Can LLMs Really Judge? A Progressive Argumentation-Mining Framework for Distinguishing Understanding from Aggregation

Fuyu Wang[1,2], **Jiangtong Li**[1,2*], **Kun Zhu**[1,2], **Changjun Jiang**[1,2]
1. Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University
2. School of Computer Science and Technology, Tongji University
{fywang, jiangtongli, kzhu00, cjjiang}@tongji.edu.cn

Abstract

Current evaluations of large language models (LLMs) mainly rely on dataset-based generation accuracy. However, generative correctness does not guarantee the discriminative capability required to verify solutions, frequently masking an inability to distinguish valid reasoning from plausible errors. While multi-agent debate inherently entails judgment, we show that uncontrolled context growth and convergence to majority voting introduce significant noise, obscuring intrinsic model judgment. To address these limitations, we propose a **progressive argumentation-mining diagnostic framework** designed to explicitly control context and isolate discriminative behaviors. Instead of indiscriminate aggregation, our approach distills and retains only the single most well-supported rationale per answer, preventing context dilution while enforcing strict quality-based selection. Applying this framework reveals a fundamental cognitive divergence: models exhibit structural susceptibility to plausible misinformation in knowledge tasks, whereas in reasoning tasks they demonstrate latent discriminative potential that remains fragile under pressure. These findings underscore the fragility of discriminative capabilities, advocating for diagnostic methodologies that prioritize judgment stability over simple generation performance.

1 Introduction

Large language models demonstrate strong performance across diverse domains as indicated by benchmark evaluations. However, existing dataset-based evaluation practices are limited, as they largely assume that generation performance alone is sufficient to define model capability (Wu et al., 2024; Liu et al., 2025a; West et al., 2024). These methods primarily assess the capacity to produce responses, often overlooking the discriminative ability of LLMs to verify information. This oversight is critical because reasoning, akin to human cognition, typically involves a cycle of generation followed by reflection and selection (Shleifer, 2012; Li et al., 2021). As a result, the ability to judge a solution's correctness often better indicates understanding than directly generating an answer.

Model Name	Gen.	Maj. Vote	Self-Sel.	Conv.
qwen-2.5-7b-instruct	74.5	79.0	78.7	96.4
qwen-2.5-14b-instruct	85.3	86.4	85.7	98.5
qwen-2.5-32b-instruct	88.3	90.4	90.4	98.2
qwen3-8b	78.2	90.1	73.9	89.9
qwen3-4b	69.7	79.4	69.1	85.3

Table 1: Performance comparison of selection strategies in the Best-of-n setting. **Gen.**: Mean accuracy across all generated candidates; **Maj. Vote**: Accuracy determined by majority voting over the candidate set; **Self-Sel.**: Accuracy of the answer chosen by the model from the candidate pool; **Conv.**: Convergence rate measuring the agreement between majority voting and self-selection.

Model	Initial Acc	3 Agents	5 Agents	7 Agents
qwen3-4b	69.7	71.9	72.2	69.7
qwen3-8b	78.2	78.9	75.7	73.8

Table 2: Impact of agent count on accuracy in a standard multi-agent debate setting.

To emulate this reflective process, multi-agent debate frameworks (Du et al., 2024; Singhal et al., 2025; Chistova et al., 2025) have been developed to enhance model performance, allowing agents to refine their answers based on peer responses. This interaction requires both generative and discriminative skills, as models must propose candidate solutions while simultaneously evaluating those provided by others. However, current approaches primarily utilize these mechanisms to boost final accuracy rather than explicitly isolating and measuring the discriminative capability. First, the debate process often generates extended contexts, introducing noise that complicates assessing the in-

*Corresponding Author.

trinsic judgment ability of LLMs. Second, the reliance on consensus frequently leads to conformity, making it difficult to distinguish genuine identification of correct answers from simple aggregation effects (Choi et al., 2025; Estornell and Liu, 2024; Kaesberg et al., 2025).

To validate these limitations, we conduct two preliminary experiments assessing selection behavior and context length effects. In the first experiment (Table 1), we employ a best-of-n ($n = 7$) setting to compare mean generation accuracy against self-selection (model-chosen answer) and majority voting. Results indicate that self-selection closely mirrors majority voting, suggesting that the selection process acts as simple aggregation rather than evidence of independent discriminative reasoning. In the second experiment (Table 2), we analyze a standard multi-agent debate (MAD) setting (Du et al., 2024) by varying the agent count (3, 5, and 7) over two debate rounds. While accuracy initially improves with fewer agents, it degrades as the count increases, revealing a non-monotonic trend as the context length expands with additional references. This suggests that while limited references aid reasoning, excessive context hinders the information processing capability of LLMs. The convergence toward majority voting and uncontrolled context growth prevent MAD frameworks from effectively isolating and evaluating discriminative ability.

Building on this analysis, we propose a **progressive argumentation-mining diagnostic framework** designed to explicitly control reasoning context and evaluation difficulty. Instead of aggregating all generated outputs, which introduces noise and favors majority voting, our approach groups responses by answer and employs a scoring mechanism to **mine** and retain only the most well-supported rationale for each distinct answer. This design enables the progressive refinement of reasoning quality while maintaining a concise context, allowing us to examine whether models improve their judgments when exposed to stronger arguments rather than simply a higher volume of information. We apply this framework to two distinct evaluation scenarios: reasoning-centric tasks (*e.g.*, mathematical reasoning) and knowledge-centric tasks (*e.g.*, professional question answering).

Experimental results reveal a fundamental cognitive divergence in model capabilities across these domains. In knowledge-centric tasks, models exhibit high susceptibility to misleading information, failing to rectify errors even under ideal evidence conditions, confirming that **reasoning cannot compensate for knowledge deficits**. In contrast, in reasoning-centric tasks, models demonstrate latent discriminative potential, effectively improving decisions through internal verification when exposed to high-quality argumentation. However, this capability remains fragile, degrading significantly as the density of plausible but incorrect reasoning paths increases. Overall, these findings suggest that while generative capabilities have advanced, the discriminative stability required to reject plausible misinformation remains a critical bottleneck. This limitation motivates the need for diagnostic frameworks that prioritize judgment consistency over simple outcome accuracy, isolating genuine understanding from rote memorization. Our contributions are summarized as:

- We distinguish discriminative capability from generative performance, demonstrating that correct generation often masks an inability to verify solutions against competing errors.
- Through experiments, we reveal that standard self-selection mechanisms often collapse into simple aggregation, failing to reflect independent judgment when facing distractors.
- We propose a progressive argumentation-mining framework that integrates debate with quality-based selection. By filtering for high-quality reasoning, this approach controls context noise to isolate discriminative bounds.
- We analyze discriminative behavior across domains, identifying a structural failure in knowledge-centric tasks where models mislead by hallucinations, contrasting with the fragile but present potential in reasoning tasks.

2 Related Work

2.1 LLM Evaluation

LLM evaluation methodologies typically fall into three categories, starting with task-oriented automatic metrics that compare outputs against references using standards like accuracy, BLEU (Papineni et al., 2002), or ROUGE (Lin, 2004). Human-centered approaches employ annotators to assess qualities like coherence (Do et al., 2025), though they remain costly and prone to inter-annotator variability (Piot et al., 2025). Behavioral evaluations probe model stability under controlled perturbations, including paraphrasing (Chataigner

et al., 2025), adversarial inputs (Raina et al., 2024), and distribution shifts (Blanchet et al., 2024). Within the first category, benchmark-based evaluation remains the dominant paradigm, utilizing general-purpose suites like MMLU (Hendrycks et al., 2021), BIG-Bench (Suzgun et al., 2023), and C-Eval (Huang et al., 2023) for standardized comparisons. As performance on early benchmarks saturates, challenging suites like BIG-Bench Hard (Kazemi et al., 2025) have been introduced to test advanced reasoning capabilities. In parallel, domain-specific benchmarks have emerged for specialized fields, including medicine (Cai et al., 2024), law (Fei et al., 2024), mathematics (Lightman et al., 2024), and science (Wang et al., 2024). However, despite this diversity, most benchmarks rely on final-answer correctness, assuming that successful generation equates to understanding.

Exiting frameworks conflate generative and discriminative capabilities (West et al., 2024; Wu et al., 2024), measuring correctness without verifying the ability to distinguish valid reasoning from errors. High benchmark accuracy often reflects shallow heuristics or memorization rather than stable reasoning criteria (West et al., 2024). This fragility appears in performance drops under semantic reformulations, highlighting sensitivity to surface forms despite unchanged meaning (Salido et al., 2025). As a result, accuracy metrics may overstate true understanding and obscure important differences in discriminative capability between models with similar scores (Jiang et al., 2025). Recent studies therefore advocate jointly assessing generation and discrimination to evaluate reasoning stability (Lin et al., 2024; Tyen et al., 2024).

2.2 Multi-Agent Debate

Multi-Agent Debate (MAD) has established itself as a paradigm leveraging inter-model interactions to enhance both task performance and analysis. Early research primarily utilized debate to improve accuracy on complex reasoning tasks, such as mathematics and multi-step inference (Du et al., 2024). Subsequent work further explored the role of diverse strategies in multi-agent debate, by encouraging agents to adopt distinct problem-solving approaches rather than merely different personas, so as to break homogeneous reasoning patterns and improve debate effectiveness with fewer rounds (Liu et al., 2025b). Studies demonstrate that iterative argumentation reveals diverse solution paths and mitigates single-model reasoning errors (Liang et al., 2024). Recent work extends MAD to probe capabilities beyond accuracy, including stability, self-correction, and consistency across competing explanations (Xiong et al., 2023), while also applying debate frameworks to better approximate real-world debate settings (Wang et al., 2025). As an evaluative setting, debate combines both generation and discrimination. Agents must produce answers while judging competing responses, even when those responses appear plausible. This discrimination is embedded in the interaction rather than separated into an independent judging stage. This coupling makes MAD an ideal framework for evaluating whether models can distinguish valid reasoning under contextual pressure.

Recent research utilizes debate frameworks for evaluation, moving beyond simple answer refinement. Methods range from adversarial argument selection (Chan et al., 2024) to factuality assessments challenging unsupported claims (Khan et al., 2024). However, current approaches often lack interpretability and rigorous process control. Simply accumulating arguments introduces noise and context sensitivity, making it difficult to distinguish genuine reasoning from input complexity(Kaesberg et al., 2025). Therefore, models are overwhelmed by argument quantity rather than quality, impeding the assessment of judgment stability.

3 Diagnostic Framework

Existing evaluation paradigms primarily focus on generative accuracy, often neglecting the model's capacity to distinguish valid reasoning from plausible alternatives. We introduce a progressive argumentation-mining framework that integrates discriminative tasks within a controlled multi-agent debate. This design isolates judgment capability by measuring how models adjust their decisions in response to high-quality arguments rather than expanded context volume.

3.1 Candidate Reference Construction

A primary challenge in debate-based evaluation lies in balancing reference quality with context constraints, as indiscriminate retention of responses causes rapid context expansion, confounding judgment with noise and excessive length. To address this, we implement a quality-controlled argumentation-mining mechanism that filters candidate integration across debate rounds. Specifically, we maintain a **candidate list** where each unique

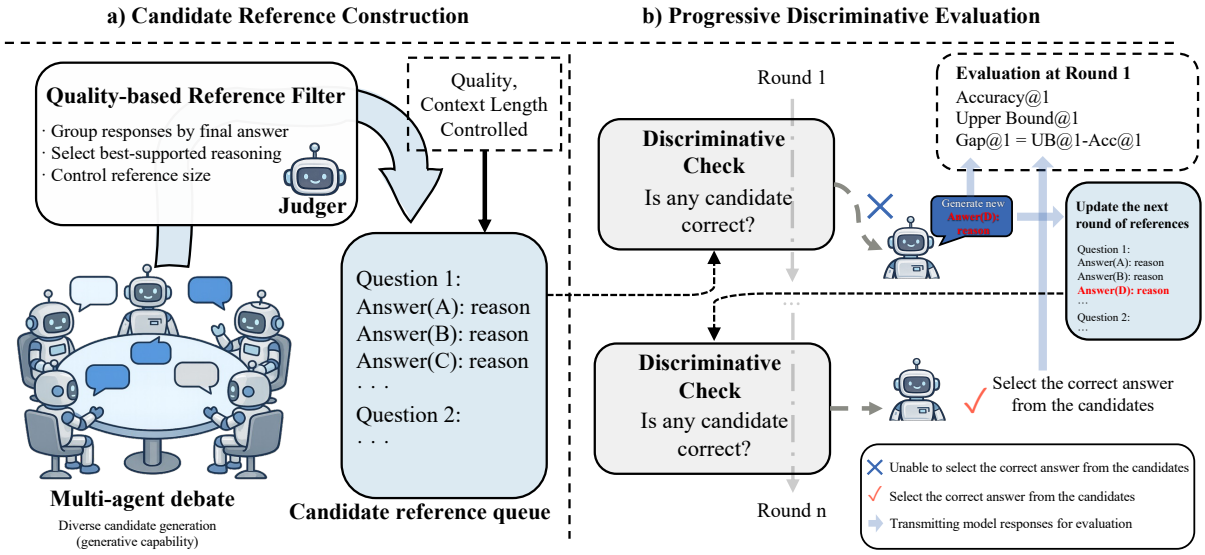


Figure 1: **Overview of the Progressive Argumentation-Mining Framework. (a) Candidate Reference Construction:** Responses are clustered by answer, and a judge mines the optimal argument per cluster to form a concise, quality-controlled reference queue. **(b) Progressive Discriminative Evaluation:** Models evaluate these references, prioritizing selection over generation. This cycle progressively tests whether exposure to mined, high-quality arguments improves judgment stability compared to simple information accumulation.

final answer is paired with a single rationale. In each round, generated responses along with candidate list are grouped by their final answer, and a judge model selects the single most well-supported rationale to represent each unique answer, updating the list to retain only the optimal rationale for each answer (See Appendix A.1 for judge prompt). This strategy achieves a dual objective: first, it curbs uncontrolled context expansion by filtering redundant or weak responses, ensuring concise and comparable inputs across rounds; second, by prioritizing quality over quantity, it attributes performance shifts to stronger reasoning exposure rather than context length or noise. This pipeline is rigorous for isolating discriminative behavior under progressively refined evidence.

3.2 Progressive Discriminative Diagnosis

Building on the updating candidate list, we integrate discriminative diagnosis directly within the multi-agent debate workflow. In each round, the model first evaluates whether any existing candidate aligns with its judgment; if a match is found, it is selected, otherwise, the model generates a new response. This structure explicitly prioritizes discrimination, compelling the model to assess competing alternatives before resorting to generation. By iterating between selection and conditional generation, the framework tightly couples discriminative judgment with generative action in a unified process. Furthermore, this design facilitates fine-grained analysis of judgment dynamics across rounds. It allows us to observe whether models leverage high-quality references, persist in redundant generation, or succumb to plausible misinformation. Therefore, the framework offers a principled, interpretable method to evaluate discriminative capability and judgment stability under controlled conditions (see Appendix A.2 for prompts).

3.3 Metrics for Discriminative Diagnosis

We propose a hierarchical metric framework aligned with our progressive debate structure, spanning individual rounds, cross-round dynamics, and cross-dataset distributions. This design assesses not only outcome correctness but also the evolution of model judgment as argument quality improves.

At the round level, we quantify how effectively a model utilizes available reference evidence via the upper-bound gap. Let m denote a model, d a dataset, and $r \in \{0, \ldots, R\}$ the debate round. We define model accuracy at round r as $\text{Acc}_{m,d,r}$, and the corresponding upper bound determined by the available reference candidates as $\text{UB}_{m,d,r}$. The per-round gap is defined as:

$$\text{Gap}_{m,d,r} = \text{UB}_{m,d,r} - \text{Acc}_{m,d,r}. \quad (1)$$

This metric measures the disparity between actual

performance and the maximum potential achievable with the candidate set. Narrower gaps indicate superior utilization of high-quality arguments.

At the cross-round level, we assess the correlation between evidence quality and model performance using *Trend Alignment* (TA). For model m and dataset d, we analyze the accuracy sequence $\{\text{Acc}_{m,d,r}\}_{r=0}^{R}$ against the upper-bound sequence $\{\text{UB}_{m,d,r}\}_{r=0}^{R}$. *Trend Alignment* is defined as:

$$\text{TA}_{m,d} = \text{corr}\big(\{\text{Acc}_{m,d,r}\}_{r=0}^{R}, \{\text{UB}_{m,d,r}\}_{r=0}^{R}\big), \quad (2)$$

where $\text{corr}(\cdot,\cdot)$ denotes the Pearson correlation coefficient. This metric determines whether improvements in argument quality consistently translate into accuracy gains. High TA indicates judgment responsive to evidence, whereas low or negative TA suggests that stronger references fail to reliably improve decision-making.

At the dataset level, we analyze systematic divergences between reasoning-centric and knowledge-centric tasks. We classify datasets into reasoning-centric (D_R) and knowledge-centric (D_K) categories, defining Category Separation as:

$$\text{Sep}_m(S) = \overline{S}_{m,D_R} - \overline{S}_{m,D_K}, \quad (3)$$

where $S_{m,D_R} = \{\text{TA}_{m,d}\}_{d \in D_R}$ and the overline denotes averaging over datasets within each category. This metric captures process-level differences in how consistently model judgments improve as argument quality increases across task categories. Together, these metrics provide a unified assessment of judgment quality, temporal dynamics, and task-dependent behavior within our framework.

4 Experiments and Results

4.1 Experiments Setting

Compared Method. We evaluate our framework using seven LLMs covering diverse parameter scales and training iterations. Specifically, we select six open-source models (Qwen-2.5-7B-Instruct (Yang et al., 2024), Qwen-2.5-14B-Instruct, Qwen-2.5-32B-Instruct, Qwen3-4B (Yang et al., 2025), Qwen3-8B, and glm-4-9b-chat (GLM et al., 2024)) alongside one proprietary model, doubao-1.5-pro-32k. These models vary in scale and architecture yet exhibit comparable instruction-following capabilities, allowing for controlled comparison under a unified protocol. We assess all models in inference-only mode, without fine-tuning. During the debate, gemini-2.5-pro and gpt-5.1 serve as external judges, tasked with curating high-quality reasoning references for subsequent rounds.

Dataset. To examine model behavior across distinct capability domains, we employ four benchmark datasets categorized by their demands.

Knowledge-centric datasets prioritize domain-specific factual mastery. We utilize two MMLU subsets (Hendrycks et al., 2021), Professional Medicine and Virology, which demand precise interpretation of specialized concepts and expose susceptibility to misinformation.

Reasoning-centric datasets target logical consistency and multi-step inference. Evaluations include MMLU-Formal Logic for abstract reasoning and Math500 (Lightman et al., 2024) for complex mathematical problem solving.

We maintain identical inference settings across all datasets. Unless noted otherwise, we set the temperature to 1.0 and conduct three independent inference trials per model.

4.2 Discriminative Capability Analysis

4.2.1 Round-Level Dynamics

At the round level, we quantify evidence utilization using the upper-bound gap. This metric measures the difference between accuracy and the maximum potential performance within candidates.

Table 3 and Table 4 present the gap dynamics across dataset categories. (Computation details provided in Appendix B.1). In knowledge-centric datasets (MMLU-Professional Medicine and MMLU-Virology), the upper-bound gap consistently widens across debate rounds for most models. Although the upper bound improves with additional references, model accuracy lags, causing the gap to expand. This trend is evident in mid- and small-scale models (*e.g.*, Qwen3-4B, GLM-4-9B), suggesting that reference noise in later rounds impedes informational gains in expert domains. Conversely, reasoning-centric datasets (Math500 and MMLU-Formal Logic) display varied gap dynamics. Stronger models (*e.g.*, Qwen-2.5-32B) maintain stable gaps, indicating effective utilization of refined reasoning traces. However, weaker models show expanding gaps, particularly on Formal Logic, where accuracy plateaus despite rapid upper-bound growth. This distinction underscores a capacity dependency: additional rounds benefit performance only when the model reliably distinguishes high-quality reasoning from distractors. Reasoning-centric datasets typically exhibit larger

Model Setting	MMLU-profession medicine					MMLU-virology				
	round1	round2	round3	round4	round5	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	19.8	20.9	22.8	24.2	25.3	10.4	11.5	15.1	15.6	15.1
qwen-2.5-14b-instruct	10.3	11.8	11.8	12.2	12.1	6.8	6.7	6.3	6.9	7.5
qwen-2.5-32b-instruct	6.3	7.7	7.2	6.4	6.7	7.5	8.1	9.0	10.2	8.8
qwen3-4b	28.6	33.3	33.7	41.4	43.7	21.6	27.5	32.4	34.0	32.9
qwen3-8b	24.0	27.0	27.9	30.6	33.6	18.2	21.5	21.8	25.6	28.4
doubao-1.5pro-32k	2.8	2.8	2.9	2.9	3.1	1.3	1.0	1.3	1.1	1.6
glm-4-9b-chat	24.3	26.1	29.4	30.4	33.5	17.5	19.3	21.8	20.6	23.1

Table 3: Upper-bound gaps across debate rounds on **knowledge-centric datasets**.

Model Setting	Math500					MMLU-formal logic				
	round1	round2	round3	round4	round5	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	14.3	15.7	17.0	15.9	16.6	20.4	23.0	22.2	24.6	26.2
qwen-2.5-14b-instruct	10.3	10.6	10.2	10.4	11.7	17.1	17.6	15.7	14.1	15.4
qwen-2.5-32b-instruct	8.2	8.3	8.8	9.1	9.6	15.9	13.5	11.1	11.9	14.0
qwen3-4b	5.3	7.0	7.6	7.9	8.0	21.8	31.2	30.8	32.3	32.9
qwen3-8b	4.9	8.2	9.5	9.8	10.1	16.6	26.0	25.2	28.4	27.0
doubao-1.5pro-32k	4.4	6.0	6.6	6.8	7.4	1.6	1.7	1.6	1.6	1.6
glm-4-9b-chat	23.4	26.3	31.3	31.3	34.2	28.1	33.3	33.8	34.0	37.2

Table 4: Upper-bound gaps across debate rounds on **reasoning-centric datasets**.

absolute gaps than knowledge-centric ones, even among stronger models. This does not imply inferior reasoning; rather, it results from the rapid expansion of the attainable upper bound in reasoning tasks. As rounds progress, the candidate set accumulates diverse reasoning paths with varying lengths, structures, and assumptions. While this diversity raises the upper bound, it challenges the model's discriminative capacity to identify the most effective reasoning. When this capacity is insufficient, the upper bound grows faster than realized accuracy, resulting in a larger observed gap.

In summary, round-level gap analysis demonstrates that increasing debate rounds does not universally improve performance. While later rounds expand potential performance, only models with sufficient discriminative ability translate this potential into actual gains. Otherwise, the widening gap reveals a failure mode where iterative aggregation increases inefficiency, an effect masked when analyzing accuracy or upper bounds in isolation.

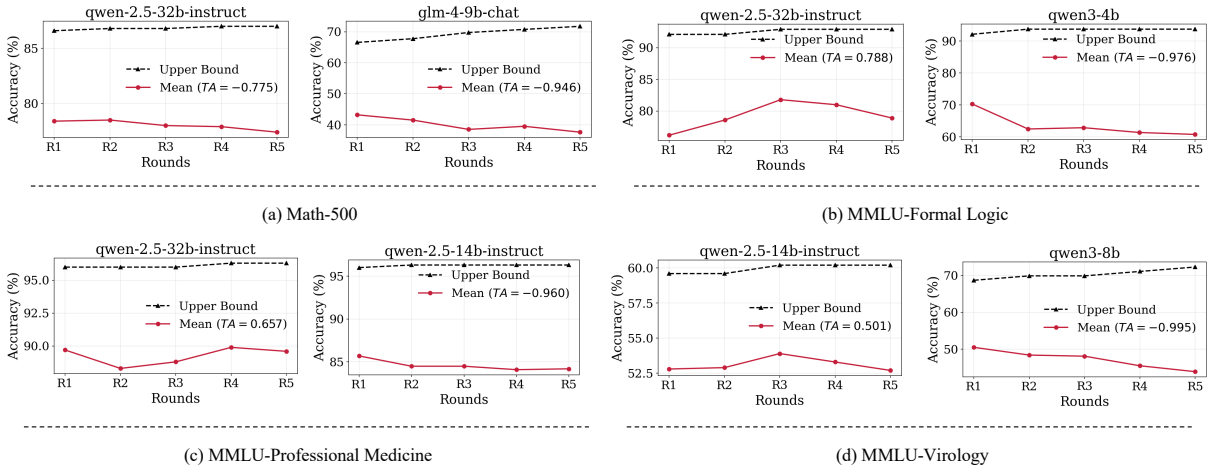
4.2.2 Cross-Round Level Dynamics

At the cross-round level, we analyze performance evolution through the trend alignment between the upper bound (*e.g.*, the maximum achievable accuracy within the candidate set) and the mean accuracy across rounds. This reveals whether models convert the expanding candidate space into realized gains or instead undergo systematic degradation. As shown in Figure 2, this alignment varies substantially across task categories.

In reasoning-centric domains (Math500 and MMLU-Formal Logic), the pattern is not determined by model capability alone. Qwen-2.5-32B, despite its relatively strong initial capability, shows positive alignment on Formal Logic (TA=0.788) but negative alignment on Math500 (TA=-0.775), indicating that higher capability does not guarantee consistent gains from the expanding solution space. Models with relatively lower initial capability can exhibit even stronger negative alignment, such as glm-4-9b-chat on Math500 (TA=-0.946) and Qwen3-4B on Formal Logic (TA=-0.976).

In knowledge-intensive domains, both negative alignment and weak positive alignment are observed. On MMLU-Professional Medicine, Qwen-2.5-14B shows severe negative alignment (TA=-0.960), whereas Qwen-2.5-32B exhibits only modest positive alignment (TA=0.657). A similar pattern appears in MMLU-Virology: Qwen3-8B shows strong negative alignment (TA=-0.995), while Qwen-2.5-14B displays weak positive alignment (TA=0.501).

Overall, these results suggest that cross-round dynamics are shaped by task characteristics and model-specific robustness rather than model capability alone. While additional debate rounds consistently raise the upper bound, this potential translates into actual gains only when models can stably retain and exploit useful evidence across rounds. Detailed results are in Appendix B.2.



(a) Math-500 (b) MMLU-Formal Logic

(c) MMLU-Professional Medicine (d) MMLU-Virology

Figure 2: **Analysis of Trend Alignment (TA) between Upper Bound and Mean Accuracy.** Positive TA indicates that the model successfully tracks the expanding solution space. Negative TA reveals systematic degradation where accuracy declines despite improved potential evidence.

4.2.3 Dataset-Level Distribution

We use Sep(S) to measure dataset-level differences between reasoning-centric and knowledge-centric domains, where positive values indicate stronger alignment in reasoning-centric tasks and negative values indicate a more generalized failure to track evidence quality across domains. Table 5 shows that most models yield positive separation, indicating that their trend alignment is generally stronger in reasoning-centric tasks than in knowledge-centric ones. Qwen-2.5-32B achieves the largest separation (0.638), suggesting the clearest relative advantage in tracking evidence quality under reasoning-focused settings. Qwen-2.5-14B (0.377), Qwen-2.5-7B (0.211), and GLM-4-9B-Chat (0.237) also exhibit positive separation, although the magnitude varies across models. In contrast, Qwen3-4B (-0.067) and Qwen3-8B (-0.056) show negative separation, indicating that their alignment difficulties are not confined to knowledge-centric tasks but also extend to reasoning-centric ones.

Model	Sep(S)
qwen-2.5-7b-instruct	0.211
qwen-2.5-14b-instruct	0.377
qwen-2.5-32b-instruct	0.638
qwen3-4b	-0.067
qwen3-8b	-0.056
glm-4-9b-chat	0.237

Table 5: **Category Separation based on Sep(S).** Positive values indicate superior evidence utilization in reasoning tasks, while negative values suggest a generalized failure to track evidence quality across domains.

4.3 Ablation Experiment

4.3.1 Effect of Different Judge Models

We assess the impact of judge models using the upper-bound gap . On MMLU-Virology (Table 25), gaps consistently widen for most models regardless of the judge. For instance, Qwen3-4b exhibits expanding gaps under both OpenAI (22.9 $\rightarrow$ 28.0) and Gemini (21.6 $\rightarrow$ 32.4) judges, indicating a failure to convert raised upper bounds into accuracy. Conversely, Doubao-1.5-pro-32k maintains minimal gaps ($\leq$ 1.3) across all settings, reflecting efficient evidence utilization. On MMLU-Formal Logic (Table 26), gap dynamics strictly follow model capacity. Strong models like Qwen-2.5-32B maintain stable gaps (11.1–15.9), while weaker models like Qwen3-4b see significant expansion (21.8 $\rightarrow$ 32.2). Doubao-1.5-pro-32k again demonstrates robust utilization with gaps remaining narrow (1.6–1.9). These consistent patterns confirm that while judge models elevate the attainable upper bound, the limiting factor remains the model's discriminative capacity to exploit these references.

4.3.2 Effect of Candidate Reference Quality

We examine the impact of reference quality on upper-bound gaps using Tables 27 and 28 . Weakly supported references systematically widen gaps, particularly for weaker models. On Formal Logic, Qwen3-4b's gap expands significantly under weak support (22.6 $\rightarrow$ 32.2), whereas well-supported references allow stronger models like Qwen-2.5-32B to narrow the gap (15.9 $\rightarrow$ 11.1). Overall,

Case study: round-wise candidate refinement for MMLU-formal logic `question_19` (GT: (B))

Question: Select the best translation into predicate logic: No artifacts are people.
(A) $\neg P(a)$
(B) $(\forall x)\,(A(x) \rightarrow \neg P(x))$
(C) $\neg A(p)$
(D) $(\forall x)\,(A(x) \wedge \neg P(x))$

Candidate reference update log:
"Action": "added","Round":1,"Answer":"(D)"
"Action": "discarded","Round":1
"ExistingScore": 8.5, "CandidateScore": 8, "Reason": "Both reasonings are clear and rigorous. However, Option 1 is more detailed in explaining why Option (D) is incorrect by explicitly stating that it makes a stronger claim than the original statement. The additional clarity gives Option 1 a slight edge".
"Action": "replaced","Round":1
"BetterOption": 2, "ExistingScore": 8, "CandidateScore": 9, "Reason": "Both reasonings are clear and rigorous. However, Option 2 has a more concise and well - structured presentation. It directly dives into analyzing each option without a separate 'Understanding the statement' section, which makes the overall reasoning flow more smoothly and is easier to follow."
...

Table 6: Case study from MMLU-Formal logic.

Case study: round-wise candidate refinement for MMLU-Virology `question_130` (GT: (A))

Question: The successful anti-cancer HPV vaccine consists of:
(A) Live virus attenuated by specific mutagenesis
(B) Whole virus chemically inactivated vaccine
(C) Self-assemble of virus L1 protein into VLP
(D) Sub unit chemically inactivated vaccine

Candidate reference update log:
"Action": "added", "Round": 1, "Answer": "(C)"
"Action": "added", "Round": 1, "Answer": "(D)"
"Action": "replaced", "Round": 1
"ExistingScore": 7.5, "CandidateScore": 8.5, "Reason": "Option 2 provides more in-depth details about the immunogenicity of VLPs and offers a more rigorous explanation."
"Action": "discarded", "Round": 2
"ExistingScore": 8, "CandidateScore": 3, "Reason": "Option 1 directly analyzes each option to reach the answer, while Option 2 mainly evaluates other references rather than solving the problem."
...

Table 7: Case study from MMLU-Virology.

Qwen3-4B Model	Round1	Round2	Round3	Round4	Round5
Overlap (H–M)	0.90	0.92	0.93	0.93	0.93
Acc. (LLM-Curated)	63.7	65.9	62.2	62.2	62.2
Acc. (Human-Curated)	62.2	65.2	63.0	63.7	61.5

Table 8: Comparison between Human and LLM-Curated Candidate Queues.

strong models like Doubao-1.5-pro-32k demonstrate robust discriminative utilization, maintaining negligible gaps ($\approx$ 1–2) regardless of reference quality. This confirms that while reference quality establishes the performance ceiling, the model's intrinsic discriminative capability ultimately determines whether that potential is realized.

4.3.3 Effect of Different Number of Agents

We analyze the impact of agent count via the upper-bound gap, as shown in Tables 29 and 30. Increasing the number of agents systematically widens the gap between the model's performance and its upper bound. On Formal Logic, Qwen3-4b's gap expands monotonically from 7.7–16.1 (3 agents) to 26.9–34.1 (7 agents). This trend intensifies on Virology, where the gap nearly doubles ($16.6 \rightarrow 40.1$). This indicates that while adding agents raises the performance ceiling, most models fail to utilize this potential, instead succumbing to the increased noise. Uniquely, Doubao-1.5-pro-32k maintains narrow gaps (1–5) across all scales, demonstrating robust evidence utilization. Conversely, the rapid gap expansion in weaker models highlights that the bottleneck is discriminative capacity, not evidence availability.

4.4 Candidate Curation with a Strong Judge

To construct a stable candidate queue, we use a stronger model as a judge to select higher-quality reasoning trajectories. The goal is not to supply oracle correctness, but to control context so that behavior below the tested model's discriminative capacity is not confounded by uncontrolled context growth or a tendency toward majority voting. By retaining only the single most well-supported rationale for each answer, the framework stabilizes the candidate set and enables a cleaner diagnosis of intrinsic discriminative behavior.

We further validate this design with human annotation on the MMLU-Professional Medicine subset. Specifically, 50% of the candidate-update instances are sampled, and the full five-round candidate-update process is replicated using human annotators as judges. As shown in Table 8, the human–model overlap (H–M) remains consistently high (0.90–0.93 across rounds), indicating strong agreement in candidate selection. We then evaluate the same model under both human-curated and LLM-curated candidate queues and observe closely aligned accuracy trends across all rounds.

This consistency suggests that the stronger judge approximates human-level candidate filtering well, without introducing artificial performance gains.

4.5 Case Study

Table 6 illustrates a trajectory where the candidate queue is constructed incrementally within the first round through sequential updates from three agent responses. An initial candidate is added first, and later candidates are then either discarded or used to replace the retained reference according to the judge's comparison. As the debate progresses, the candidate queue is incrementally updated by comparing newly generated candidates against the retained references from previous rounds. The candidate queue continues to receive newly generated candidates in later rounds. The table shown here presents only an abbreviated snapshot of this ongoing process and therefore omits additional candidate comparisons and queue updates that occur in subsequent rounds.

Table 7 illustrates a contrasting failure trajectory in which the candidate queue is updated incrementally, but the retained references remain incorrect throughout the process. In the first round, multiple incorrect candidates enter the queue, and a later candidate replaces the retained reference because its explanation is judged more detailed and rigorous. However, this replacement improves only the presentation of the reasoning in surface form, without correcting the underlying conclusion. In later rounds, additional candidates continue to be evaluated against the retained references, yet the shortened record shows that no retained reference corresponds to the correct answer at any stage, indicating that the update process never succeeds in producing a correct trajectory.

5 Conclusion

We introduced a Progressive Argumentation-Mining Framework designed to serve as a rigorous diagnostic tool for LLM capabilities. By strictly controlling reference quality and filtering contextual noise, our approach effectively isolates discriminative capability from generative randomness. Our experiments expose a fundamental **cognitive fracture** in current LLMs: (1) In **Reasoning-centric domains**, strong models demonstrate a latent ability to improve judgments when exposed to high-quality argumentation, validating the utility of debate for logic refinement. (2) In contrast, **Knowledge-centric domains** reveal a systemic failure mode. Regardless of evidence quality, models exhibit high susceptibility to plausible misinformation, indicating that *reasoning cannot compensate for knowledge deficits*. These findings serve as a cautionary tale: current multi-agent paradigms relying on "collective intelligence" may amplify hallucinations in knowledge-heavy tasks. Future evaluations should focus on stability metrics to distinguish understanding from memorization.

Limitations

While our framework enables a more controlled and process-aware evaluation of LLM judgment, it still has several limitations. The reliance on multi-agent debate and external judges introduces additional computational cost, which may restrict scalability to larger models or broader benchmarks. Moreover, although our analysis distinguishes knowledge-centric and reasoning-centric behaviors, the framework does not directly explain the underlying causes of these differences at the representation or training level. Finally, our experiments are limited to a small set of datasets and configurations. Extending the framework to more diverse tasks and adaptive debate settings remains an important direction for future work.

Ethical Considerations

Our work focuses exclusively on technical and methodological aspects of large language model evaluation. The proposed framework targets the assessment of models' judgment and discrimination capabilities in controlled debate settings and does not involve sensitive political, social, or cultural content. By emphasizing controlled context and progressively refined arguments, the framework encourages more cautious and evidence-based decision-making, which can help reduce the influence of misleading information. Moreover, it is intended as a diagnostic methodology for model behavior rather than a mechanism for producing or amplifying harmful content. Overall, our evaluation design supports safer and more reliable model assessment without introducing ethical risks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62402341, 62302337) and the Postdoctoral Fellowship Program of CPSF (GZC20241225).

References

Jose Blanchet, Peng Cui, Jiajin Li, and Jiashuo Liu. 2024. Stability evaluation via distributional perturbation analysis. In *ICLM 2024*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *AAAI 2024*, volume 38, pages 17709–17717.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *ICLR 2024*.

Cléa Chataigner, Rebecca Ma, Prakhar Ganesh, Yuhao Chen, Afaf Taïk, Elliot Creager, and Golnoosh Farnadi. 2025. Say it another way: Auditing llms with a user-grounded automated paraphrasing framework. In *NeurIPS 2025*.

Elena Chistova, Philipp Cimiano, Shima Haddadan, Gabriella Lapesa, and Ramon Ruiz-Dolz, editors. 2025. *Proceedings of the 12th Workshop on Argument Mining*. Association for Computational Linguistics.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? In *NeurIPS 2025*.

Heejin Do, Jaehui Hwang, Dongyoon Han, Seong Joon Oh, and Sangdoo Yun. 2025. What defines good reasoning in llms? dissecting reasoning steps with multi-aspect evaluation. *arXiv preprint arXiv:2510.20603*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *ICML 2024*.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *NeurIPS 2024*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *EMNLP 2024*, pages 7933–7962.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *ICLR 2021*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. 2025. Self-[in] correct: Llms struggle with discriminating self-generated responses. In *AAAI 2025*, volume 39, pages 24266–24275.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *ACL 2025*, pages 11640–11671.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Yuanzhu Peter Chen, et al. 2025. Big-bench extra hard. In *ACL 2025*, pages 26473–26501.

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 *ICML 2024*.

Jiangtong Li, Liu Liu, Li Niu, and Liqing Zhang. 2021. Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. *IEEE Transactions on Image Processing*, pages 9193–9207.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP 2024*, pages 17889–17904.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *ICLR 2024*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. In *ACL 2024*, pages 1552–1587.

Xinxin Liu, Aaron Thomas, Cheng Zhang, Jianyi Cheng, Yiren Zhao, and Xitong Gao. 2025a. Refining salience-aware sparse fine-tuning strategies for language models. In *ACL 2025*, pages 31932–31945.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025b. Breaking mental set to improve reasoning through diverse multi-agent debate. In *ICLR 2025*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Paloma Piot, David Otero, Patricia Martín-Rodilla, and Javier Parapar. 2025. Can llms evaluate what they cannot annotate? revisiting llm reliability in hate speech detection. *arXiv preprint arXiv:2512.09662*.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *EMNLP 2024*, pages 7499–7517.

Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. 2025. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks. *arXiv preprint arXiv:2502.12896*.

Andrei Shleifer. 2012. Psychologists at the gate: a review of daniel kahneman's thinking, fast and slow. *Journal of Economic Literature*, 50(4):1080–1091.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL 2023*, pages 13003–13051.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *ACL 2024*, pages 13894–13908.

Fuyu Wang, Jiangtong Li, Kun Zhu, and Changjun Jiang. 2025. InspireDebate: Multi-dimensional subjective-objective evaluation-guided reasoning and optimization for debating. In *ACL 2025*, pages 27525–27544.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *ICML 2024*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative ai paradox: "what it can create, it may not understand". In *ICLR 2024*.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2024. Large language models can self-correct with key condition verification. In *EMNLP 2024*, pages 12846–12867.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *EMNLP 2023*, pages 7572–7590.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A System Prompt

A.1 Candidate Reference Evaluation Prompt for Judge Models

To systematically filter candidate reasoning references, we employ a judge prompt that prioritizes reasoning quality over answer accuracy. For each instance, the judge compares two reasoning traces arriving at an identical final answer: an established reference and a new candidate. The evaluation determines the superior reasoning based on clarity, logical flow, and the identification of flawed or unsupported steps. Since both candidates share the same result, the assessment targets only the internal coherence and validity of the derivation. The model allocates a 0-10 score to each path and chooses the preferred option, outputting a JSON object with the selection, scores, and a short rationale. We use this judge model strictly to maintain reference quality during evaluation; it remains separate from the set of models being tested. The full judge prompt used in our experiments is shown in Table 9.

Prompt for Judger

System prompt:
You are acting as a judge to evaluate which reasoning is better. Both reasonings lead to the same answer, but you need to determine which one is more clear, rigorous, and has fewer logical gaps. Give each option an overall score (0-10).
Respond strictly in JSON format:
{"BetterOption": 1 or 2, "ExistingScore": <float>, "CandidateScore": <float>, "Reason": brief explanation}

Option 1 is the existing reasoning, Option 2 is the candidate reasoning. Prefer higher score and better reasoning

User Prompt:
Problem:{question}
Answer: {answer}
Option 1 (Existing - {existing source}):{existing reasoning}
Option 2 (Candidate - {candidate source}):{candidate reasoning}
Which reasoning is better? Respond with JSON:
{"BetterOption": 1 or 2, "Reasoning":...}

Table 9: Prompt for judge to filter candidates

A.2 Prompts for Multi-Agent Debate Process

We execute the multi-agent debate (MAD) via a unified prompting strategy where agents, given reference candidates, evaluate existing solutions to adopt a suitable match or generate a fresh response otherwise. In the absence of references, the agents directly derive the solution. Table 10 outlines the specific prompt employed for this procedure.

Prompt for the Multi-Agent Debate (MAD) Process

Reference-aware debate.
When reference candidates are available, agents examine each reference solution step by step, evaluate its reasoning process and final answer, and select an answer if a suitable reference is identified. If none of the references are considered correct, agents generate a new answer based on their own analysis.
System prompt:
You are a assistant. Your task is to carefully examine each reference solution's analysis process and answer. Evaluate each reference step by step to determine which one provides the correct answer. If none of the references contain the correct answer, you should generate your own correct answer based on your medical knowledge. Your final answer should be in the form (X) at the end of your response, where X is one of A, B, C, or D.
User Prompt:
{reference context}
Problem: {question}
Please examine each reference solution's analysis process and answer one by one. Evaluate whether each reference's reasoning is sound and whether its final answer is correct. If you find a reference with the correct answer, select that answer. If none of the references contain the correct answer, generate your own correct answer based on your analysis. Explain your evaluation process and reasoning. Finish with (X) at the end of your response, where X is one of A, B, C, or D.

Reference-free generation.
When no reference candidates are provided, agents solve the given multiple-choice question independently and produce a reasoned answer without relying on external references.
System prompt:
You are a assistant. Answer the multiple-choice question carefully and explain your reasoning. Your final answer should be in the form (X) at the end of your response, where X is one of A, B, C, or D.
User Prompt:
Problem:{question}. Solve this multiple-choice question carefully and finish with (X) at the end of your response, where X is one of A, B, C, or D.

Table 10: Prompt design used in the multi-agent debate (MAD) process. The table summarizes agent behavior under reference-aware and reference-free settings.

B Experimental Results

B.1 Results of accuracy changing with rounds

Tables 15 and 16 present model accuracy across debate rounds alongside corresponding upper bounds (parenthesized), facilitating analysis of discriminative capability with progressively refined references. Tables 17 and 18 provide detailed results for the five debate rounds.

B.1.1 Knowledge-centric Datasets

For MMLU-Professional Medicine (Table 15), the upper bound rises consistently across rounds, generally moving from the 93–94 range in Round 1 to

Model Setting	initial acc	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	74.5	73.6 (93.4)	73.6 (94.5)	72.1 (94.9)	71.0 (95.2)	70.3 (95.6)
qwen-2.5-14b-instruct	85.3	85.7 (96.0)	84.5 (96.3)	84.5 (96.3)	84.1 (96.3)	84.2 (96.3)
qwen-2.5-32b-instruct	88.3	89.7 (96.0)	88.3 (96.0)	88.8 (96.0)	89.9 (96.3)	89.6 (96.3)
qwen3-4b	69.7	66.3 (94.9)	62.3 (95.6)	62.3 (96.0)	54.6 (96.0)	52.3 (96.0)
qwen3-8b	78.2	73.4 (97.4)	71.5 (98.5)	70.2 (98.5)	67.9 (98.5)	64.9 (98.5)
doubao-1.5pro-32k	95.7	95.7 (98.5)	95.7 (98.5)	95.6 (98.5)	95.6 (98.5)	95.4 (98.5)
glm-4-9b-chat	70.4	72.4 (96.7)	70.9 (97.0)	68.0 (97.4)	67.4 (97.8)	64.3 (97.8)

Table 11: Model performance dynamics across debate rounds on **MMLU-professional medicine**. Performance on each round are represented as Averaged Acc (Upper Bound).

Model Setting	initial acc	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	51.0	52.3 (62.7)	52.4 (63.9)	51.2 (66.3)	50.7 (66.3)	51.2 (66.3)
qwen-2.5-14b-instruct	51.2	52.8 (59.6)	52.9 (59.6)	53.9 (60.2)	53.3 (60.2)	52.7 (60.2)
qwen-2.5-32b-instruct	53.0	52.7 (60.2)	52.7 (60.8)	52.5 (61.5)	51.3 (61.5)	52.7 (61.5)
qwen3-4b	45.5	42.3 (63.9)	40.0 (67.5)	36.9 (69.3)	35.3 (69.3)	36.4 (69.3)
qwen3-8b	50.1	50.5 (68.7)	48.4 (69.9)	48.1 (69.9)	45.5 (71.1)	43.9 (72.3)
doubao-1.5pro-32k	59.5	60.8 (62.1)	61.1 (62.1)	60.8 (62.1)	61.0 (62.1)	60.5 (62.1)
glm-4-9b-chat	45.5	48.8 (66.3)	48.8 (68.1)	46.9 (68.7)	48.1 (68.7)	48.0 (71.1)

Table 12: Model performance dynamics across debate rounds on **MMLU-virology**. Performance on each round are represented as Averaged Acc (Upper Bound).

96–98 in Round 3. Observed accuracy, however, diverges from this trajectory. For instance, qwen-2.5-7b-instruct holds steady at 73.6 through Round 2 before falling to 72.1 in Round 3, even as the upper bound climbs from 93.4 to 94.9. Therefore, the performance gap relative to the upper bound widens from 19.8 in Round 1 to 22.8 by Round 3. Qwen3-4b displays a similar pattern; its accuracy declines from 66.3 to 62.3 while the upper bound improves from 94.9 to 96.0, expanding the gap from 28.6 to 33.7. Gains remain marginal even for capable models. Qwen-2.5-32b-instruct rises slightly to 89.7 in Round 1 but settles at 88.8 in Round 3, despite a stable upper bound of 96.0. Only doubao-1.5-pro-32k approaches saturation, maintaining a minimal gap of approximately 2.8–2.9 throughout the rounds.

This divergence is more acute on MMLU-Virology. Qwen3-8b accuracy falls from 50.5 in Round 1 to 48.1 in Round 3, whereas the upper bound grows from 68.7 to 69.9, stretching the gap from 18.2 to 21.8. Smaller models like qwen3-4b suffer a continuous decline (42.3 to 36.9), despite the upper bound rising from 63.9 to 69.3. Overall, while stronger reference arguments consistently elevate the upper bound, most models fail to narrow the performance gap. The gap frequently persists or expands during the debate, suggesting that the models struggle to leverage stronger evidence and lack sufficient discriminative capability in knowledge-centric contexts.

B.1.2 Reasoning-centric Datasets

In contrast, reasoning-centric datasets (Table 16) display distinct round-level dynamics, defined by early and sustained reductions in the upper-bound gap. On Math500, several models convert stronger reference arguments into immediate accuracy gains. For instance, Qwen-2.5-32b-instruct improves from an initial 75.2 to 78.4 in Round 1, whereas the upper bound increases marginally from 86.6 to 86.8, narrowing the per-round gap from 11.4 to 8.4. Although accuracy dips to 78.0 by Round 3, the final gap (8.8) remains well below the initial value, reflecting persistent utilization of stronger references. Qwen3-4b follows a similar trajectory; accuracy rises from 77.7 to 81.3 in Round 1 as the upper bound moves from 86.6 to 88.0, compressing the gap from 8.9 to 6.7. Even if accuracy stabilizes or declines slightly in later rounds, the gap stays consistently smaller than at initialization, implying the model capitalizes on higher-quality reasoning early in the process. On MMLU–Formal Logic, round-level gap reduction becomes more distinct for larger models. Qwen-2.5-32b-instruct shows a monotonic accuracy increase from 74.3 to 81.8, while the upper bound rises modestly from 92.1 to 92.9, shrinking the gap from 17.8 to 11.1. Qwen-2.5-14b-instruct exhibits a comparable, though smaller, gap reduction from 17.5 to 15.7. These results suggest that in

Model Setting	initial acc	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	70.6	71.9 (86.2)	70.7 (86.4)	69.6 (86.6)	70.9 (86.8)	70.2 (86.8)
qwen-2.5-14b-instruct	73.1	74.5 (84.8)	74.4 (85.0)	74.8 (85.0)	74.8 (85.2)	73.9 (85.6)
qwen-2.5-32b-instruct	75.2	78.4 (86.6)	78.5 (86.8)	78.0 (86.8)	77.9 (87.0)	77.4 (87.0)
qwen3-4b	77.7	81.3 (86.6)	81.0 (88.0)	80.8 (88.4)	80.5 (88.4)	80.4 (88.4)
qwen3-8b	73.9	78.3 (83.2)	77.4 (85.6)	76.3 (85.8)	76.0 (85.8)	75.7 (85.8)
doubao-1.5pro-32k	84.2	85.2 (89.6)	84.8 (90.8)	84.2 (90.8)	84.0 (90.8)	83.4 (90.8)
glm-4-9b-chat	42.5	43.2 (66.6)	41.5 (67.8)	38.5 (69.8)	39.5 (70.8)	37.6 (71.8)

Table 13: Model performance dynamics across debate rounds on **Math500**. Performance on each round are represented as Averaged Acc (Upper Bound).

Model Setting	initial acc	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	59.1	61.4 (81.8)	60.3 (83.3)	61.1 (83.3)	60.3 (84.9)	60.3 (86.5)
qwen-2.5-14b-instruct	69.8	70.2 (87.3)	69.7 (87.3)	71.6 (87.3)	73.2 (87.3)	73.5 (88.9)
qwen-2.5-32b-instruct	74.3	76.2 (92.1)	78.6 (92.1)	81.8 (92.9)	81.0 (92.9)	78.9 (92.9)
qwen3-4b	66.0	70.3 (92.1)	62.5 (93.7)	62.9 (93.7)	61.4 (93.7)	60.8 (93.7)
qwen3-8b	72.1	78.6 (95.2)	74.0 (100)	74.8 (100)	71.6 (100)	73.0 (100)
doubao-1.5pro-32k	94.1	95.2 (96.8)	95.1 (96.8)	95.2 (96.8)	95.2 (96.8)	95.2 (96.8)
glm-4-9b-chat	56.7	56.8 (84.9)	54.0 (87.3)	56.7 (90.5)	57.3 (91.3)	54.9 (92.1)

Table 14: Model performance dynamics across debate rounds on **MMLU-formal logic**. Performance on each round are represented as Averaged Acc (Upper Bound).

reasoning-centric tasks, models consistently translate stronger references into improved decisions rather than absorbing them as noise. Reasoning-centric datasets feature systematic gap compression rather than expansion at the round level. As stronger reference arguments appear, most models narrow or maintain the upper-bound gap instead of diverging from the optimum. This behavior contrasts with knowledge-centric settings, indicating that discriminative capability is far more stable when evaluation emphasizes structured reasoning over factual recall.

Jointly, Tables 15 and 16 reveal a clear category-level separation in discriminative behavior. Measured by final accuracy, models consistently outperform on reasoning-centric tasks compared to knowledge-centric ones. Importantly, when evaluated via process-level metrics like trend alignment, models show a superior ability to track and respond to argument quality improvements in reasoning contexts. In contrast, knowledge-centric tasks expose a structural weakness: models frequently fail to distinguish correct information from plausible distractors, even when higher-quality references are explicitly available. This analysis demonstrates that progressive debate does not uniformly improve model performance. Instead, it functions as a diagnostic mechanism exposing where discriminative capability genuinely exists and where it collapses. While current LLMs can use structured reasoning to refine judgments under argumentative pressure, they remain highly susceptible to noise in knowledge-intensive scenarios. These findings emphasize the necessity of evaluating LLMs not only by their generation output but by their reliability in discriminating among competing alternatives under controlled conditions.

B.2 Trend Alignment Visualizations

We provide complete trend alignment (TA) visualizations for all evaluated datasets and models. Figures 3, 4, 5, and 6 display the Pearson correlation between the per-round upper bound and model accuracy, quantifying the adherence of model performance to the attainable frontier across rounds. These plots supplement the aggregated statistics in the main text by isolating alignment patterns specific to each model and dataset.

B.3 Effect of Different Judge Models

To evaluate the sensitivity of our framework regarding the judge model selection, we execute a controlled comparison on two representative datasets: MMLU-Virology (knowledge-centric) and MMLU-Formal Logic (reasoning-centric). For each dataset, we perform parallel evaluations employing gemini-2.5-pro and openai-5.1pro as external judges. In both configurations, the judge functions solely to filter and select the best-supported candidate references, serving as progressively stronger evidence in subsequent debate rounds. We analyze the performance dynamics of four evaluated models under

Model Setting	MMLU-profession medicine				MMLU-virology			
	initial acc	round1	round2	round3	initial acc	round1	round2	round3
qwen-2.5-7b-instruct	74.5	73.6(93.4)	73.6(94.5)	72.1(94.9)	51.0	52.3(62.7)	52.4(63.9)	51.2(66.3)
qwen-2.5-14b-instruct	85.3	85.7(96.0)	84.5(96.3)	84.5(96.3)	51.2	52.8(59.6)	52.9(59.6)	53.9(60.2)
qwen-2.5-32b-instruct	88.3	89.7(96.0)	88.3(96.0)	88.8(96.0)	53.0	52.7(60.2)	52.7(60.8)	52.5(61.5)
qwen3-4b	69.7	66.3(94.9)	62.3(95.6)	62.3(96.0)	45.5	42.3(63.9)	40.0(67.5)	36.9(69.3)
qwen3-8b	78.2	73.4(97.4)	71.5(98.5)	70.2(98.5)	50.1	50.5(68.7)	48.4(69.9)	48.1(69.9)
doubao-1.5pro-32k	95.7	95.7(98.5)	95.7(98.5)	95.6(98.5)	59.5	60.8(62.1)	61.1(62.1)	60.8(62.1)
glm-4-9b-chat	70.4	72.4(96.7)	70.9(97.0)	68.0(97.4)	45.5	48.8(66.3)	48.8(68.1)	46.9(68.7)

Table 15: Model performance dynamics across debate rounds on **knowledge-centric datasets** under our evaluation framework. The table reports how model capabilities evolve as progressively stronger reference arguments are introduced.

Model Setting	Math500				MMLU-formal logic			
	initial acc	round1	round2	round3	initial acc	round1	round2	round3
qwen-2.5-7b-instruct	70.6	71.9(86.2)	70.7(86.4)	69.6(86.6)	59.1	61.4(81.8)	60.3(83.3)	61.1(83.3)
qwen-2.5-14b-instruct	73.1	74.5(84.8)	74.4(85.0)	74.8(85.0)	69.8	70.2(87.3)	69.7(87.3)	71.6(87.3)
qwen-2.5-32b-instruct	75.2	78.4(86.6)	78.5(86.8)	78.0(86.8)	74.3	76.2(92.1)	78.6(92.1)	81.8(92.9)
qwen3-4b	77.7	81.3(86.6)	81.0(88.0)	80.8(88.4)	66.0	70.3(92.1)	62.5(93.7)	62.9(93.7)
qwen3-8b	73.9	78.3(83.2)	77.4(85.6)	76.3(85.8)	72.1	78.6(95.2)	74.0(100)	74.8(100)
doubao-1.5pro-32k	84.2	85.2(89.6)	84.8(90.8)	84.2(90.8)	94.1	95.2(96.8)	95.1(96.8)	95.2(96.8)
glm-4-9b-chat	42.5	43.2(66.6)	41.5(67.8)	38.5(69.8)	56.7	56.8(84.9)	54.0(87.3)	56.7(90.5)

Table 16: Model performance dynamics across debate rounds on **reasoning-centric datasets** under our evaluation framework. The table illustrates how judgment quality changes as models are exposed to progressively refined reasoning rather than increased context.

these two judge configurations. Tables 19 and 20 summarize the results for MMLU-Virology and MMLU-Formal Logic, respectively.

The data reveal a distinct contrast in model behavior between knowledge-centric and reasoning-centric environments under different judge models. For MMLU-Virology (19), model performance appears unstable across debate rounds and sensitive to reference selection quality. Even with the introduction of progressively stronger references, most models fail to consistently surpass their initial accuracy. Specifically, Qwen3-4b and GLM-4-9b-chat exhibit noticeable degradation across rounds under both judges; this suggests that exposure to additional arguments, despite filtration, can mislead models in knowledge-heavy scenarios. Although doubao-1.5-pro-32k shows modest gains under both judges, these improvements remain limited, suggesting that current models struggle to distinguish correct domain knowledge from plausible but incorrect alternatives. In contrast, results on the MMLU-Formal Logic dataset (Table 20) display a divergent pattern. For several models, particularly qwen-2.5-32b-instruct and doubao-1.5-pro-32k, accuracy improves in early debate rounds, indicating that models utilize progressively refined reasoning references to refine judgment in logic-intensive tasks. Although performance fluctuates or degrades slightly in later rounds for some models, the general trend implies stronger discriminative capability when reasoning structure, rather than factual recall, constitutes the primary requirement. Crucially, these trends are consistent across both gemini-2.5-pro and openai-5.1 judges, implying that the observed differences arise from task characteristics rather than judge-specific biases. Collectively, these results expose a structural asymmetry: current LLMs exhibit stronger yet fragile judgment ability in reasoning-centric settings, while remaining highly vulnerable to misinformation and reference noise in knowledge-centric evaluation.

B.4 Effect of Candidate Reference Quality

Tables 21 and 22 examine the impact of candidate reference quality by contrasting debate dynamics between weakly supported and well-supported reference queues across reasoning-centric and knowledge-centric tasks. This controlled reverse-selection setup enables a direct assessment of whether improvements in debate-based evaluation stem from stronger arguments or merely from additional interaction. A consistent pattern appears across both datasets: well-supported references yield stable, superior performance trajectories,

Model Setting	MMLU-profession medicine					MMLU-virology				
	round1	round2	round3	round4	round5	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	73.6(93.4)	73.6(94.5)	72.1(94.9)	71.0(95.2)	70.3(95.6)	52.3(62.7)	52.4(63.9)	51.2(66.3)	50.7(66.3)	51.2(66.3)
qwen-2.5-14b-instruct	85.7(96.0)	84.5(96.3)	84.5(96.3)	84.1(96.3)	84.2(96.3)	52.8(59.6)	52.9(59.6)	53.9(60.2)	53.3(60.2)	52.7(60.2)
qwen-2.5-32b-instruct	89.7(96.0)	88.3(96.0)	88.8(96.0)	89.9(96.3)	89.6(96.3)	52.7(60.2)	52.7(60.8)	52.5(61.5)	51.3(61.5)	52.7(61.5)
qwen3-4b	66.3(94.9)	62.3(95.6)	62.3(96.0)	54.6(96.0)	52.3(96.0)	42.3(63.9)	40.0(67.5)	36.9(69.3)	35.3(69.3)	36.4(69.3)
qwen3-8b	73.4(97.4)	71.5(98.5)	70.6(98.5)	67.9(98.5)	64.9(98.5)	50.5(68.7)	48.4(69.9)	48.1(69.9)	45.5(71.1)	43.9(72.3)
doubao-1.5pro-32k	95.7(98.5)	95.7(98.5)	95.6(98.5)	95.6(98.5)	95.4(98.5)	60.8(62.1)	61.1(62.1)	60.8(62.1)	61.0(62.1)	60.5(62.1)
glm-4-9b-chat	72.4(96.7)	70.9(97.0)	68.0(97.4)	67.4(97.8)	64.3(97.8)	48.8(66.3)	48.8(68.1)	46.9(68.7)	48.1(68.7)	48.0(71.1)

Table 17: Model performance dynamics across five debate rounds on **knowledge-centric datasets** under our evaluation framework. The table reports how model capabilities evolve as progressively stronger reference arguments are introduced.

Model Setting	Math500					MMLU-formal logic				
	round1	round2	round3	round4	round5	round1	round2	round3	round4	round5
qwen-2.5-7b-instruct	71.9(86.2)	70.7(86.4)	69.6(86.6)	70.9(86.8)	70.2(86.8)	61.4(81.8)	60.3(83.3)	61.1(83.3)	60.3(84.9)	60.3(86.5)
qwen-2.5-14b-instruct	74.5(84.8)	74.4(85.0)	74.8(85.0)	74.8(87.0)	73.9(87.0)	70.2(87.3)	69.7(87.3)	71.6(87.3)	73.2(87.3)	73.5(88.9)
qwen-2.5-32b-instruct	78.4(86.6)	78.5(86.8)	78.0(86.8)	77.9(85.2)	77.4(85.6)	76.2(92.1)	78.6(92.1)	81.8(92.9)	81.0(92.9)	78.9(92.9)
qwen3-4b	81.3(86.6)	81.0(88.0)	80.8(88.4)	80.5(88.4)	80.4(88.4)	70.3(92.1)	62.5(93.7)	62.9(93.7)	61.4(93.7)	60.8(93.7)
qwen3-8b	78.3(83.2)	77.4(85.6)	76.3(85.8)	76.0(85.8)	75.7(85.8)	78.6(95.2)	74.0(100)	74.8(100)	71.6(100)	73.0(100)
doubao-1.5pro-32k	85.2(89.6)	84.8(90.8)	84.2(90.8)	84.0(90.8)	83.4(90.8)	95.2(96.8)	95.1(96.8)	95.2(96.8)	95.2(96.8)	95.2(96.8)
glm-4-9b-chat	43.2(66.6)	41.5(67.8)	38.5(69.8)	39.5(70.8)	37.6(71.8)	56.8(84.9)	54.0(87.3)	56.7(90.5)	57.3(91.3)	54.9(92.1)

Table 18: Model performance dynamics across five debate rounds on **reasoning-centric datasets** under our evaluation framework. The table illustrates how judgment quality changes as models are exposed to progressively refined reasoning rather than increased context.

whereas weakly supported references frequently degrade model accuracy over the debate rounds. On the reasoning-centric MMLU-Formal Logic dataset (Table 21), several models maintain or modestly improve performance given high-quality references, suggesting they utilize structured, reliable reasoning to refine judgments. In contrast, under exposure to weakly supported candidates, performance frequently stagnates or declines, indicating that low-quality reasoning disrupts the refinement process even in logic-oriented tasks. This divergence intensifies on the knowledge-centric MMLU-Virology dataset (Table 22). With weakly supported references, most models suffer pronounced performance drops across debate rounds, reflecting high vulnerability to misleading or poorly grounded information. Although well-supported references partially mitigate this effect to yield stable outcomes, gains remain limited, exposing the inherent difficulty models face in distinguishing correct domain knowledge from plausible incorrect alternatives.

Collectively, these results demonstrate that debate effectiveness depends critically on reference quality rather than the debate process itself. Poor-quality references can actively mislead models and mask true judgment capability, whereas well-supported candidates create conditions where discriminative behavior is observable. This finding reinforces the necessity of quality-aware candidate selection, validating our framework's focus on progressive argument refinement over indiscriminate context accumulation.

B.5 Effect of Different Number of Agent

To examine the influence of debate scale on model behavior, we conduct an ablation study by varying the number of participating agents. We evaluate each model using 3, 5, and 7 agents on both a reasoning-centric dataset (MMLU-Formal Logic) and a knowledge-centric dataset (MMLU-Virology). This experiment investigates whether increasing the agent count improves judgment quality or introduces noise that impedes effective discrimination. On the MMLU-Formal Logic dataset (Table 23), moderate agent counts yield performance superior to or more stable than smaller or larger configurations. With 3 agents, several models (e.g., qwen-2.5-32b-instruct and doubao-1.5-pro-32k) show clear improvements over initial accuracy, indicating that limited debate aids models in utilizing refined reasoning references. Increasing the agent count to 5 maintains or improves performance for select models, whereas scaling to 7 often results in stagnation or mild degradation. This trend suggests that while additional agents introduce diverse reasoning paths, excessive inter-

Model Setting	MMLU-virology(openai-5.1)				MMLU-virology(gemini-2.5-pro)			
	initial acc	round1	round2	round3	initial acc	round1	round2	round3
qwen-2.5-32b-instruct	53.0	53.3(60.2)	53.0(61.5)	53.4(62.1)	53.0	52.7(60.2)	52.7(60.8)	52.5(61.5)
qwen3-4b	45.5	41.0(63.9)	41.5(67.5)	40.1(68.1)	45.5	42.3(63.9)	40.0(67.5)	36.9(69.3)
doubao-1.5pro-32k	59.5	60.8(62.1)	61.0(62.1)	61.2(62.1)	59.5	60.8(62.1)	61.1(62.1)	60.8(62.1)
glm-4-9b-chat	45.5	50.7(66.3)	48.9(69.7)	49.5(70.5)	45.5	48.8(66.3)	48.8(68.1)	46.9(68.7)

Table 19: Performance dynamics under different judge models on the **MMLU-Virology** dataset. Results compare four evaluated models across debate rounds when using **gemini-2.5-pro** and **openai-5.1** as the external judge.

Model Setting	MMLU-formal logic(openai-5.1)				MMLU-formal logic(gemini-2.5-pro)			
	initial acc	round1	round2	round3	initial acc	round1	round2	round3
qwen-2.5-32b-instruct	74.3	80.0(92.1)	78.6(93.7)	78.4(93.7)	74.3	76.2(92.1)	78.6(92.1)	81.8(92.9)
qwen3-4b	66.0	69.5(92.1)	67.9(96.8)	64.6(96.8)	66.0	70.3(92.1)	62.5(93.7)	62.9(93.7)
doubao-1.5pro-32k	94.0	95.1(96.8)	95.1(96.8)	94.9(96.8)	94.1	95.2(96.8)	95.1(96.8)	95.2(96.8)
glm-4-9b-chat	56.7	52.4(84.9)	54.9(88.1)	54.4(88.9)	56.7	56.8 (84.9)	54.0(87.3)	56.7(90.5)

Table 20: Performance dynamics under different judge models on the **MMLU-Formal Logic** dataset. Results compare four evaluated models across debate rounds when using **gemini-2.5-pro** and **openai-5.1** as the external judge.

action may dilute the value of refined arguments, complicating the judgment process. In contrast, results on the MMLU-Virology dataset (Table 24) show a more acute sensitivity to agent count. For most models, increasing the agent count fails to yield consistent gains and often causes performance degradation, particularly in later rounds. Even capable models show limited improvement beyond the initial round, whereas smaller models become increasingly susceptible to errors as the agent count rises. This behavior indicates that in knowledge-intensive settings, additional agents introduce noise that overwhelms discriminative capacity rather than reinforcing judgment.

Collectively, these findings expose a distinct asymmetry between reasoning-centric and knowledge-centric tasks. For reasoning tasks, a moderate agent count supports better judgment by presenting complementary reasoning structures, yet this benefit diminishes as the scale grows. In knowledge-centric tasks, a higher agent count typically exacerbates misinformation and reference noise, undermining discriminative performance. These results validate the necessity of controlling debate scale and context complexity, supporting our progressive argument refinement design over unconstrained multi-agent expansion.

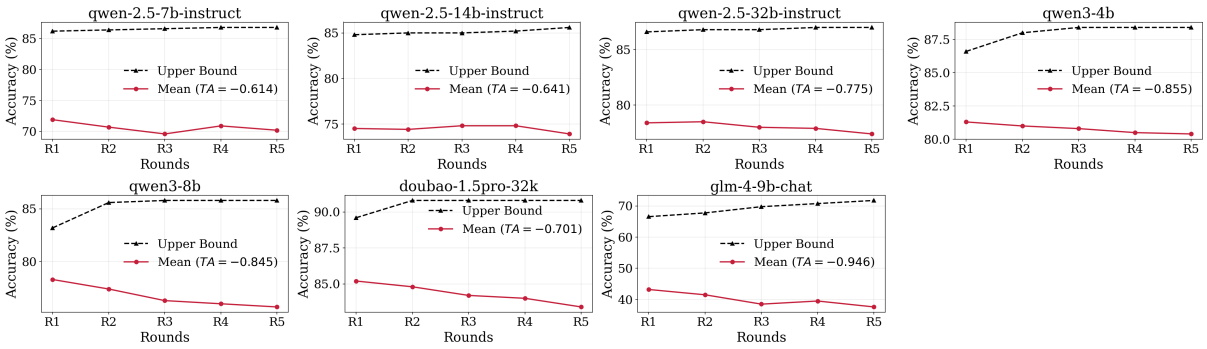


Figure 3: Trend alignment between upper bound and accuracy across debate rounds on Math500.

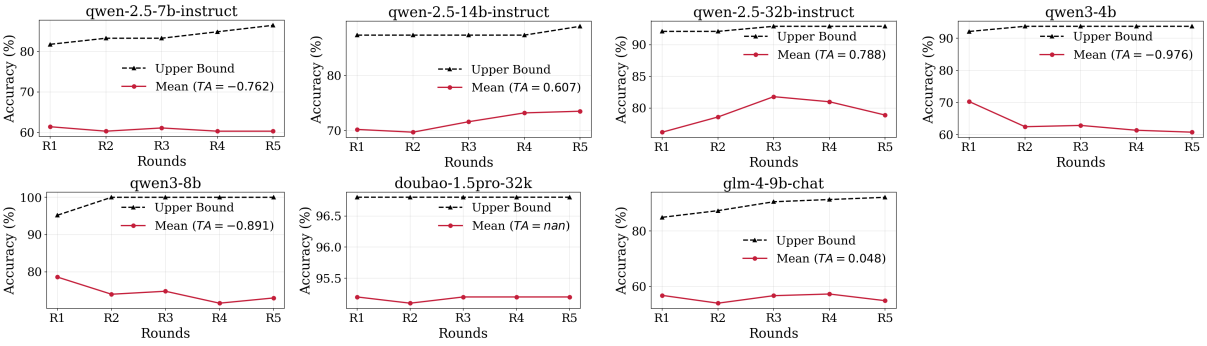


Figure 4: Trend alignment between upper bound and accuracy across debate rounds on MMLU-Formal Logic.

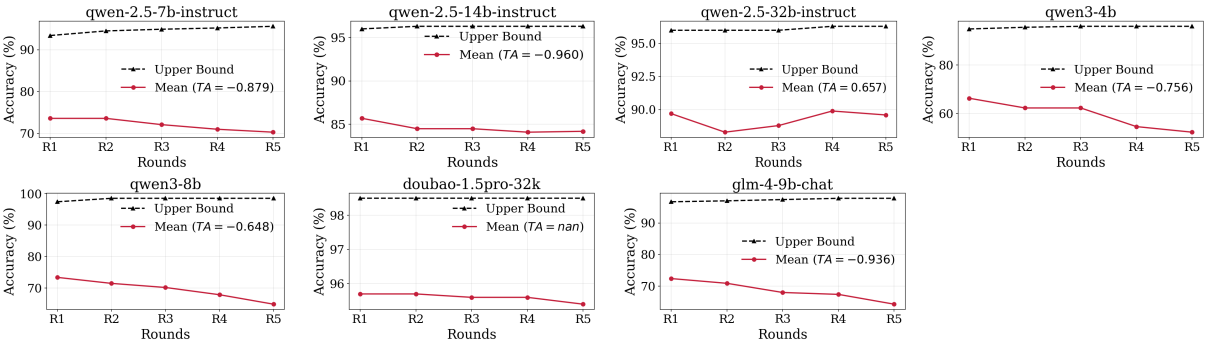


Figure 5: Trend alignment between upper bound and accuracy across debate rounds on MMLU-Professional Medicine.

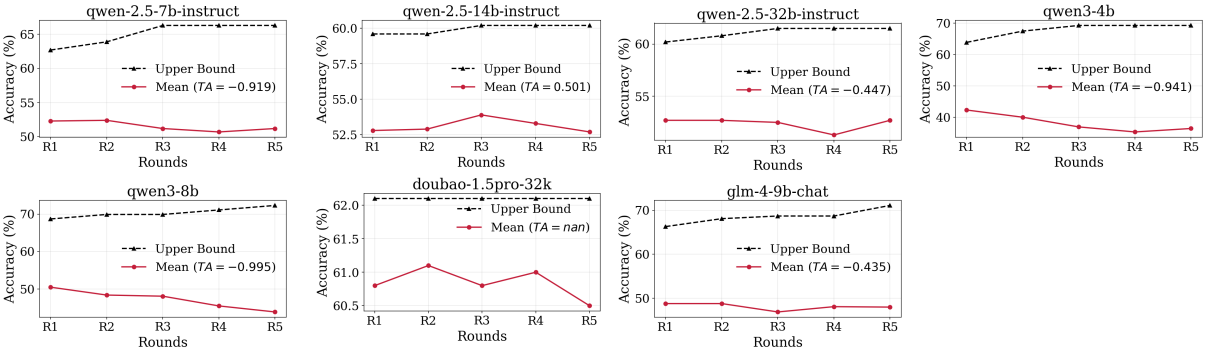


Figure 6: Trend alignment between upper bound and accuracy across debate rounds on MMLU-Virology.

Model Setting	MMLU-formal logic(wak supported)				MMLU-formal logic(well-supported)			
	initial acc	round1	round2	round3	initial acc	round1	round2	round3
qwen-2.5-32b-instruct	74.3	80.0(92.1)	78.6(95.2)	78.4(95.2)	74.3	76.2(92.1)	78.6(92.1)	81.8(92.9)
qwen3-4b	66.0	69.5(92.1)	67.9(95.2)	64.6(96.8)	66.0	70.3(92.1)	62.5(93.7)	62.9(93.7)
doubao-1.5pro-32k	94.0	95.1(96.8)	95.1(97.6)	94.9(97.6)	94.1	95.2(96.8)	95.1(96.8)	95.2(96.8)
glm-4-9b-chat	56.7	52.4(84.9)	54.9(88.9)	54.4(90.5)	56.7	56.8 (84.9)	54.0(87.3)	56.7(90.5)

Table 21: Performance dynamics on **MMLU-Formal Logic** under different candidate reference qualities. We compare debate-round accuracy trajectories when the reference queue is constructed using **weakly supported** (low-scoring) versus **well-supported** (high-scoring) candidates.

Model Setting	MMLU-virology (weak supported)				MMLU-virology (well-supported)			
	initial acc.	round1	round2	round3	initial acc.	round1	round2	round3
qwen-2.5-32b-instruct	53.0	53.1(60.2)	51.5(62.1)	51.0(62.1)	53.0	52.7(60.2)	52.7(60.8)	52.5(61.5)
qwen3-4b	45.5	44.9(63.9)	38.4(68.1)	39.3(68.7)	45.5	42.3(63.9)	40.0(67.5)	36.9(69.3)
doubao-1.5pro-32k	59.5	60.8(61.2)	60.2(62.1)	60.2(62.1)	59.5	60.8(62.1)	61.1(62.1)	60.8(62.1)
glm-4-9b-chat	45.5	48.8(66.3)	45.9(69.3)	42.8(71.7)	45.5	48.8(66.3)	48.8(68.1)	46.9(68.7)

Table 22: Performance dynamics on **MMLU-Virology** under different candidate reference qualities. We compare debate-round accuracy trajectories when the reference queue is constructed using **weakly supported** (low-scoring) versus **well-supported** (high-scoring) candidates.

Model Setting	3 Agents			5 Agents			7 Agents		
	initial acc	round1	round 2	initial acc	round 1	round2	initial acc	round1	round2
qwen-2.5-32b-instruct	78.0	81.8(90.5)	80.7(90.5)	74.3	76.2(92.1)	78.6(92.1)	75.3	77.0(96.8)	76.5(96.8)
qwen3-4b	76.5	81.2(88.9)	79.4(95.2)	66.0	70.3(92.1)	62.5(93.7)	66.2	68.3(95.2)	66.0(98.4)
doubao-1.5pro-32k	93.9	95.0(96.8)	94.2(96.8)	94.1	95.2(96.8)	95.1(96.8)	93.7	93.0(97.6)	92.6(97.6)
glm-4-9b-chat	55.0	55.0(78.6)	54.0(79.4)	56.7	56.8 (84.9)	54.0(87.3)	57.1	56.7(89.7)	56.0(92.9)

Table 23: Impact of agent count on model performance on the **MMLU-Formal Logic** dataset under the proposed evaluation framework.

Model Setting	3 Agents			5 Agents			7 Agents		
	initial acc	round1	round 2	initial acc	round 1	round2	initial acc	round1	round2
qwen-2.5-32b-instruct	53.0	54.0(59.0)	53.4(59.6)	53.0	52.7(60.2)	52.7(60.8)	51.7	52.2(57.8)	52.2(58.4)
qwen3-4b	44.6	42.4(59.0)	42.8(62.1)	45.5	42.3(63.9)	40.0(67.5)	50.3	43.9(68.1)	38.4(73.5)
doubao-1.5pro-32k	59.8	60.4(62.1)	60.4(62.1)	59.5	60.8(62.1)	61.1(62.1)	59.7	58.9(62.7)	59.0(62.7)
glm-4-9b-chat	43.0	47.8(61.5)	50.4(63.3)	45.5	48.8(66.3)	48.8(68.1)	45.0	47.1(67.5)	45.9(69.9)

Table 24: Impact of agent count on model performance on the **MMLU-Virology** dataset under the proposed evaluation framework.

Model Setting	MMLU-virology (openai-5.1)			MMLU-virology (gemini-2.5-pro)		
	round1	round2	round3	round1	round2	round3
qwen-2.5-32b-instruct	6.9	8.5	8.7	7.5	8.1	9.0
qwen3-4b	22.9	26.0	28.0	21.6	27.5	32.4
doubao-1.5pro-32k	1.3	1.1	0.9	1.3	1.0	1.3
glm-4-9b-chat	15.6	20.8	21.0	17.5	19.3	21.8

Table 25: Upper-bound gaps across debate rounds on **MMLU-Virology**, under different judge models.

Model Setting	MMLU-formal logic (openai-5.1)			MMLU-formal logic (gemini-2.5-pro)		
	round1	round2	round3	round1	round2	round3
qwen-2.5-32b-instruct	12.1	15.1	15.3	15.9	13.5	11.1
qwen3-4b	22.6	28.9	32.2	21.8	31.2	30.8
doubao-1.5pro-32k	1.7	1.7	1.9	1.6	1.7	1.6
glm-4-9b-chat	32.5	33.2	34.5	28.1	33.3	33.8

Table 26: Upper-bound gaps across debate rounds on **MMLU-Formal Logic**, under different judge models.

Model Setting	MMLU-formal logic (weak supported)			MMLU-formal logic (well-supported)		
	round1	round2	round3	round1	round2	round3
qwen-2.5-32b-instruct	12.1	16.6	16.8	15.9	13.5	11.1
qwen3-4b	22.6	27.3	32.2	21.8	31.2	30.8
doubao-1.5pro-32k	1.7	2.5	2.7	1.6	1.7	1.6
glm-4-9b-chat	32.5	34.0	36.1	28.1	33.3	33.8

Table 27: Upper-bound gaps on **MMLU-Formal Logic** under different candidate reference qualities. Results compare gap dynamics when the reference queue is constructed from **weakly supported** versus **well-supported** candidates.

Model Setting	MMLU-virology (weak supported)			MMLU-virology (well-supported)		
	round1	round2	round3	round1	round2	round3
qwen-2.5-32b-instruct	7.1	10.6	11.1	7.5	8.1	9.0
qwen3-4b	19.0	29.7	29.4	21.6	27.5	32.4
doubao-1.5pro-32k	0.4	1.9	1.9	1.3	1.0	1.3
glm-4-9b-chat	17.5	23.4	28.9	17.5	19.3	21.8

Table 28: Upper-bound gaps on **MMLU-Virology** under different candidate reference qualities. Results compare gap dynamics when the reference queue is constructed from **weakly supported** versus **well-supported** candidates.

Model Setting	3 Agents			5 Agents			7 Agents		
	round1	round2	round3	round1	round2	round3	round1	round2	round3
qwen-2.5-32b-instruct	8.7	9.8	10.1	15.9	13.5	11.1	19.8	20.3	18.1
qwen3-4b	7.7	15.8	16.1	21.8	31.2	30.8	26.9	32.4	34.1
doubao-1.5pro-32k	1.8	2.6	1.8	1.6	1.7	1.6	4.6	5.0	5.2
glm-4-9b-chat	23.6	25.4	29.6	28.1	33.3	33.8	33.0	36.9	37.7

Table 29: Upper-bound gaps (upper bound minus mean accuracy) on the **MMLU-Formal Logic** dataset under different agent counts.

Model Setting	3 Agents			5 Agents			7 Agents		
	initial acc	round1	round 2	initial acc	round 1	round2	initial acc	round3	round2
qwen-2.5-32b-instruct	5.0	6.2	6.2	7.5	8.1	9.0	5.6	6.2	5.9
qwen3-4b	16.6	19.3	21.1	21.6	27.5	32.4	24.2	35.1	40.1
doubao-1.5pro-32k	1.7	1.7	1.5	1.3	1.0	1.3	3.8	3.7	3.5
glm-4-9b-chat	13.7	12.9	19.3	17.5	19.3	21.8	20.4	24.0	25.2

Table 30: Upper-bound gaps (upper bound minus mean accuracy) on the **MMLU-Virology** dataset under different agent counts.