_{\text{PRIOR}}$ and computes its confidence score.

3. **Final Answer Selection**: A decision rule chooses between $\hat{a}_{\text{CTX}}$, $\hat{a}_{\text{PRIOR}}$, or ABSTAIN, based on the context verdict and confidence level.

3.1 Asymmetric Context Verification Debate

To assess the reliability of external context passages, we introduce *ACVD*—a structured MAD framework that builds on prior MAD approaches (Du et al., 2023; Michael et al., 2023; Lang et al., 2025), but introduces informational asymmetry between agents. Unlike standard MAD setups where both agents receive the same input, ACVD withholds the context from one participant, enabling an adversarial test of whether the passage contributes trustworthy information.

We instantiate three roles:

- **Defender (Agent A)** sees (q, c) and defends the context-based answer $\hat{a}_{\text{CTX}} := f_\theta(q, c)$, arguing that passage c is coherent and helpful.

- **Critic (Agent B)** sees only the question q and supports the prior answer $\hat{a}_{\text{PRIOR}}$, arguing that context c is misleading or fabricated.

- **Judge (Agent C)** observes the full debate transcript up to round r, issuing a verdict $\mathcal{V}^{(r)} \in \{\text{REASONABLE}, \text{UNREASONABLE}\}$ based on the evolving dialogue.

The debate proceeds over $R = 6$ rounds. In round $r = 0$, both Defender and Critic submit opening statements. In subsequent rounds $r \geq 1$, the Critic speaks first and the Defender replies. Each agent can access the full transcript $\mathcal{T}^{(r)}$ up to that round and may quote or challenge prior arguments. The final verdict is defined as the earliest stabilized outcome:

$$\mathcal{V}^* := \mathcal{V}^{(r^*)}, \text{ where } \mathcal{V}^{(r)} = \mathcal{V}^{(r+1)} = \cdots = \mathcal{V}^{(R)}.$$

ACVD allows us to assess not only whether models use context but whether they can recognize when context is verifiably beneficial or harmful—using a setup that foregrounds the asymmetry of information and the epistemic role of debate.

3.2 Self-Confidence–Aware Belief Update

Given a question q, the model generates a prior prediction $\hat{a}_{\text{PRIOR}}$ under zero-context conditions. Let $p_\theta(a \mid q)$ denote the model's predictive distribution. The self-confidence score is defined as:

$$\text{Conf}(q, \hat{a}_{\text{PRIOR}}) := p_\theta(\hat{a}_{\text{PRIOR}} \mid q),$$

computed via normalized log-probability over token sequences. We threshold this score at $\tau = 0.90$:

$$\text{CONF}(q) = \begin{cases} \text{HIGH}, & \text{if } \text{Conf}(q, \hat{a}_{\text{PRIOR}}) \geq \tau, \\ \text{LOW}, & \text{otherwise}. \end{cases}$$

This binary confidence label is cached per instance and used to inform final decisions without repeated inference.

Finally, the model selects its answer using a gating decision that combines the verdict $\mathcal{V}^*$ and confidence level:

$$\hat{a}_{\text{FINAL}} = \begin{cases} \hat{a}_{\text{CTX}}, & \text{if } \mathcal{V}^* = \text{REASONABLE}, \\ \hat{a}_{\text{PRIOR}}, & \text{if } \mathcal{V}^* = \text{UNREASONABLE} \\ & \wedge \text{CONFIDENCE}(q) = \text{HIGH}, \\ \text{ABSTAIN}, & \text{otherwise}. \end{cases}$$

This decision rule ensures that the model trusts external context only when it is explicitly judged to be reliable, and otherwise relies on its internal belief only if that belief is expressed with high confidence.
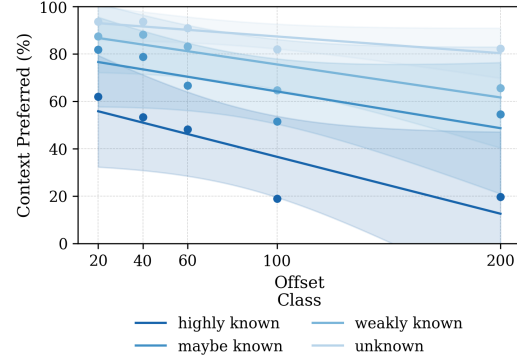

Figure 2: Context preference rates of the few-shot baseline across perturbation levels and knowledge categories, where perturbation levels are represented by offset magnitudes: subtle (offset = 20), mild (40), moderate (60), and blatant (100 & 200).

4 Experiments

Dataset and Implementations. We utilize *ClashEval* (Wu et al., 2024), a benchmark specifically constructed to study factual conflicts between an LLM's prior knowledge and contradictory external context. We adopt the official development/test splits and construct a controlled subset by subsampling and stratifying perturbation levels. All models are evaluated using the *exact match (EM)* metric: a prediction $\hat{a}$ is correct only if it exactly matches the ground truth a. More experimental details are in Appendix A, and additional results are available in Appendix C.

Baselines. Five strong baselines and the Golden Baseline are selected: (1) *Few-shot prompting* provides the model with 4 randomly selected exemplars along with the context (Brown et al., 2020). (2) *Self-Ask* (Press et al., 2022) prompts the model to generate intermediate sub-questions before answering the main query. (3) *RCI* (Kim et al., 2023) requires the model to iteratively critique and improve its output. (4) *Judge Debate* (Liang et al., 2023) classical MAD framework involves two debaters and a judge: one debater receives and supports the context-derived answer as initial stand, while the other supports prior-knowledge-derived answer as initial stand. (5) *Naive Debate* (Du et al., 2023) asks multiple LLMs to propose individual answers and engage in multi-round debate, ultimately converging on a final answer. (6) *Golden Baseline* represents a theoretical upper bound in which models only trust the unperturbed, ground-truth context. When facing perturbed contexts, models rely entirely on their prior knowledge. This setup simulates an ideal scenario where LLMs perfectly

discern context correctness and serves as a reference for their best achievable performance under our experimental setting.

4.1 Few-Shot Context Reliance under Perturbation

We begin by investigating how a standard few-shot prompting strategy behaves in the presence of increasingly misleading context. We prompt the model with several random in-domain exemplars and present it with the input pair (q, c), where c contains a perturbed version of the ground-truth answer. No adversarial interaction is used at this stage. For each example, we record whether the model's answer aligns with the (incorrect) contextual claim—i.e., whether it "prefers" the context over its prior knowledge. We then aggregate this context preference rate across different offset levels and stratify by knowledge category.

Fig. 2 shows that *Highly-Known* items rapidly reject perturbed context with increasing offsets. In contrast, *Unknown* and *Weakly Known* examples exhibit persistent reliance on context, remaining $> 70\%$ agreement even under extreme perturbations. This overreliance on erroneous context in the absence of prior certainty motivates our SR-DCR design.

4.2 Linking Self-Confidence and Prior Knowledge

Next, we analyze the relationship between two measures of model familiarity: (i) sampling-based knowledge categories and (ii) single-shot self-confidence scores $\mathrm{Conf}(q, \hat{a})$. We conduct this analysis across 5 LLMs: GPT-3.5-Turbo, GPT-4o (OpenAI, 2023), Claude Sonnet 3.7, Claude Haiku 3.5, and Llama 3.3-70B (Grattafiori et al., 2024).

For each model, we first compute $C(\tau) = \{(q, a) \mid \mathrm{Conf}(q, \hat{a}) \geq \tau\}$ and define:

$$P(\text{Highly-Known} \mid \mathrm{Conf} \geq \tau) = \frac{|S_{\text{Highly}} \cap C(\tau)|}{|C(\tau)|}.$$

where S_{Highly} is the set of QA pairs categorized as Highly-Known.

Fig. 3 and 7 show that across all models, this conditional probability exceeds 0.88 once $\tau \geq 0.90$ and surpasses 0.95 at $\tau \geq 0.95$, confirming that high-confidence predictions are strong indicators of prior knowledge. In contrast, confidence below 0.70 yields poor overlap with Highly-Known pairs.

These findings validate our key hypothesis: when $\mathrm{Conf}(q, \hat{a})$ crosses a threshold (e.g., 0.90),
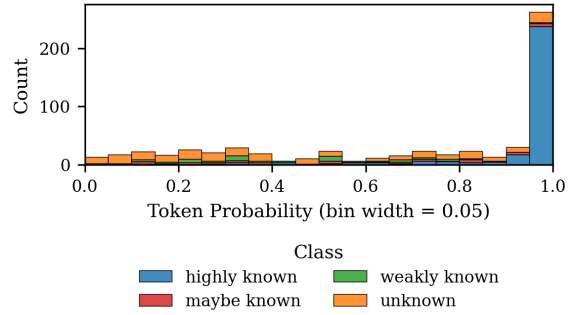

Figure 3: The distribution of self-confidence scores of GPT-4o, colored by sampling-based knowledge categories. Highly-Known instances cluster at the high end of the distribution (≥ 0.95), while Unknown examples dominate the low-confidence region, demonstrating self-confidence's efficacy as a proxy for factual recall.
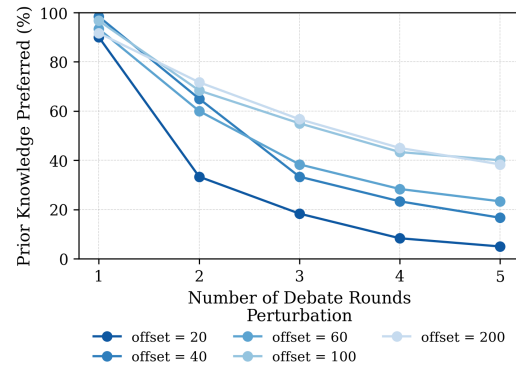

Figure 4: Prior Knowledge preference rate in judge debate on Sonnet 3.7 across debate rounds, stratified by perturbation levels (represented by offset). Judges exhibit strong prior bias in early rounds, but shift toward context over rounds.

the model is highly likely to "know" the answer. We use this insight to guide dynamic routing in SR-DCR.

4.3 Evaluating Multi-Agent Debate under Knowledge Conflict

We compare the Classical MAD framework (Judge Debate) against five baselines with standard and perturbed contexts. Results of 1, 3, 5 rounds of debate on five different models are shown in Tab. 1, from which we draw several conclusions.

Insight 1 — Debate rounds shift judgment from prior to context. We observe a consistent trend in which early rounds of judge debate favor prior knowledge answers, while longer debates increasingly support context-derived responses. At $r = 1$, judges tend to reject the context. With GPT-4o, judge debate ($r = 1$) achieves only 70% accuracy on standard (unperturbed) context questions,

5

Table 1: Accuracy of each method across standard and perturbed contexts. Best results per column within each model block are **bold**, second-best are underlined.

Model	Method	Standard Context	Perturbed Context	Overall
GPT 3.5 Turbo	Few-Shots	**99.30%**	9.00%	54.15%
	Self-Ask	95.00%	14.66%	54.83%
	RCI	96.70%	12.32%	54.51%
	Judge Debate (Round 1)	36.67%	<u>23.99%</u>	30.33%
	Judge Debate (Round 3)	83.30%	14.02%	48.66%
	Judge Debate (Round 5)	90.33%	17.00%	53.66%
	Naive Debate	<u>98.33%</u>	20.33%	<u>59.33%</u>
	SR-DCR	95.70%	**29.66%**	**62.68%**
	Golden Baseline	98.67%	31.67%	65.17%
GPT 4o	Few-Shots	**99.33%**	42.27%	70.80%
	Self-Ask	97.00%	31.67%	64.33%
	RCI	91.67%	22.33%	57.00%
	Judge Debate (Round 1)	70.00%	**57.33%**	63.67%
	Judge Debate (Round 3)	96.00%	45.67%	<u>70.83%</u>
	Judge Debate (Round 5)	96.00%	45.33%	70.67%
	Naive Debate	<u>97.33%</u>	44.33%	<u>70.83%</u>
	SR-DCR	94.67%	<u>54.51%</u>	**74.59%**
	Golden Baseline	99.00%	50.00%	74.50%
Claude Haiku 3.5	Few-Shots	**99.67%**	16.33%	58.00%
	Self-Ask	98.00%	21.67%	59.83%
	RCI	93.33%	22.00%	57.66%
	Judge Debate (Round 1)	55.67%	26.33%	41.00%
	Judge Debate (Round 3)	94.67%	21.67%	58.17%
	Judge Debate (Round 5)	97.00%	21.00%	59.00%
	Naive Debate	<u>98.67%</u>	<u>29.67%</u>	<u>64.17%</u>
	SR-DCR	95.33%	**49.89%**	**72.61%**
	Golden Baseline	99.67%	48.00%	73.83%
Claude Sonnet 3.7	Few-Shots	<u>98.67%</u>	43.04%	70.85%
	Self-Ask	97.00%	45.33%	<u>71.17%</u>
	RCI	95.00%	22.66%	58.83%
	Judge Debate (Round 1)	61.67%	**50.33%**	56.00%
	Judge Debate (Round 3)	<u>98.67%</u>	30.00%	64.33%
	Judge Debate (Round 5)	**99.33%**	24.67%	62.00%
	Naive Debate	98.00%	47.34%	**72.67%**
	SR-DCR	91.04%	<u>47.94%</u>	69.76%
	Golden Baseline	99.67%	47.67%	73.67%
Llama 3.3-70B	Few-Shots	**98.67%**	9.67%	54.17%
	Self-Ask	<u>98.33%</u>	23.67%	<u>61.00%</u>
	RCI	94.96%	25.00%	59.98%
	Judge Debate (Round 1)	56.67%	35.66%	46.17%
	Judge Debate (Round 3)	92.33%	**28.33%**	60.33%
	Judge Debate (Round 5)	93.33%	25.67%	59.49%
	Naive Debate	**98.67%**	22.33%	60.50%
	SR-DCR	94.97%	<u>28.39%</u>	**61.67%**
	Golden Baseline	99.00%	33.33%	66.17%

compared to $> 90\%$ with other baselines. This skepticism toward context provides strong protection against misinformation: on perturbed inputs, GPT-4o's judge debate ($r = 1$) achieves significantly higher accuracy (57.33%) than others, with Self-Ask and RCI trailing far behind at 31.7% and 22.3%, respectively.

As the number of debate rounds increases, this behavior gradually reverses. Judges become more willing to trust contextual evidence—especially when it is accurate. On standard (unperturbed) examples, accuracy rises steadily with more rounds, reaching nearly 100% by $r = 3$. This benefit is not without tradeoffs. While longer debates improve acceptance of correct context, they also reduce robustness to misleading information. accuracy on perturbed inputs declines as the number of rounds increases: for GPT-4o, accuracy on perturbations drops from 57.3% at $n = 1$ to 45.3% at $n = 5$.

Insight 2 — Longer debates increase context reliance, and reduce robustness to small inconsistencies. A more detailed analysis by perturbation level reveals that this degradation is not uniform. From Fig. 4 and Fig. 9, we can see that for GPT-4o, sonnet 3.7, and LLaMA 70B, data points with a lower perturbation level (e.g., subtle (offset = 20)) show a steeper decrease in accuracy, suggesting that minor inconsistencies are more easily overlooked as debates grow longer. In contrast, large perturbations (e.g., blatant (offset >= 100))

Table 2: Performance of different methods under a group of prevalent LLMs across varying offset levels. Best and second-best results are shown in **bold** and underlined, respectively.

Model	Method	Offset 20	Offset 40	Offset 60	Offset 100	Offset 200	Overall
GPT 3.5 Turbo	Few-Shots	1.70	8.30	11.70	8.30	15.00	9.00
	Self-Ask	1.00	13.30	18.30	23.30	18.30	14.66
	RCI	8.30	8.30	15.00	16.70	13.30	12.32
	Judge Debate (Round 1)	25.00	21.67	28.30	18.33	26.67	23.99
	Judge Debate (Round 3)	5.00	11.70	20.00	11.70	21.70	14.02
	Judge Debate (Round 5)	10.00	13.33	16.67	20.00	25.00	17.00
	Naive Debate	8.33	13.33	15.00	25.00	**40.00**	20.33
	SR-DCR	**26.70**	**25.00**	**30.00**	28.30	38.30	**29.66**
	Golden Baseline	31.67	30.00	33.33	21.67	41.67	31.67
GPT 4o	Few-Shots	23.33	31.00	38.33	57.00	61.67	42.27
	Self-Ask	28.33	25.00	21.67	40.00	43.33	31.67
	RCI	15.00	16.67	21.67	23.33	35.00	22.33
	Judge Debate (Round 1)	**56.67**	**53.33**	**58.33**	**58.33**	60.00	**57.33**
	Judge Debate (Round 3)	35.00	45.00	48.33	48.33	51.67	45.67
	Judge Debate (Round 5)	33.33	41.67	48.33	48.33	55.00	45.33
	Naive Debate	26.67	33.33	36.67	53.33	**71.67**	44.33
	SR-DCR	47.46	47.37	52.73	55.00	70.00	54.51
	Golden Baseline	50.00	50.00	53.33	46.67	50.00	50.00
Claude Haiku 3.5	Few-Shots	8.33	18.33	6.67	23.33	25.00	16.33
	Self-Ask	10.00	16.67	25.00	25.00	31.67	21.67
	RCI	8.33	13.33	15.00	30.00	43.33	22.00
	Judge Debate (Round 1)	28.33	26.67	25.00	18.33	33.33	26.33
	Judge Debate (Round 3)	15.00	20.00	23.33	20.00	30.00	21.67
	Judge Debate (Round 5)	15.00	15.00	21.67	21.67	31.67	21.00
	Naive Debate	18.33	21.67	26.67	35.00	46.67	29.67
	SR-DCR	**55.56**	**41.38**	**53.45**	**47.27**	**51.79**	**49.89**
	Golden Baseline	50.00	41.67	55.00	43.33	50.00	48.00
Claude Sonnet 3.7	Few-Shots	26.67	31.67	38.33	55.00	63.52	43.04
	Self-Ask	25.00	35.00	41.67	63.33	61.67	45.33
	RCI	13.33	18.33	20.00	23.33	38.33	22.66
	Judge Debate (Round 1)	**46.67**	**43.33**	**56.67**	48.33	56.67	**50.33**
	Judge Debate (Round 3)	15.00	18.33	30.00	41.67	45.00	30.00
	Judge Debate (Round 5)	3.33	10.00	23.33	40.00	46.67	24.67
	Naive Debate	21.67	31.67	45.00	**66.67**	**71.67**	47.34
	SR-DCR	33.33	40.00	43.63	55.00	68.42	47.95
	Golden Baseline	50.00	43.33	53.33	41.67	50.00	47.67
Llama 3.3-70B	Few-Shots	5.00	3.33	13.33	11.67	15.00	9.67
	Self-Ask	15.00	18.33	21.67	28.33	35.00	23.67
	RCI	13.33	18.33	25.00	31.67	36.67	25.00
	Judge Debate (Round 1)	**35.00**	**40.00**	33.33	28.33	**41.67**	**35.67**
	Judge Debate (Round 3)	16.67	18.33	31.67	35.00	40.00	28.33
	Judge Debate (Round 5)	15.00	15.00	**33.33**	31.67	33.33	25.67
	Naive Debate	11.67	13.33	20.00	31.67	35.00	22.33
	SR-DCR	20.34	25.00	30.51	32.76	33.33	28.39
	Golden Baseline	50.00	41.67	55.00	43.33	50.00	48.00

remain consistently detectable: Their accuracy remains relatively stable across all debate rounds. This indicates that judge debate is effective in rejecting an obviously flawed context but becomes more vulnerable to subtle misinformation as it attempts to reconcile competing claims.

These observations reveal a tension in Judge Debate: short debates reinforce prior bias, while longer debates risk accepting false context.

4.4 Evaluation of SR-DCR

Evaluation Setups. We assembled a composite testbed of 600 question–answer instances: a half use *standard contexts*, where the retrieved passage is correct; the other half employ *perturbed contexts*, simulating erroneous RAG outputs at 4 calibrated disturbance levels (subtle to blatant perturbations).

This split measures the ability to leverage valid context and the robustness to misleading excerpts.

Results. Tab. 1 reports EM across standard and perturbed contexts. Across all model families, we observe a consistent pattern: few-shot prompting and retrieval-based approaches (e.g., Self-Ask, RCI) perform well on standard contexts but degrade sharply under adversarial perturbation. For instance, GPT-3.5 Turbo's accuracy drops from 99.3% to 9.0% with Few-Shot, and from 95.0% to 14.7% with Self-Ask. Classical MAD (e.g., Judge Debate at Round 5) improves robustness but often sacrifices standard-context accuracy. For example, Judge Debate (R5) recovers perturbed accuracy to 17.0%, but standard accuracy falls to 90.3%.

SR-DCR consistently outperforms prior methods in handling conflicting information. On GPT-3.5 Turbo, it achieves 29.7% on perturbed inputs—nearly matching the golden baseline's 31.7%—while maintaining 95.7% on standard contexts. This yields an overall accuracy of 62.7%, a +3.4-point gain over Naive Debate, the best-performing baseline. On stronger models like GPT-4o and Haiku 3.5, SR-DCR matches or exceeds the performance of Naive Debate and classical MAD, with the largest gains observed in perturbed contexts: +10.2 points over Judge Debate (R5) on GPT-4o and +20.2 points over RCI on Haiku 3.5.

To better understand robustness under varying degrees of contradiction, Tab. 2 reports accuracy across perturbation offsets. On GPT-3.5 Turbo, SR-DCR achieves state-of-the-art performance at every offset level, including 26.7% at Offset 20 and 38.3% at Offset 200—far surpassing all baselines. Similarly, SR-DCR attains the best average perturbed accuracy on Sonnet (47.9%) and Haiku (49.9%), and ranks second only to Judge Debate (R1) on GPT-4o. This demonstrates SR-DCR's capacity to retain valid contextual information while resisting misleading perturbations.

We further analyze SR-DCR's ability to selectively trust valid contexts and reject corrupted ones. On GPT-3.5 Turbo, SR-DCR yields a true positive rate (standard context accuracy) of 95.7% and a true negative rate (perturbed context accuracy) of 29.7%. In contrast, Naive Debate achieves a higher standard-context rate (98.3%) but a lower rejection rate (20.3%), while Judge Debate (R5) performs worse on both fronts. SR-DCR thus achieves a more balanced trade-off, leveraging external context when warranted and falling back on internal knowledge when necessary.

Summary and Analysis. Empirically, this self-reflective control yields three key advantages: (i) a +7.7-point improvement in robustness to corrupted context, (ii) a +5.6-point recovery in accuracy on clean, context-consistent cases, and (iii) a 2× reduction in prior bias during extended debates. Crucially, SR-DCR achieves this improved arbitration between parametric and contextual knowledge with minimal overhead, requiring only one additional forward pass for confidence estimation and a debate that is already necessary for context evaluation.

Together, these results confirm SR-DCR as a principled and efficient solution for resolving factual conflicts, outperforming both confidence-based and debate-only strategies across models and perturbation settings.

5 Related Work

Knowledge Conflict. In in-context learning, conflicts between an LLM's internal knowledge and external context can cause interference (Jin et al., 2024). Models often over-rely on coherent external evidence even when it contradicts their memory (Xie et al., 2024), especially under low internal confidence (Xu et al., 2024; Chatziveroglou et al., 2025). To address this, prior work has proposed Knowledge-Aware Fine-Tuning (Li et al., 2023), opinion-based prompts and counterfactuals (Zhou et al., 2023), and fact duration prediction (Zhang and Choi, 2023). Some prioritize context (Li et al., 2023; Zhou et al., 2023; Zhang and Choi, 2023), while others favor memory (Hong et al., 2024). However, rigid faithfulness to either source is suboptimal. We extend this line by jointly varying perturbation strength and prior confidence to better arbitrate between conflicting knowledge sources.

Multi-Agent Debate. MAD has been shown to improve LLM reasoning and evaluation (Khan et al., 2024). Adversarial critique can enhance factuality, and aggregation strategies like RECONCILE (Chen et al., 2024) use confidence-weighted voting across agents. Prior work demonstrates that adversarial dialogue can reveal false assumptions (Du et al., 2023), improve robustness (Michael et al., 2023), and enhance factual accuracy even with weaker judges (Kenton et al., 2024). However, most MAD frameworks debate all queries uniformly, ignoring model confidence—sometimes leading to unnecessary correction when the model is already correct. To address this, we integrate self-confidence signals with context reasonableness and introduce a judge agent to yield more balanced and credible answers.

6 Conclusion

We presented SR-DCR, a self-reflective framework that combines model confidence with asymmetric debate to resolve conflicts between parametric knowledge and contextual input. By selectively trusting context when it is judged reliable—and deferring to priors when confidence is high—SR-DCR improves factual robustness and accuracy under perturbation, outperforming classical debate and confidence-only baselines. Limitations are in Appendix D.

References

Mahak Agarwal and Divyam Khanna. 2025. When persuasion overrides truth in multi-agent llm debates: Introducing a confidence-weighted persuasion override rate (cw-por). *arXiv preprint arXiv:2504.00374*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chi-Min Chan and 1 others. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Giannis Chatziveroglou, Richard Yun, and Maura Kelleher. 2025. Exploring llm reasoning through controlled prompt variations. *Preprint*, arXiv:2504.02111.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Zorik Gekhman and 1 others. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2404.14219*.

Taicheng Guo and 1 others. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495. Association for Computational Linguistics.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Zachary Kenton and 1 others. 2024. On scalable oversight with weak llms judging strong llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 75229–75276.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 23662–23733. PMLR. ArXiv:2402.06782.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. In *Advances in Neural Information Processing Systems (NeurIPS) 2023*, volume 37, pages 21510–21523. ArXiv:2303.17491.

Hao Lang, Fei Huang, and Yongbin Li. 2025. Debate helps weak-to-strong generalization. *arXiv preprint arXiv:2501.13124*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793. Association for Computational Linguistics.

Tian Liang and 1 others. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of EMNLP*.

Julian Michael and 1 others. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

Cheng Niu and 1 others. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

OpenAI. 2023. Gpt-4 technical report. Technical report, arXiv preprint arXiv:2303.08774.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Lei Wang and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

9

Xuezhi Wang and 1 others. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Kevin Wu, Eric Wu, and James Y. Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, pages 33402–33422.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303.

Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226, Singapore. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

A Experimental Details

A.1 Base Models and Dataset

Our experiments were conducted using a variety of state-of-the-art language models sourced from different platforms. From the Anthropic API, we used two versions of Claude models: `claude-3-5-haiku-20241022` and `claude-3-7-sonnet-20250219`. From OpenAI's API platform, we employed `gpt-3.5-turbo-0125` and the more recent `gpt-4o-2024-08-06`. For the LLaMA models, we used the `LLaMA 3.1 8B`, `LLaMA 3.3 70B`, and `LLaMA 3.1 405B`.

To support experiments with local LLaMA models, we used three NVIDIA H100 GPUs. Running a full 5-round, 3-agent debate setup over our 600-question dataset required approximately 7 hours of wall-clock time on this setup. This runtime includes both the multi-agent dialogue stages and final judge evaluations.

The benchmark dataset, ClashEval, spans various knowledge domains—including biographical facts, geography, medical dosages, Olympic records, historical data, and breaking news—ensuring wide coverage in both content and difficulty.

A.2 Baseline Implementation Details

We implemented six baseline methods alongside SR-DCR to evaluate the effectiveness of different prompting and reasoning strategies for answering factual questions under a potentially perturbed context. Each method was run under standardized conditions across models and datasets to ensure fair comparison. Unless otherwise stated, all experiments used full-precision models with temperature set to zero and default decoding parameters.

Few-shot Prompting: We provided the model with four randomly selected in-domain exemplars, along with the question and accompanying context and short instruction.

Self-Ask (Press et al., 2022): It is designed to encourage structured intermediate reasoning by prompting the model to first generate a clarifying sub-question, answer that sub-question, and then commit to a final answer. We adopt a fixed template to ensure consistency across examples. The model is instructed to respond in exactly three lines—Sub-question, Sub-answer, and Final answer—with no additional output and make a best-guess answer even when uncertain. To ensure that the model

fully completes all three required lines of output, we set the maximum token limit to 400 during decoding.

RCI (Recursive Criticism and Improvement) (Kim et al., 2023): It encourages iterative self-reflection and refinement by prompting the model to first generate an initial answer, then critique that answer, and finally revise it based on its critique. This process simulates a single-agent self-debate loop. Our implementation follows a three-step pipeline. First, the model is prompted with the question and context and asked to produce its best guess (initial answer). Next, it receives the same input along with its previous response and is asked to identify any factual or logical issues in the form of a short, bulleted critique. Finally, the model is prompted to revise its answer using the critique, outputting only the updated final answer. Each of these three stages can response with a maximum of 300 tokens to ensure sufficient space for generation.

Naive Debate: We use three independent agents of the same language model, all operating with a temperature of 0 and a maximum token limit of 250. The agents engage in a maximum of five debate rounds to iteratively refine their answers. Each agent begins by independently reading the context and question and giving its best answer.

In each subsequent round, agents are shown their peers' answers from previous rounds. They are prompted to (1) evaluate the plausibility of those answers, (2) reflect on their reasoning, and (3) decide whether to revise their answer or stick with their original. The agents are explicitly instructed to avoid vague responses or uncertainty and to always produce a single answer. The debate terminates early if all three agents converge on the same answer. If no consensus is reached after five rounds, we adopt a majority vote to determine the final output.

Classical Debate & Judge Debate: We simulate a debate between three agents instantiated from the same language model, using a temperature of 0 and a maximum token limit of 300. Before the debate starts, we first obtain two initial answers: one is based on context, the other is purely based on prior knowledge. Two agents serve as debaters (Agent A and Agent B), and the third acts as a neutral judge. The debaters are initialized with the same instructions but different initial standings.

Specifically, in our *Judge Debate* setting, Agent A is instructed to support and defend an answer grounded in the context. In contrast, Agent B is instructed to support the answer based on prior knowledge.

Following their initial responses, the two agents engage in a multi-round debate where they alternate presenting arguments, challenging each other's positions, and defending their answers. After 1, 3, or 5 rounds of debate, the judge agent is presented with the question, context, and complete debate transcript and tasked with selecting a winning answer. The judge's final output must include a single answer followed by a concise justification.

A.3 Knowledge Categories

We depict the details of knowledge categories in Tab. 3, where four types are included based on the sampled accuracy.

B More Illustrative Examples

Consider the ClashEval question "Which city hosted the 1904 Summer Olympics?" with ground-truth *St. Louis*. When prompted with four random exemplars and sampled 32 times at $T = 0.5$, the model answers "St. Louis" in 28 trials, yielding a sampled accuracy of 0.875 and a *Highly-Known* label (Wu et al., 2024). In contrast, the question "Which team won the 1974 World Series?" might yield only 6 correct out of 32 (≈ 0.19), classifying it as *Weakly-Known*.

As shown Fig. 5, this illustrates how sampling with random four-shot prompts uncovers differing degrees of model familiarity.

C Additionial Results

C.1 Verification of Self-consistency Scores

For closed-source models that do not expose token-level log-probabilities (e.g., Anthropic's Claude models), we approximate model self-confidence using a sampling-based self-consistency measure. Specifically, we first query the model using few-shot prompt without context to obtain a deterministic answer ($T = 0$). Then, we sample the same question **16 times** using a higher temperature setting ($T = 0.5$). We compute the proportion of sampled outputs that match the deterministic answer, which we interpret as the model's *self-consistency* on that question. This score serves as a proxy for the model's confidence in its original answer: higher consistency indicates higher certainty.

Table 3: Knowledge qualification categories based on the sampled recall rate from 32 completions at $T = 0.5$. This categorization adapts the schema from SLICK (Gekhman et al., 2024).

Type	Category	Definition	Intuition
Known	Highly Known	$P_{correct}(q, a; M, T > 0) \geq 0.85$	Model almost always recalls the fact.
	Maybe Known	$0.30 \leq P_{correct}(q, a; M, T > 0) < 0.85$	Fact is recalled, but inconsistently.
	Weakly Known	$0 < P_{correct}(q, a; M, T > 0) < 0.30$	The model shows only sporadic recall.
Unknown	Unknown	$P_{correct}(q, a; M, T > 0) = 0$	Model never recalls the fact.
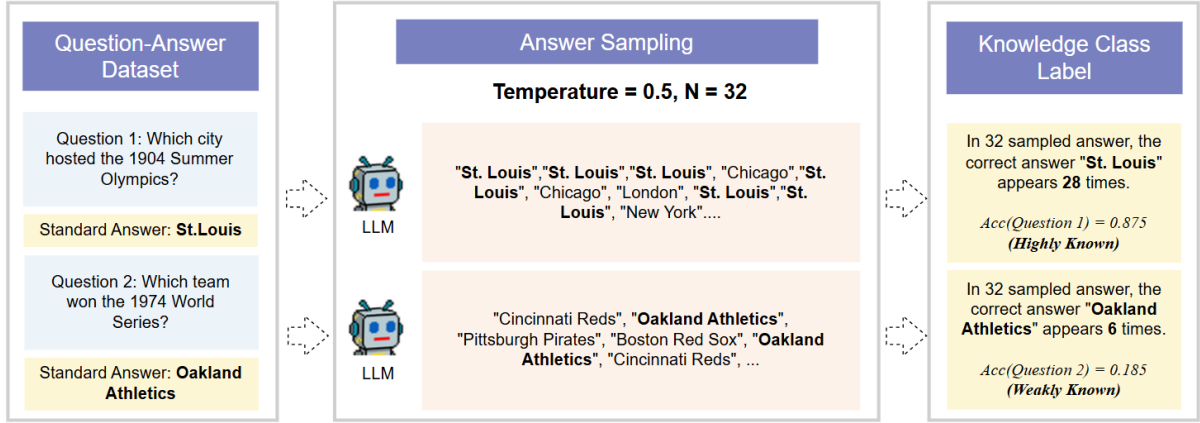
Figure 5: Two illustrative examples of how questions are classified based on the model's familiarity

The distribution of self-consistency scores using a stacked histogram is shown in Fig. 6, which shows a similar pattern to the self-confidence scores derived from models with access to log-probabilities. This alignment supports the effectiveness of self-consistency as an approximation for model confidence, proving that our framework is also effective on closed-source models.

C.2 More Experiment Results

Fig. 6 presents the distribution of self-consistency scores (our proxy for self-confidence) for GPT-3.5 Turbo, GPT-4o, LLaMA 3.1 405B, and LLaMA 3.3 70B, extending the analysis shown in Fig. 6. All four models exhibit similar self-confidence patterns, supporting the generality and robustness of our sampling-based approximation method.

Fig. 8 extends our earlier analysis of the relationship between context preference and knowledge familiarity to six language models, including both open and closed-source LLMs. Each curve represents how often a model aligns its answer with the perturbed (and incorrect) context across varying offset magnitudes, stratified by knowledge category. The consistent trends across models—higher context preference in "unknown" or "unsure" cases, and lower preference in "highly

known" cases—further validate the generality of our framework.

Fig. 9 extends the analysis of context susceptibility by visualizing prior knowledge preference rates across five models, stratified by perturbation strength. Complementing the trends shown in Fig. 4, this figure further demonstrates that models are more likely to accept the perturbed context as the debate rounds increase. Subtle inconsistencies (e.g., offset = 20) often lead to higher context preference after 3 rounds. For GPT 3.5 Turbo and Sonnet 3.7, the Prior knowledge preference rate is even below 20% at the end of the debate when provided with context with minor error.

D Limitations and Future Work

Limitations. SR-DCR relies on fixed confidence thresholds, which may not generalize optimally across all domains or tasks. Additionally, the current implementation assumes access to deterministic judge behavior, which may not hold in real-time deployment with stochastic models. The asymmetric debate structure also presumes the availability of prior-free inference, which can be challenging in closed-source or limited-access APIs.
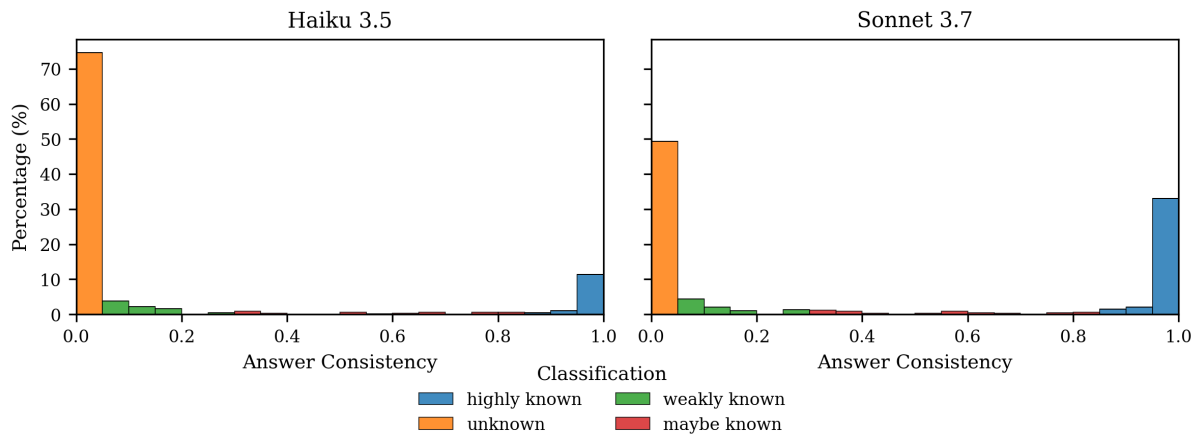
Figure 6: Distribution of self-consistency scores (estimated self-confidence score based on sampling) on our dataset, colored by sampling-based knowledge categories.

Future Work. Future directions include learning adaptive confidence thresholds, training a dedicated judge model with supervised debate data, and extending SR-DCR to multi-hop reasoning and document-level tasks. Incorporating human-in-the-loop feedback or interactive oversight could further enhance its applicability to high-stakes or ambiguous domains.
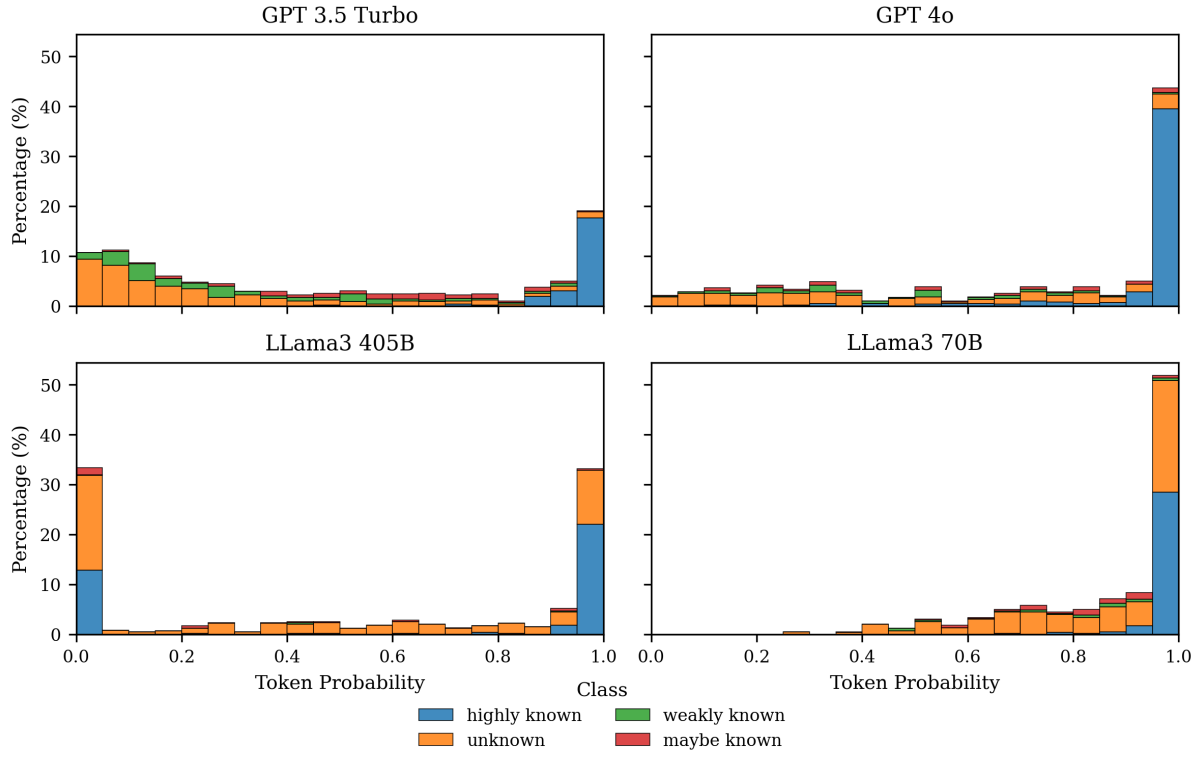
Figure 7: Distribution of self-confidence scores on more models
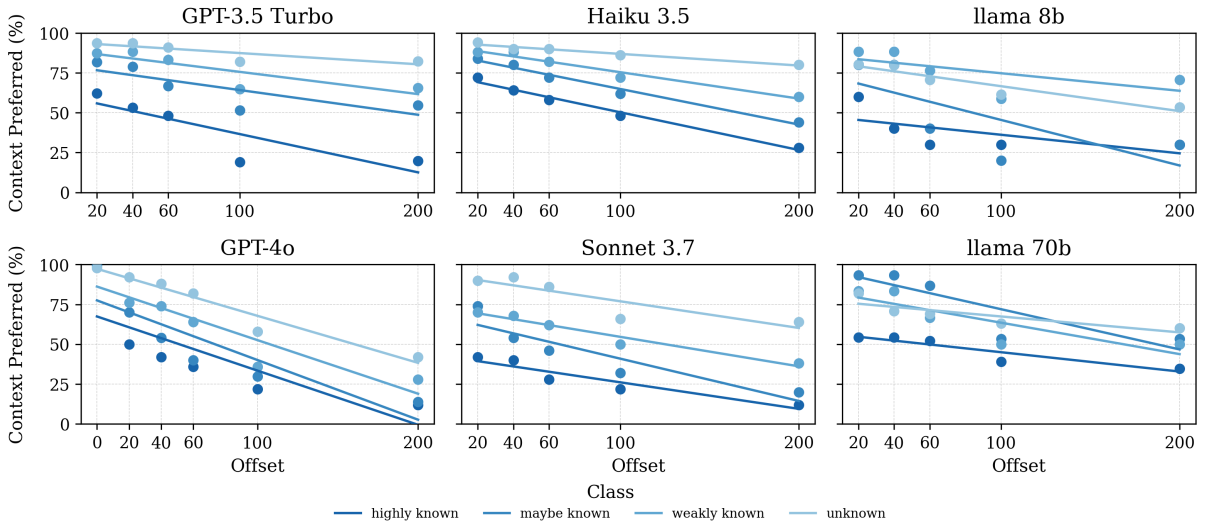

Figure 8: Context preference rates of a 4-shot baseline across offset magnitudes and knowledge categories of six different models on our dataset. Each curve indicates the proportion of examples where the model's answer aligns with the (incorrect) perturbed context.
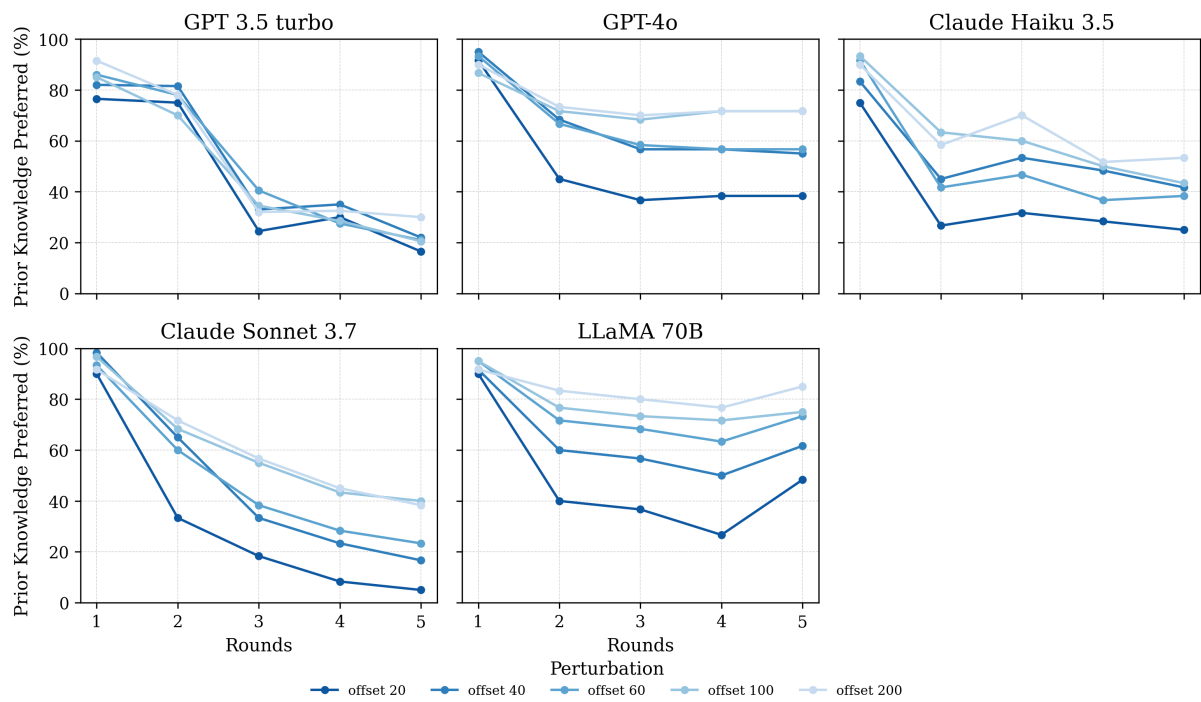
Figure 9: Prior Knowledge preference rate in judge debate of five models on perturbed context across debate rounds, stratified by five levels of perturbation strengths—subtle (offset = 20), mild (40), moderate (60), and blatant (100 & 200).