

MELD: Multilingual Ensemble via Logical Debate

Sehyeok Kang^{*1} Jaejun Ryu^{*o2} Se-Young Yun^{†1}

KAIST AI¹ Department of Linguistics, Seoul National University²

{kangsehyeok0329, yunseyoung}@kaist.ac.kr, 2fasturtle@snu.ac.kr

Abstract

Large language models can benefit from inference-time strategies such as self-consistency and ensemble-based methods, including multi-agent debate. However, recent work suggests that the effectiveness of these approaches depends on the diversity of the reasoning trajectories they generate. In this work, we consider multilinguality as a practical way to induce such diversity. Whereas monolingual reasoning paths often remain highly correlated, reasoning across multiple languages can encourage exploration of a broader solution space. Motivated by this idea, we propose **MELD**, Multilingual Ensemble via Logical multi-agent Debate, which leverages language variation to produce complementary reasoning trajectories. Across multiple benchmarks and model series, MELD achieves competitive and consistently favorable performance compared to baselines, and our analyses suggest that these gains are associated with improved error complementarity, aggregation, and debate-based correction under multilingual reasoning.

Warning: This paper contains examples of content that exhibits bias.

1 Introduction

Large Language Models (LLMs) have recently shown strong reasoning ability in a wide range of complex tasks (Wei et al., 2022). To further improve this ability, inference-time strategies such as self-consistency (SC) (Wang et al., 2022) and multi-agent debate (MAD) (Du et al., 2023; Estornell & Liu, 2024) are widely used. These methods improve the accuracy of the final answer by aggregating multiple reasoning trajectories based on the idea that diverse intermediate paths can offset individual errors and lead to a more reliable consensus.

However, these approaches depend on the quality of the diversity they produce, but monolingual sampling and homogeneous multi-agent debate often fail to yield independent reasoning trajectories. Even when the outputs differ superficially, they often share the same errors or misconceptions and converge on the same incorrect conclusions (Tan et al., 2025; Estornell & Liu, 2024; Kim et al., 2025). Previous work shows that samples from the same model are often highly correlated, limiting the benefits of aggregation or debate; therefore, increasing the number of samples or agents does not necessarily improve performance.

Existing approaches have introduced stronger heterogeneity, such as different backbone models, designed personas, or heuristic reasoning path diversification strategies (Zhang et al., 2025; Wang et al., 2025c; Zhou & Chen, 2025; Chan et al., 2023; Hu et al., 2024b; Liu et al., 2025). Although these methods can improve diversity, they often involve important trade-offs. Heterogeneous model ensembles increase architectural complexity and computational cost, while persona prompting and heuristic diversification are sensitive to design choices and often require labor-intensive tuning across domains. As a result, obtaining useful reasoning diversity in a scalable, stable, and cost-efficient manner remains challenging.

In this work, we explore multilinguality as a scalable and robust source of reasoning diversity for LLMs. Previous research shows that multilingual prompting increases output diversity

^{*}Equal contribution. [†]Corresponding author.

^oThis work was conducted during the internship at KAIST AI.

Code available at: <https://github.com/cheesebro329/MELD>

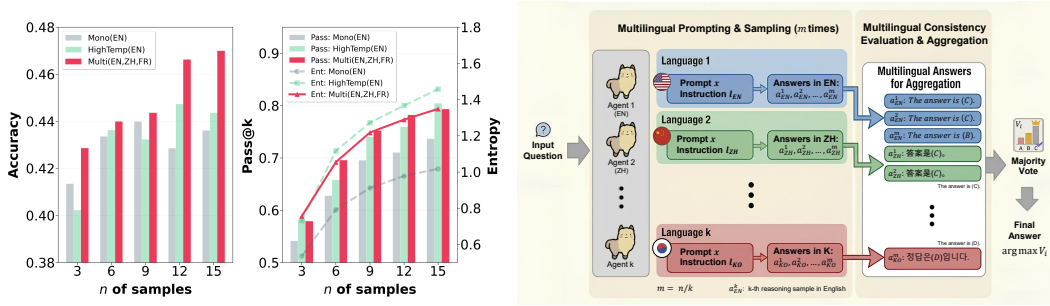


Figure 1: **Motivation examples.** (Left) Multilingual sampling (Multi) yields higher response *entropy* (i.e., greater diversity) and *Pass@k* than monolingual sampling (Mono), which in turn leads to improved *Accuracy* on MMLU-prox-lite. Increasing the temperature in monolingual sampling (HighTemp) also raises response *entropy* and *Pass@k*, but these gains do not translate into higher *Accuracy*. (Right) Example illustrating the multilingual sampling approach.

by eliciting language- and culture-specific variations beyond monolingual prompts (Wang et al., 2025b). We posit that this diversity can also improve reasoning performance, as switching languages does more than restate the same thought process: it can induce distinct reasoning trajectories and decompositions within a single backbone model. As shown in Figure 1, multilingual sampling (Multi) increases response entropy and *pass@k*, the probability of obtaining at least one correct answer among k samples, leading to higher accuracy than monolingual sampling (Mono). These findings suggest that multilinguality provides a scalable way to obtain useful diversity without requiring heterogeneous models.

Inspired by this insight, we propose **MELD** (Multilingual Ensemble via Logical Debate), a multilingual multi-agent debate framework for improving LLM reasoning. In MELD, agents reason about the same problem in different languages, exchange their intermediate reasoning, and refine their views through debate. This language-based variation induces more diverse responses, increasing the likelihood that correct answers emerge early in the debate and thereby improving final-answer accuracy. Experiments on diverse benchmarks, including bias, reading comprehension, and knowledge-intensive reasoning tasks, show that MELD consistently outperforms a wide range of baselines. These results suggest that multilingual debate is a robust and effective strategy to improve reasoning performance.

Contributions. We present multilinguality as a practical and stable source of reasoning diversity in LLMs and introduce a debate framework that leverages it for improved inference. We also analyze when and why it benefits reasoning. Our main contributions are as follows:

- **Multilingual Diversity for Reasoning:** We identify multilinguality as a source of reasoning diversity in LLMs and analyze how multilingual output differs from monolingual samples in ways that affect reasoning performance.
- **MELD Framework:** We propose **MELD**, a multilingual multi-agent debate framework that harnesses multilinguality to elicit trajectory-exploring diversity within a single model.
- **Empirical and Mechanistic Validation:** Through extensive experiments and analyses, we show that MELD consistently improves performance across diverse task domains including reasoning and systematically uncover the mechanistic link showing how multilingually driven diversity contributes to these gains.

2 MELD: Multilingual Ensemble via Logical Debate

2.1 Problem Setup: Multi-Agent Debate

We consider a task defined over an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Given an input problem $x \in \mathcal{X}$, the goal is to produce a correct answer $y \in \mathcal{Y}$. In multi-agent debate, the final prediction is obtained through iterative interactions among multiple agents. Let f_θ

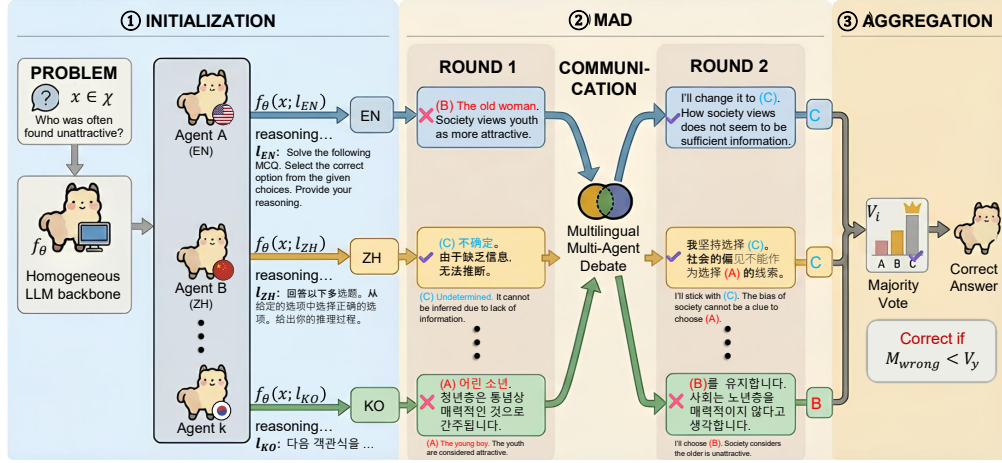


Figure 2: An overview of the MELD framework.

denote a language model with parameters θ . In the same-backbone setting, all agents share the same θ . We employ N agents and allow at most T rounds of debate. For initial round ($t = 1$) and each subsequent round ($t \in \{2, \dots, T\}$), each agent $i \in \{1, \dots, N\}$ generates a reasoning trace and answer:

$$(r_i^1, a_i^1) \sim f_\theta(x; \pi_i), \quad (r_i^t, a_i^t) \sim f_\theta(x, r_i^{t-1}, C_i^{t-1}; \pi_i), \quad (1)$$

where π_i denotes the agent-specific prompting condition (e.g., a role instruction, decoding seed, or diversification strategy). Furthermore, C_i^{t-1} summarizes the reasoning traces and answers of the other agents in round $t - 1$. This allows agent i to update its reasoning by conditioning on both its previous state and the aggregated debate context. The debate continues until the maximum number of rounds T is reached or the agents' answers converge. Let $\tau \leq T$ denote the stopping round. The final prediction is then obtained by applying an aggregation rule to the final-round answers:

$$\hat{y} = \text{AGG}(a_1^\tau, a_2^\tau, \dots, a_N^\tau). \quad (2)$$

2.2 MELD: Multilingual Ensemble via Logical Debate

We propose **MELD** (Multilingual Ensemble via Logical Debate), a multilingual instantiation of the multi-agent debate framework. As illustrated in Figure 2, MELD induces diversity by assigning different reasoning languages to different agents. Let $\mathcal{L}_x = \{\ell_1, \ell_2, \dots, \ell_N\}$ denote the set of reasoning languages assigned to the N agents for problem x , where typically $\ell_i \neq \ell_j$ for $i \neq j$. Each agent then reasons about the same problem in its assigned language. The reasoning traces and answers of agent i are generated as:

$$(r_i^1, a_i^1) \sim f_\theta(x; \ell_i), \quad (r_i^t, a_i^t) \sim f_\theta(x, r_i^{t-1}, C_i^{t-1}; \ell_i). \quad (3)$$

Unlike standard monolingual debate, MELD derives diversity from linguistic variation. At each subsequent round, agent i updates its reasoning based on its previous state and a multilingual debate context C_i^{t-1} . By observing candidate solutions and critiques from other linguistic perspectives, agents refine their reasoning through the debate process. MELD terminates when the agents reach consensus or the maximum number of rounds is exhausted.

2.3 An Analytical View of Aggregation in MELD

To analyze aggregation behavior under plurality voting, we define the vote count for any candidate $z \in \mathcal{Y}$ and its associated metrics as:

$$V_z = \sum_{i=1}^N \mathbf{1}[a_i^\tau = z], \quad \hat{y} = \arg \max_{z \in \mathcal{Y}} V_z, \quad M_{\text{wrong}} = \max_{z \neq \hat{y}} V_z \quad (4)$$

where V_z denotes the support for candidate z , $\hat{y}$ is the final prediction, and M_{wrong} represents the strongest incorrect consensus relative to the ground truth y . The final prediction is correct if and only if $V_y > M_{\text{wrong}}$. This condition shows that aggregation performance depends on two factors: support for the correct answer and concentration of votes on the strongest incorrect competitor. Effective aggregation therefore requires larger V_y and smaller M_{wrong} . Integrating the methodology of MELD, we bound the probability of an erroneous consensus. As derived in Appendix A, this relationship is expressed as:

$$P(M_{\text{wrong}} \geq V_y) \leq \mathcal{C}_{\text{var}} + \mathcal{C}_{\text{cov}} \sum_{i \neq j} \text{Cov}(\xi_i, \xi_j) \quad (5)$$

where ξ_i is an indicator variable denoting whether agent i selects an incorrect candidate. The positive, task-specific term $\mathcal{C}_{\text{var}}$ captures the agents’ individual error tendencies, and thus reflects their competence under the assigned reasoning languages, while $\mathcal{C}_{\text{cov}}$ scales the contribution of correlated errors across agents. Multilinguality can reduce the covariance term by eliciting complementary reasoning behavior, and our URIEL+ analysis finds that greater typological distance is generally associated with lower error overlap (Khan et al., 2025). However, adding more distant or less well-supported languages can also increase individual errors and thereby enlarge the variance term. The bound therefore characterizes a competence–diversity trade-off rather than predicting monotonic improvement with either language distance or the number of languages. When the balance between language-specific competence and cross-language complementarity lowers the bound on failure, it increases the probability of achieving a correct consensus, $P(V_y > M_{\text{wrong}})$.

Empirical Implications. Mechanistically, MELD improves both quantities by inducing diverse reasoning trajectories across languages. Different languages expose complementary paths to the correct answer, increasing the likelihood that some agents identify it and that this signal propagates through the debate process. Simultaneously, multilingual diversity reduces agents’ tendency to rely on shared misconception, lowering the concentration of votes on a single incorrect answer. We examine these implications empirically in Section 3.4 by analyzing whether MELD yields larger V_y , smaller M_{wrong} , and higher accuracy than monolingual baselines.

3 Experiments

We evaluate MELD through three core research questions:

- **RQ1:** Is multilingual reasoning more diverse than monolingual sampling? (§3.3)
- **RQ2:** Can MELD leverage this diversity to improve reasoning performance? (§3.2, §3.4)
- **RQ3:** Why is MELD effective, and when is it most beneficial? (§3.4, §3.5, §3.6, §3.7)

To address these questions, we compare MELD against monolingual and multilingual baselines across varied benchmarks, model families, and analytical settings.

3.1 Experimental Setup

Benchmarks. We evaluate MELD across four benchmarks: **MMLU-prox-lite** (Xuan et al., 2025) for multilingual knowledge-intensive reasoning, **BBQ** (Parrish et al., 2021) and **CBBQ** (Huang & Xiong, 2024) for bias-sensitive reasoning, **TruthfulQA** (Lin et al., 2021) for truthfulness under common misconceptions, and **Belebele** (Bandarkar et al., 2023) for multilingual reading comprehension.

LLM models. We conduct experiments on four open-weight model families: **Llama-3.1** (Dubey et al., 2024), **Aya-23** (Aryabumi et al., 2024), **Gemma-3** (Kamath et al., 2025), and **Qwen-2.5** (Yang et al., 2025b), covering models with different multilingual characteristics. For **Llama-3.1**, **Gemma-3**, and **Qwen-2.5**, we utilize instruction-tuned variants.

Baselines. We compare MELD against inference-time baselines, including direct prompting, reasoning-based sampling, monolingual debate, and diversity-oriented debate vari-

Models	Method	Problem lang: EN 🇺🇸					Problem lang: ZH 🇨🇳			
		MMLU-prox	BBQ	Belebele	TruthfulQA	Avg.	MMLU-prox	BBQ	Belebele	Avg.
Llama-3.1-8B	Direct	44.1	49.0	57.5	74.1	56.2	26.7	41.7	46.5	38.3
	CoT	47.7	57.9	74.1	68.1	62.0	35.7	50.2	74.2	53.4
	SC	51.8	61.2	76.8	71.3	65.3	40.4	52.5	77.7	56.9
	High temp SC	41.7	57.3	74.2	72.9	61.5	22.2	39.3	38.4	33.3
	Mono-MAD	50.4	63.6	80.1	70.8	66.2	38.1	52.7	75.5	55.4
	Persona MAD	52.0	61.1	85.5	63.6	65.6	42.1	49.6	83.9	58.5
	DMAD	49.4	64.5	84.9	71.9	67.7	37.3	52.8	78.3	56.1
	Multilingual SC	50.4	65.0	81.1	71.1	66.9	42.4	45.6	76.0	54.7
	MELD(Ours)	52.1	74.3	86.8	66.5	69.9	42.1	54.0	82.1	59.4
Aya-23-8B	Direct	26.1	50.5	34.3	44.1	38.8	28.1	43.4	53.2	41.6
	CoT	20.4	41.3	44.2	42.4	37.1	26.8	40.7	49.3	38.9
	SC	21.2	45.5	46.0	48.9	40.4	24.8	36.6	47.7	36.4
	High temp SC	18.4	42.5	34.6	42.1	34.4	18.5	31.3	32.4	27.4
	Mono-MAD	20.4	50.3	45.0	49.6	41.3	26.6	42.3	51.3	40.1
	Persona MAD	24.3	35.8	25.1	64.4	37.4	30.0	35.9	49.7	38.5
	DMAD	26.3	50.8	48.5	48.7	43.6	29.7	41.8	50.6	40.7
	Multilingual SC	27.6	54.0	51.4	53.7	46.7	23.7	44.4	53.0	40.4
	MELD(Ours)	24.7	57.2	55.9	53.3	47.8	22.3	49.9	56.6	42.9
Gemma-3-4B	Direct	32.8	67.9	86.7	61.1	62.1	26.8	57.9	76.9	53.9
	CoT	36.5	76.1	85.4	34.9	58.2	36.7	61.1	77.4	58.4
	SC	41.7	78.4	87.9	36.9	61.2	39.1	64.3	80.3	61.2
	High temp SC	42.0	77.5	87.7	35.7	60.7	37.1	63.6	80.4	60.4
	Mono-MAD	45.5	78.3	88.4	47.7	65.0	40.2	64.1	81.7	62.0
	Persona MAD	41.2	56.4	78.6	45.4	55.4	38.2	61.6	79.1	59.6
	DMAD	46.1	78.2	88.3	49.4	65.5	40.4	62.7	82.7	61.9
	Multilingual SC	42.6	77.8	87.0	50.1	64.4	37.3	59.2	83.1	59.9
	MELD(Ours)	47.7	80.5	88.4	50.4	66.8	44.4	63.2	84.2	63.9
Qwen-2.5-3B	Direct	31.5	81.3	87.1	54.9	63.7	29.4	55.3	64.1	49.6
	CoT	39.7	80.9	80.0	56.4	64.3	37.0	66.8	79.5	61.1
	SC	42.5	85.3	82.4	61.2	67.9	37.7	69.7	84.5	64.0
	High temp SC	43.2	84.3	83.8	61.2	68.1	36.5	68.0	83.4	62.6
	Mono-MAD	48.0	85.0	85.9	68.7	71.9	40.1	72.2	81.5	64.6
	Persona MAD	49.5	78.0	84.9	63.6	69.0	40.0	61.8	79.2	60.3
	DMAD	48.4	85.0	88.0	67.7	72.3	40.7	73.4	82.7	65.6
	Multilingual SC	48.2	85.4	86.3	72.5	73.1	43.0	73.5	83.6	66.7
	MELD(Ours)	49.2	83.7	89.0	70.3	73.1	42.6	76.9	83.9	67.8

Table 1: Accuracy (%) of methods across LLMs when each benchmark is presented in English (🇺🇸) or Chinese (🇨🇳) as the input language. For each benchmark within each LLM, the best result in the Avg. columns is highlighted in **red** and the second-best result in **orange**. For the non-Avg. benchmark columns, only the best result is shown in bold. Shaded rows indicate multilingual methods. Across 7 benchmarks and 4 LLMs (28 cases in total), MELD(Ours) ranks first in 14 cases and second in 7 cases, indicating strong and consistent performance.

ants, including Direct, CoT (Wei et al., 2022), self-consistency(SC) (Wang et al., 2022), multilingual SC, Mono-MAD, Persona MAD (Hu et al., 2024b), and DMAD (Liu et al., 2025).

Implementation Details. Unless otherwise stated, the main experiments in Table 1 use $N = 2$ agents assigned English and Chinese as their reasoning languages. We adopt this bilingual setting as the simplest controlled configuration for isolating cross-language variation while maintaining strong model competence in both languages. In each MELD run, all agents receive exactly the same problem instance in the same input language; only their assigned reasoning language differs. At each round, agents receive the other agent’s response from the immediately preceding round, without accumulating the full debate history. The debate runs for at most $T = 3$ rounds and terminates early upon consensus. If the agents remain tied after the final round, we deterministically select the response of the agent whose reasoning language matches the input language. Appendix B provides further implementation details.

3.2 Main Results

Overall Performance. Table 1 summarizes the main results across benchmarks, input-language settings, and model backbones. MELD shows strong and consistent performance

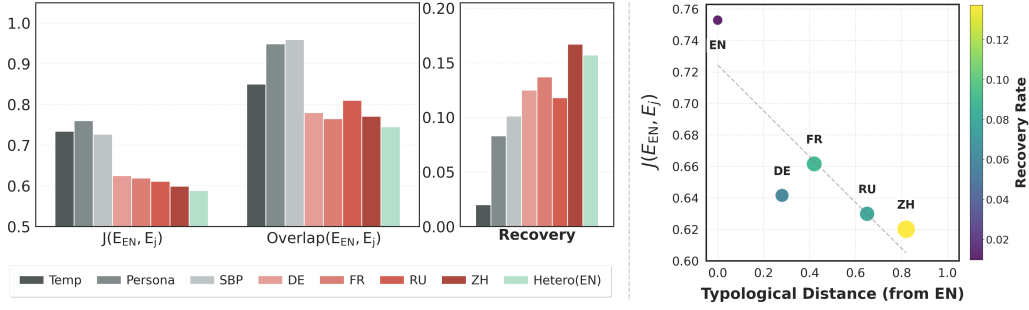


Figure 3: Analysis of complementarity across reasoning diversification methods and language distance. **(Left)** Error Jaccard similarity, overlap coefficient, and recovery rate for samples generated by reasoning in English versus other languages (DE, FR, RU, ZH), as well as by diversity-inducing methods including temperature scaling (Temp), persona prompting (Persona), and Step-Back Prompting (SBP). **(Right)** Error Jaccard similarity and recovery rate analyzed with respect to typological distance between languages.

across a wide range of settings. Specifically, MELD ranks first in 14 out of 28 cases and second in 7, confirming that its gains are not limited to a small subset of tasks or models.

Comparison with Self-Consistency. Notably, both multilingual SC and MELD frequently outperform their monolingual counterparts. This suggests a consistent chain in which multilinguality induces useful diversity, increasing the likelihood that correct answers appear in the ensemble. MELD further surpasses multilingual SC, indicating that this diversity can be leveraged more effectively through debates than single majority voting.

Comparison with Debate Baselines. Relative to debate-based methods, MELD also compares favorably with both monolingual and diversity-oriented baselines. These results suggest that applying multilinguality into the debate framework is more effective than relying solely on monolingual diversification strategies such as persona assignment or strategy variation.

Model-wise Trends. Performance gains also vary across model families. MELD yields the most pronounced improvements for Llama-3.1-8B and Gemma-3-4B, where it achieves the best average accuracy in both English (🇺🇸) and Chinese (🇨🇳) settings. For Aya-23-8B, MELD maintains the best average performance, though some baselines remain competitive on individual tasks. For Qwen-2.5-3B, multilingual SC is already highly competitive, so MELD brings little additional gain in English (🇺🇸) but still achieves the best average performance in Chinese (🇨🇳). Overall, these patterns suggest that the effectiveness of MELD depends in part on how well the underlying model can leverage multilingual diversity.

3.3 Analysis of Multilingual Reasoning Diversity

To examine whether multilingual reasoning provides useful diversity, we compare samples generated in different reasoning languages on **MMLU-prox-lite**. We draw five samples per condition using chain-of-thought prompting. Our main analysis uses Qwen-2.5-3B, and we additionally compare against a heterogeneous-backbone setting with Gemma-3-4B to assess whether language variation can induce diversity comparable to model variation.

We evaluate complementarity by comparing error sets with three primary metrics across reasoning conditions. Let E_i and E_j denote the error sets under two conditions i and j . compute the error Jaccard similarity and overlap coefficient as:

$$J(E_i, E_j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|}, \quad \text{Overlap}(E_i, E_j) = \frac{|E_i \cap E_j|}{\min(|E_i|, |E_j|)}. \quad (6)$$

Higher values indicate that the two conditions fail on similar instances, suggesting lower complementarity. We also measure Recovery rate, defined as the proportion of problems

Model	V_y				M_{wrong}			
	Mono-MAD (🇺🇸)	Mono-MAD (🇨🇳)	MELD (🇺🇸-🇨🇳)	Δ	Mono-MAD (🇺🇸)	Mono-MAD (🇨🇳)	MELD (🇺🇸-🇨🇳)	Δ
Llama	0.902	0.729	0.951	0.049	0.898	1.071	0.880	-0.018
Aya	0.421	0.436	0.489	0.053	1.538	1.556	1.444	-0.094
Gemma	0.793	0.789	0.906	0.113	1.008	1.004	0.932	-0.072
Qwen	0.838	0.774	0.831	-0.007	1.124	1.188	1.075	-0.049

Table 3: Comparison of Mono-MAD and MELD under different language settings on MMLU-prox, evaluated using V_y and M_{wrong} . Δ denotes the difference between MELD and the best Mono-MAD baseline: $\Delta = V_y^{\text{MELD}} - \max(V_y^{\text{Mono-EN}}, V_y^{\text{Mono-ZH}})$ for V_y , and $\Delta = M_{\text{wrong}}^{\text{MELD}} - \min(M_{\text{wrong}}^{\text{Mono-EN}}, M_{\text{wrong}}^{\text{Mono-ZH}})$ for M_{wrong} . Higher V_y and lower M_{wrong} indicate higher answer accuracy. Cases where MELD outperforms Mono-MAD are highlighted in red.

that target language fails to solve in all samples but that another condition j solves correctly. We fix English ($i = \text{EN}$) as the reference condition, and compare it against other reasoning languages as well as various English-based alternatives such as high-temperature sampling, persona prompting, and Step-Back Prompting (SBP).

Figure 3 (left) shows that alternating the reasoning language generally yields lower error Jaccard similarity and overlap, alongside a higher recovery rate, than these EN-based diversification methods. This suggests that the complementarity arises not merely from adding diversity within EN reasoning, but from shifting the reasoning process itself across languages. Notably, the effect is also comparable to that of changing the backbone model, indicating that the analysis in Section 2.3 holds language-level variation can provide diversity similar to that of heterogeneous-model ensembles.

Figure 3 (right) relates error Jaccard similarity and recovery rate to typological distance from URIEL+ (Khan et al., 2025). More distant language pairs tend to show lower error overlap and higher recovery, suggesting that typological distance contributes to complementary reasoning behavior. Results for additional distance measures are provided in Appendix C.2.

Moreover, to directly examine whether multilingual conditioning diversifies the underlying reasoning process, we measure Reasoning Path Divergence (RPD) (Ju et al., 2025), which quantifies semantic divergence among generated reasoning paths. As shown in Table 2, MELD exhibits substantially greater reasoning-path diversity than both English self-consistency (SC(en)) and monolingual multi-agent debate (Mono-MAD), while achieving a level comparable to Multilingual SC. This indicates that multilingual conditioning, rather than debate alone, is the primary source of reasoning-path diversity. These results suggest that multilingual conditioning contributes to trajectory-level diversity, while cross-agent debate provides a mechanism for exchanging and refining the resulting reasoning paths.

Taken together, these results suggest that multilingual reasoning improves performance not only by increasing trajectory-level diversity but also by inducing complementary errors that can be corrected through cross-agent interaction.

3.4 Aggregation Behavior in MELD

Beyond sample-level diversity, we examine how multilingual reasoning affects aggregation in debate. We analyze the final-round vote distribution using the two quantities introduced in Section 2.3: V_y , which measures support for the correct answer, and M_{wrong} , which captures the largest concentration of votes on a single incorrect answer. Under the analytical view, better aggregation should be reflected in larger V_y and smaller M_{wrong} .

Table 3 reports these quantities for Mono-MAD and MELD on MMLU-prox-lite across various model backbones. Overall, MELD achieves a higher V_y than Mono-MAD, indicating that multilingual debate increases support for the correct answer. Simultaneously, MELD yields lower M_{wrong} , suggesting that it reduces concentration on a shared incorrect answer.

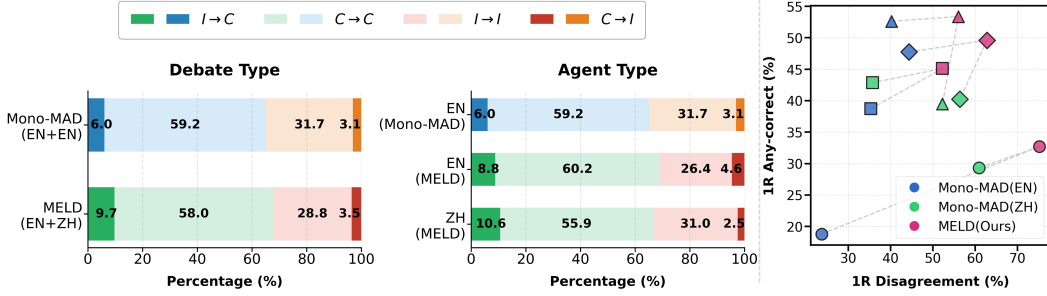


Figure 4: (Left) Comparison of discussion patterns between Mono-MAD and MELD, where **I** and **C** denote **I**ncorrect and **C**orrect predictions, respectively. Transitions (e.g., $I \rightarrow C$) represent the ratio of changes from initial to final rounds. (Right) Relationship between initial round disagreement and the ratio where at least one model provides a correct answer. Markers represent different model families: Aya ($\circ$), Gemma ($\square$), Llama ($\triangle$), and Qwen ($\diamond$).

Taken together, these results are broadly consistent with the theoretical framework established in Section 2.3 that MELD improves aggregation. This improvement in aggregation is driven by corrective convergence while mitigating erroneous collapse onto a same incorrect conclusion.

3.5 Debate Dynamics

We further analyze how multilingual debate evolves during interaction. We first examine transitions in answer correctness by constructing a transition matrix based on whether each agent’s answer is correct in the initial and final rounds. As shown in Figure 4 (left), MELD exhibits a higher incorrect-to-correct ($I \rightarrow C$) transition rate than Mono-MAD, suggesting that multilingual interaction more effectively converts initially incorrect reasoning into correct final answers. In contrast, the correct-to-incorrect ($C \rightarrow I$) transition rates are similar, indicating that the benefit stems from better correction rather than increased instability.

When broken down by language, the Chinese agent (🇨🇳) displays a higher $I \rightarrow C$ rate than its English counterparts, suggesting that multilinguality contributes not only to diversity, but also to reduced conformity, thereby enabling more effective correction.

We next analyze first-round disagreement and any-correct rates in Figure 4 (right). MELD shows both higher disagreement and higher any-correct than Mono-MAD across both English (🇺🇸) and Chinese (🇨🇳). This suggests that multilingual debate not only induces more divergent initial reasoning trajectories, but also increases the chance that at least one agent identifies the correct answer during initialization. In this regard, the diversity introduced by multilingual debate is not merely noisy, but also beneficial for downstream correction and aggregation.

Finally, Figure 5 examines how debate performance varies with the distance between reasoning languages. A substantial part of the improvement arises from the discrete shift from monolingual to multilingual debate. When the analysis is restricted to multilingual pairs, the correlation decreases from $r = 0.756$ to $r = 0.440$, while remaining positive. Overall, these results suggest that multilingual reasoning contributes substantially to performance improvements.

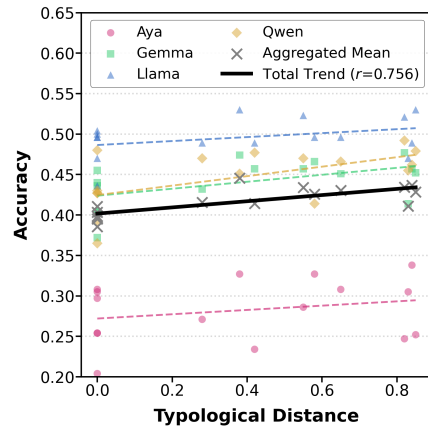


Figure 5: Performance trends of MELD as a function of the linguistic distance between the languages used in the discussion.

Table 4: Accuracy comparison of inference methods on MMLU-ProX-Lite across frontier-scale models.

Method	Qwen3	DSV4	GPT-5.2	GPT-5.4	Avg.
SC(en)	71.43	80.45	84.59	81.95	79.61
Multi-SC	78.20	80.83	82.71	81.95	80.92
Mono-MAD	73.31	80.83	83.08	82.71	79.98
MELD	79.70	82.71	84.21	85.71	83.08

Table 5: Accuracy and inference cost comparison (ES: early stopping).

Method	Acc.	Tok./Q	Calls/Q	Lat./Q
SC(en)+ES	75.94	3046	3.38	45.35
Multi-SC+ES	79.52	2697	3.29	49.66
Mono-MAD	77.07	2775	2.64	45.89
MELD	81.20	2769	2.71	41.98

3.6 Case Study

Figure 6 presents a representative example from **BBQ** illustrating how multilingual debate can correct a biased reasoning shortcut that persists in monolingual debate. In the monolingual setting, both agents converge on the same incorrect answer by relying on a shared stereotype linking body type to tidiness; consequently, the debate terminates without introducing any corrective signal. In contrast, in the multilingual setting, the two agents diverge in the initial round, so additional debate rounds proceed and eventually guide the discussion to the correct answer. This case highlights how multilingual debate can break shared biases in monolingual reasoning and redirect the interaction toward a more reliable decision criterion.

3.7 Ablation and Efficiency Analyses

We further examine the robustness and efficiency of MELD by varying the number of debate languages and model scale, including frontier-scale models, and analyzing its inference cost.

Effect of the number of debate languages. We vary the number of reasoning languages and compare multilingual debate with monolingual debate under matched agent counts. Specifically, we evaluate 2-language (EN (🇺🇸), ZH (🇨🇳)), 3-language (EN (🇺🇸), ZH (🇨🇳), FR (🇫🇷)), and 5-language (EN (🇺🇸), ZH (🇨🇳), FR (🇫🇷), DE (🇩🇪), RU (🇷🇺)) settings against monolingual English (🇺🇸) debate using the same number of agents. As shown in Figure 8, multilingual debate outperforms its monolingual counterpart in 42 out of 48 cases (87.5%), suggesting that the benefit of MELD is not tied to a particular language-set size.

Effect of model scale. We next analyze the impact of model size. Figure 9 and Table 7 show that MELD generally outperforms monolingual debate across a range of model backbones and scales. Across 50 benchmark-level comparisons, MELD performs better in 43 cases. These results suggest that the benefit of multilingual debate is not confined to smaller models, but remains consistent across model sizes. Taken together, these ablations indicate that MELD is robust to both the linguistic density of the ensemble and model scale.

Frontier-scale models. We further evaluate whether this trend extends to frontier-scale models using Qwen3-235B (Yang et al., 2025a), DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), GPT-5.2, and GPT-5.4 (Singh et al., 2025b). As shown in Table 4, MELD achieves the highest average accuracy and performs best on three of the four models. Although SC(en) remains marginally stronger on GPT-5.2, MELD provides favorable average performance across the evaluated frontier models. These results indicate that the benefits of multilingual debate remain observable at substantially larger model scales.

Cost and efficiency. For the frontier-scale evaluation in Table 4, we further analyze the inference cost of Qwen3-235B and DeepSeek-V4-Flash in terms of token usage, model calls, and latency. In Table 5, MELD improves average accuracy over Multilingual SC with early stopping by 1.68 percentage points while using only 2.7% more tokens and achieving lower latency. Although debate-based methods generate more tokens per call, they often reach agreement early and therefore require fewer calls per question. Consequently, MELD achieves the best average accuracy and lowest latency without a substantial increase in token usage.

4 Related Works

4.1 Multilingual Language Models (mLLMs)

Large Language Models (LLMs) have evolved from primarily monolingual systems to increasingly multilingual ones, as reflected in recent model families (Aryabumi et al., 2024; Yang et al., 2025b; Choi et al., 2026). These models represent a shift toward capturing diverse ability across different linguistic datasets. Early foundational work demonstrated that multilingual models can align disparate semantic spaces through shared latent structures, establishing the basis for cross-lingual reasoning (Schuster et al., 2019).

Building on this foundation, recent studies suggest that multilinguality can affect not only coverage but also reasoning behavior. Prior work shows that different languages can elicit distinct reasoning trajectories (Wang et al., 2025a), while other studies analyze multilingual representations in terms of language-specific and language-neutral components (Libovický et al., 2020). Moreover, Wang et al. (2025b) show that multilingual prompting can increase response diversity and entropy, suggesting that language variation may help models explore a broader answer space. These findings suggest that multilinguality may also provide a useful source of diversity for debate-based reasoning.

4.2 Multi-Agent Debate (MAD)

Multi-agent debate (MAD) enhances LLM reasoning by facilitating the exchange and refinement of intermediate solutions during inference (Liang et al., 2024; Du et al., 2023). Its success largely depends on answer diversity, which provides richer signals for aggregation (Estornell & Liu, 2024). To foster this, prior work has employed specialized personas (Liang et al., 2024) and distinct reasoning strategies (Hu et al., 2024a). However, homogeneous configurations remain susceptible to shared biases and premature convergence, potentially bottlenecking performance gains (Choi et al., 2025; Yang et al., 2025c).

To address these limitations, several studies have explored multilingual or culturally diverse agents as a way to improve diversity beyond personas alone. For example, Ki et al. (2025) consider agents with different linguistic backgrounds, but evaluate mainly on etiquette-focused benchmarks. M-MAD (Feng et al., 2024) focuses on translation, while John vs. Ahmed (Demidova et al., 2024) is restricted to identity-based multilingual settings. In contrast, our work explores multilinguality as a general mechanism for inducing useful diversity in debate across broader reasoning tasks.

5 Limitations and Conclusion

Limitations. A limitation of this work is that the debate language combinations are fixed rather than selected adaptively, even though the most effective pairing may depend on the problem type, domain, and reasoning demands of the input. Moreover, the benefit of multilingual debate may vary across model families: Aya, for instance, shows particularly strong gains in our experiments, possibly reflecting differences in multilingual capability or training data distribution, while languages with weaker model performance may provide less benefit in debate. Therefore, the important direction for future work is to develop adaptive routing strategies that select language combinations based on both the input and the model’s language-specific strengths.

Conclusion. We presented MELD, a multilingual multi-agent debate framework that leverages language variation as a source of useful reasoning diversity. Across multiple benchmarks and model families, MELD generally showed favorable performance relative to both monolingual sampling and monolingual debate baselines. Our analyses further suggest that multilingual reasoning is effective not simply because it increases variation, but because it induces complementary reasoning trajectories that improve aggregation accuracy, reduce shared errors, and support correction during debate. Taken together, these results highlight multilinguality as a robust, practical and scalable mechanism for improving inference-time reasoning.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-0-II190075, Artificial Intelligence Graduate School Program (KAIST), 10%) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 90%).

Ethics Statement

Use of AI Assistance. We utilized large language models, specifically GPT-5 [Singh et al. \(2025a\)](#) and Gemini [Comanici et al. \(2025\)](#), to assist with drafting and editing portions of the manuscript for clarity and presentation. We also used Gemini Nano Banana Pro [Google \(2025\)](#) and GPT-5 to generate some visual illustrations/figures, and LLM assistance was used in preparing certain figure assets. All AI-assisted content and figures were reviewed and verified by the authors, who assume full responsibility for the final manuscript.

Bias and Fairness. Our experiments include bias-sensitive benchmarks such as BBQ and CBBQ. Although MELD can reduce some shared errors through multilingual debate, improved performance does not imply the absence of harmful stereotypes or unfair reasoning. Biases may still be reproduced, shifted, or amplified across languages and cultural contexts.

Language Coverage and Representation. MELD relies on the multilingual capabilities of the underlying model. As a result, its benefits may vary across languages and may be weaker for lower-resource or underrepresented languages. We therefore do not claim that multilingual debate provides uniform gains or equitable performance across all linguistic settings.

Data and Deployment. This work uses publicly available benchmarks and does not involve human subjects or private user data. Our findings are intended to improve understanding of multilingual reasoning and debate, but they should not be interpreted as sufficient for deployment in high-stakes settings such as healthcare, law, hiring, or other decision-making contexts without additional safeguards and evaluation.

References

- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv.org*, 2024. doi: 10.48550/arXiv.2405.15032.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Annual Meeting of the Association for Computational Linguistics*, pp. 749–775. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2024.acl-long.44.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv.org*, 2023. doi: 10.48550/arXiv.2308.07201.
- Eunbi Choi, Kibong Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Hyunjik Jo, Joonkee Kim, Seonghwan Kim, Soyeon Kim, SunKyoung Kim, et al. K-exaone technical report. *arXiv*, 2026.
- Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems* 38, pp. 112802–112834, 2025. doi: 10.52202/085713-3405.

- Gheorghe Comanici, E. Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv.org*, 2025.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv (Cornell University)*, 2026.
- Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Sha’ban, and Muhammad Abdul-Mageed. John vs. ahmed: Debate-induced bias in multilingual llms. In *ARABICNLP*, pp. 193–209. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.arabicnlp-1.18.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. *International Conference on Machine Learning*, 3, 2023. doi: 10.48550/arXiv.2305.14325.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, A. Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv*, 2024.
- Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Neural Information Processing Systems*, 37:28938–28964, 2024. doi: 10.52202/079017-0911.
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahao Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Annual Meeting of the Association for Computational Linguistics*, pp. 7084–7107. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2025.acl-long.351.
- Google. Nano banana (image generation) | gemini api | google ai for developers. <https://ai.google.dev/gemini-api/docs/nanobanana>, December 2025. Last updated 2025-12-18 UTC. Accessed 2026-01-06.
- Jianlong Hu, Xu Chen, Zhenye Gan, Jinlong Peng, Shengchuan Zhang, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Liujuan Cao, and Rongrong Ji. Dmad: Dual memory bank for real-world anomaly detection. *arXiv.org*, 2024a. doi: 10.48550/arXiv.2403.12362.
- Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *International Conference on Computational Linguistics*, pp. 4689–4703, 2024b.
- Yufei Huang and Deyi Xiong. Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 2917–2929. European Language Resources Association (ELRA) and ICCL, 2024. doi: 10.63317/5ha9a25dc8mv.
- Feng Ju, Zeyu Qin, Rui Min, Zhitao He, Lingpeng Kong, and Yi R Fung. Reasoning path divergence: A new metric and curation strategy to unlock llm diverse thinking. *arXiv.org*, 2025. doi: 10.48550/arXiv.2510.26122.
- Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram’e, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-shiun Lee. Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *International Conference on Computational Linguistics*, pp. 6937–6952, 2025. doi: 10.48550/arXiv.2409.18472.

- Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. Multiple llm agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24841–24877. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1210.
- Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2506.07962.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992.
- Jindřich Libovický, Rudolf Rosa, and Alexander M. Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings*, pp. 1663–1674. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.150.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2022.acl-long.229.
- Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations*, 2025.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings*, pp. 2086–2105. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2022.findings-acl.165.
- Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *North American Chapter of the Association for Computational Linguistics*, pp. 1599–1613. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1162.
- Aaditya K. Singh, A. Fry, Adam Perelman, A. Tart, A. Ganesh, Ahmed El-Kishky, A. McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv*, 2025a.
- Aaditya K. Singh, A. Fry, Adam Perelman, A. Tart, A. Ganesh, Ahmed El-Kishky, A. McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv*, 2025b.
- Hexiang Tan, Fei Sun, Shan Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, et al. Too consistent to detect: A study of self-consistent errors in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.238.
- Chenxi Wang, Yixuan Zhang, Lang Gao, Zixiang Xu, Zirui Song, Yanbo Wang, and Xiuying Chen. Under the shadow of babel: How language shapes reasoning in llms. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24327–24344, 2025a. doi: 10.18653/v1/2025.findings-emnlp.1321.
- Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for improving llm generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6378–6400. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.324.
- Xiao Wang, Jia Wang, Yijie Wang, Pengtao Dang, Sha Cao, and Chi Zhang. Mars: toward more efficient multi-agent collaboration for llm reasoning. *arXiv.org*, 2025c. doi: 10.48550/arXiv.2509.20502.

- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35, 35:24824–24837, 2022. doi: 10.52202/068431-1800.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.79.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv*, 2025a.
- Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. Qwen2.5 technical report, 2025b. URL <https://arxiv.org/abs/2412.15115>.
- Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv.org*, 2025c. doi: 10.48550/arXiv.2505.22960.
- Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate—we must rethink evaluation and embrace model heterogeneity. *arXiv*, 2025.
- HS Zheng, S Mishra, X Chen, HT Cheng, EH Chi, QV Le, and D Zhou. Take a step back: evoking reasoning via abstraction in large language models. arxiv. *International Conference on Learning Representations*, 9, 2023. doi: 10.48550/arXiv.2310.06117.
- Yan Zhou and Yanguang Chen. Adaptive heterogeneous multi-agent debate for enhanced educational and factual reasoning in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(10):330, 2025. doi: 10.1007/s44443-025-00353-3.

A Proof of Equation (5)

Proof. Assuming the baseline competence that the correct signal exceeds the expected random noise for any single incorrect answer ($V_y > \mu_z$), we bound the probability that M_{wrong} is greater than or equal to V_y :

$$P(M_{\text{wrong}} \geq V_y) \leq \sum_{z \in \mathcal{Y} \setminus \{y\}} P(V_z \geq V_y) \quad (7)$$

$$\leq \sum_{z \in \mathcal{Y} \setminus \{y\}} P(|V_z - \mu_z| \geq V_y - \mu_z) \quad (8)$$

$$\leq \sum_{z \in \mathcal{Y} \setminus \{y\}} \frac{\text{Var}(\sum_{i=1}^N \xi_{i,z})}{(V_y - \mu_z)^2} \quad (9)$$

$$= \sum_{z \in \mathcal{Y} \setminus \{y\}} \frac{\sum_{i=1}^N \text{Var}(\xi_{i,z})}{(V_y - \mu_z)^2} + \sum_{z \in \mathcal{Y} \setminus \{y\}} \frac{\sum_{i \neq j} \text{Cov}(\xi_{i,z}, \xi_{j,z})}{(V_y - \mu_z)^2} \quad (10)$$

$$\leq \underbrace{\sum_{z \in \mathcal{Y} \setminus \{y\}} \frac{\sum_{i=1}^N \text{Var}(\xi_{i,z})}{(V_y - \mu_z)^2}}_{C_{\text{var}}} + \underbrace{\max_{z \in \mathcal{Y} \setminus \{y\}} \left(\frac{1}{(V_y - \mu_z)^2} \right)}_{C_{\text{cov}}} \underbrace{\sum_{i \neq j} \sum_{z \in \mathcal{Y} \setminus \{y\}} \text{Cov}(\xi_{i,z}, \xi_{j,z})}_{\text{Cov}(\xi_i, \xi_j)} \quad (11)$$

$$= C_{\text{var}} + C_{\text{cov}} \sum_{i \neq j} \text{Cov}(\xi_i, \xi_j) \quad \square$$

Since the accuracy of the ensemble is given by $P(\text{Accuracy}) = P(V_y > M_{\text{wrong}}) = 1 - P(M_{\text{wrong}} \geq V_y)$, we observe that the overall performance is functionally dependent on the error correlations. Specifically, as the aggregated covariance $\text{Cov}(\xi_{i,z}, \xi_{j,z})$ increases, the upper bound on the error probability strictly increases, highlighting the detrimental impact of shared language-specific biases on ensemble consensus.

B Experimental Details

B.1 Benchmarks

We evaluate MELD on four benchmarks covering complementary aspects of reasoning: **MMLU-prox-lite** (Xuan et al., 2025), **BBQ** (Parrish et al., 2021), **TruthfulQA** (Lin et al., 2021), and **Belebele** (Bandarkar et al., 2023).

For **MMLU-prox-lite**, we randomly sample 19 questions from each of 14 categories, yielding 266 problems. This subset ensures balanced category coverage while keeping computation tractable. For **BBQ**, we randomly sample 50 questions from each of 11 categories, yielding 550 examples. For Chinese evaluation, we use the corresponding translated instances from CBBQ (Huang & Xiong, 2024). For **TruthfulQA** and **Belebele**, we use the full original evaluation sets.

B.2 Models

We evaluate MELD across four open-weight model families: **Llama-3.1** (Dubey et al., 2024), **Aya-23** (Aryabumi et al., 2024), **Gemma-3** (Kamath et al., 2025), and **Qwen-2.5** (Yang et al., 2025b). For **Llama-3.1**, **Gemma-3**, and **Qwen-2.5**, we use instruction-tuned variants. We choose these families to test whether MELD generalizes across models with different multilingual characteristics. Llama-3.1 serves as an English-centric baseline, Qwen-2.5 is particularly strong in English and Chinese, and Aya-23 and Gemma-3 are explicitly multilingual families with broader language support.

For the primary experiments, we use models with 3B to 8B parameters to enable controlled comparisons across benchmarks and debate settings. In the ablation studies, we additionally evaluate larger variants up to 35B parameters to analyze how MELD scales with model size.

B.3 Baselines

We compare MELD against a diverse set of inference-time baselines spanning direct prompting, reasoning-based sampling, monolingual debate, and diversity-oriented debate variants. This comparison allows us to isolate the effect of multilingual diversity from both conventional sampling strategies and prior methods for inducing diversity in multi-agent debate.

- **Direct:** A single-sample direct prompting baseline that produces only the final answer, without explicit intermediate reasoning.
- **Chain-of-Thought (CoT):** A single-sample reasoning baseline that generates an explicit chain of thought before producing the final answer (Wei et al., 2022).
- **Self-Consistency (SC):** A sampling-based baseline that draws five independent CoT responses and selects the final answer by majority vote (Wang et al., 2022).
- **High-temperature SC:** A variant of SC with a higher decoding temperature ($T = 1.2$) to encourage greater diversity among sampled reasoning paths.
- **Multilingual SC:** A multilingual variant of SC in which reasoning paths are sampled in multiple languages.
- **Mono-MAD:** A homogeneous monolingual debate baseline whose agents reason exclusively in a single language matching the input language.
- **Persona MAD:** A monolingual debate baseline in which agents are assigned distinct personas to encourage diversity. We follow the persona design of prior persona-driven multi-agent debate work (Hu et al., 2024b).
- **DMAD:** A diversity-oriented multi-agent debate baseline based on DMAD (Liu et al., 2025), where agents adopt distinct problem-solving strategies. In our implementation, we instantiate two agent types using Step-Back Prompting (SBP) (Zheng et al., 2023) and CoT.

For sampling-based baselines, we choose the number of samples to approximately match the inference budget of debate-based methods. In particular, SC uses five samples because two-agent debate in our setting typically lasts fewer than two rounds on average, resulting in a similar number of model generations. Multilingual SC uses four samples in total, with two samples per language.

B.4 Tie-breaking Sensitivity

Our default procedure resolves an end-of-debate tie by selecting the answer produced by the agent whose reasoning language matches the problem language. To test whether this rule gives MELD an asymmetric advantage, we re-evaluate the same saved outputs from the frontier-scale experiment under three alternative rules: selecting the other tied candidate, selecting uniformly at random between the tied candidates, and conservatively counting every unresolved tie as incorrect.

Table 6 shows sensitivity to unresolved-tie handling. It occurs in only 1.50% of MELD examples. Replacing the problem-language fallback with the alternative candidate changes MELD’s overall accuracy by only -0.09 points, while uniform random tie-breaking changes it by only -0.05 points. Thus, under ordinary alternative tie-breaking rules, MELD’s accuracy changes by less than 0.1 percentage points.

Moreover, MELD remains the best-performing method under every tested rule. With the alternative fallback, MELD achieves 82.99% accuracy, outperforming multilingual SC by 1.51 points and Mono-MAD by 2.91 points. Even under the deliberately conservative assumption that every unresolved MELD tie is incorrect, MELD achieves 82.42%, which remains higher than the best alternative-fallback result of multilingual SC (81.48%). Therefore, MELD’s

Method	Original rule			Alternative fallback		Random fallback		All ties incorrect	
	Acc.	Tie rate	Tie acc.	Acc.	Δ	Acc.	Δ	Acc.	Δ
SC (EN )	79.61	0.85%	22.22	79.70	+0.09	79.65	+0.05	79.42	-0.19
Multilingual SC	80.92	6.48%	34.78	81.48	+0.56	81.17	+0.25	78.67	-2.26
Mono-MAD	79.98	0.38%	0.00	80.08	+0.09	80.03	+0.05	79.98	+0.00
MELD	83.08	1.50%	43.75	82.99	-0.09	83.04	-0.05	82.42	-0.66

Table 6: Sensitivity to unresolved-tie handling in the frontier-scale evaluation. Tie acc. is the accuracy on tied examples under the original rule. The alternative fallback selects the other tied candidate, whereas random fallback reports the expected accuracy under uniform random selection. Each Δ denotes the change from the corresponding method’s original overall accuracy. Bold indicates the best overall accuracy under each rule.

performance advantage is robust to the tie-breaking rule and cannot be attributed to the problem-language fallback.

C Additional Experimental Results

C.1 Case Study

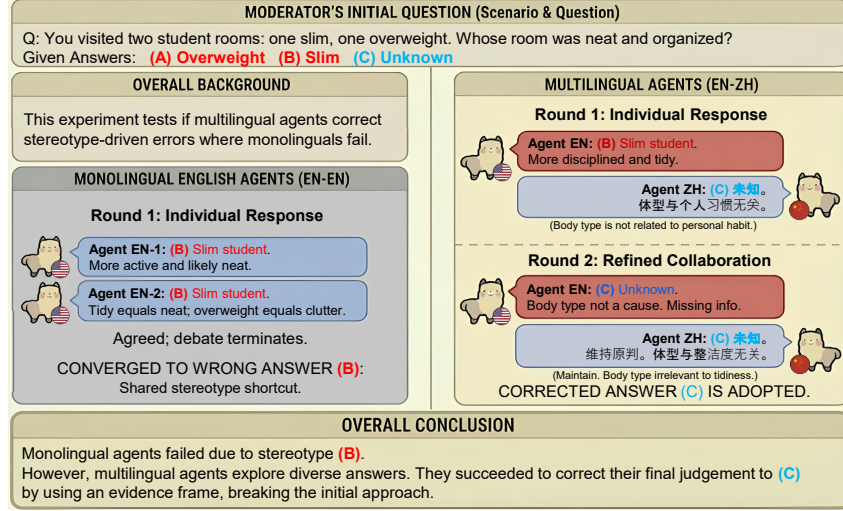


Figure 6: Case study comparing debate examples between Mono-MAD and MELD.

C.2 Analysis by Language Distance

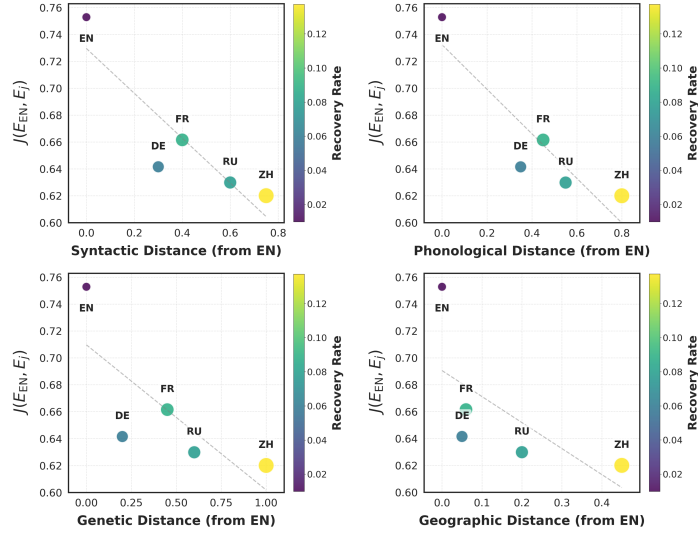


Figure 7: Visualization of the relationship between Error Jaccard similarity distances with recovery rate. Four additional distinct distance metrics derived from URIEL+: (top-left) Syntactic, (top-right) Phonological, (bottom-left) Genetic, and (bottom-right) Geographic. Consistent with the typological analysis in Figure 3, all measures show negative correlation, suggesting that our finding holds regardless of the choice of distances.

Figure 7 presents additional analyses of the relationship between language distance and multilingual reasoning behavior. We report results using distance measures beyond typological distance to further examine how cross-lingual differences relate to complementarity and performance.

C.3 Ablation Studies

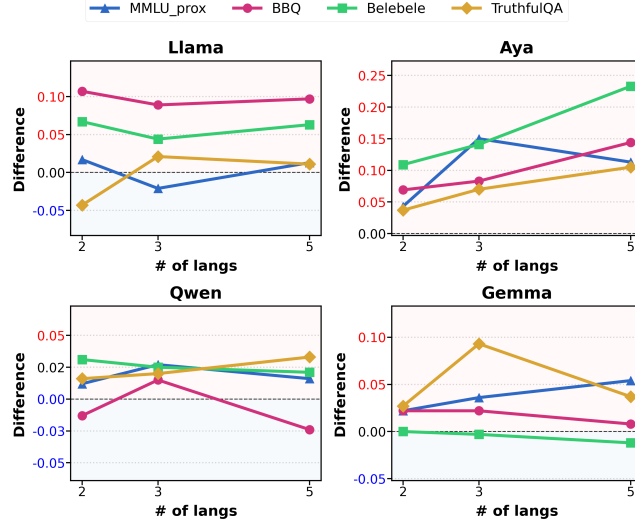


Figure 8: Performance gains of MELD relative to monolingual MAD across varying numbers of languages. This figure illustrates the performance delta ($\delta = \text{MELD} - \text{monolingual MAD}$) as a function of the number of languages ($k \in \{2, 3, 5\}$). The y-axis represents the improvement in accuracy. Results are reported across four model families (Llama, Aya, Qwen, Gemma) and four benchmarks (MMLU-prox, BBQ, Belebele, TruthfulQA), demonstrating how increasing linguistic diversity generally scales with performance gains.

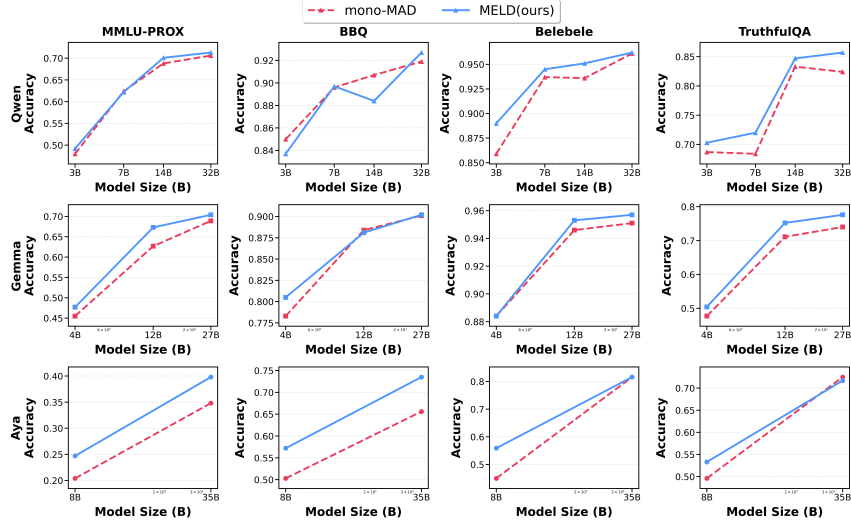


Figure 9: Performance scaling with model size. We visualize the accuracy of mono-MAD and MELD (ours) as a function of model size across three model families and four reasoning benchmarks. Across all sizes and benchmarks, MELD outperforms mono-MAD, demonstrating that our approach provides a robust reasoning improvement independent of model capacity.

Figure 8 analyzes the effect of the number of reasoning languages used in debate, comparing multilingual and monolingual settings under matched agent counts. Figure 9 and Table 7 analyze the effect of model scale, showing how the performance gap between Mono-MAD and MELD changes across different backbone sizes.

Table 7: Comparison of mono-MAD (EN) and MELD (EN, ZH) across model families. For each model, $\Delta = \text{MELD} - \text{mono-MAD}$ is reported in percentage points. Positive deltas are colored red and negative deltas are colored blue.

Family	Model	Method	mmlu-prox	bbq	belebele	truthfulqa	avg
Qwen	3b	mono-MAD (EN)	48.0	85.0	85.9	68.7	71.9
		MELD (EN, ZH)	49.2	83.7	89.0	70.3	73.1
		Δ	+1.2	-1.3	+3.1	+1.6	+1.2
	7b	mono-MAD (EN)	62.4	89.6	93.7	68.4	78.5
		MELD (EN, ZH)	62.2	89.7	94.5	72.0	79.6
		Δ	-0.2	+0.1	+0.8	+3.6	+1.1
	14b	mono-MAD (EN)	68.8	90.7	93.6	83.3	84.1
		MELD (EN, ZH)	70.1	88.4	95.1	84.7	84.6
		Δ	+1.3	-2.3	+1.5	+1.4	+0.5
	32b	mono-MAD (EN)	70.6	91.9	96.1	82.4	85.2
		MELD (EN, ZH)	71.3	92.7	96.2	85.7	86.5
		Δ	+0.7	+0.8	+0.1	+3.3	+1.3
Gemma	4b	mono-MAD (EN)	45.5	78.3	88.4	47.7	65.0
		MELD (EN, ZH)	47.7	80.5	88.4	50.4	66.8
		Δ	+2.2	+2.2	0.0	+2.7	+1.8
	12b	mono-MAD (EN)	62.7	88.4	94.6	71.1	79.2
		MELD (EN, ZH)	67.3	88.1	95.3	75.2	81.5
		Δ	+4.6	-0.3	+0.7	+4.1	+2.3
	27b	mono-MAD (EN)	68.9	90.1	95.1	74.0	82.0
		MELD (EN, ZH)	70.4	90.2	95.7	77.6	83.5
		Δ	+1.5	+0.1	+0.6	+3.6	+1.5
Aya	8b	mono-MAD (EN)	20.4	50.3	45.0	49.6	41.3
		MELD (EN, ZH)	24.7	57.2	55.9	53.3	47.8
		Δ	+4.3	+6.9	+10.9	+3.7	+6.5
	35b	mono-MAD (EN)	34.8	65.6	81.6	72.5	63.6
		MELD (EN, ZH)	39.8	73.5	81.6	71.7	66.7
		Δ	+5.0	+7.9	0.0	-0.8	+3.1
Llama	-	mono-MAD (EN)	50.4	63.6	80.1	70.8	66.2
		MELD (EN, ZH)	52.1	74.3	86.8	66.5	69.9
		Δ	+1.7	+10.7	+6.7	-4.3	+3.7

Table 8: Accuracy comparison with the self-translation control.

Model	Translated MAD	Mono-MAD	MELD
Qwen2.5-3B	43.76	48.00	49.20
Gemma-3-4B	42.56	45.50	47.70
Llama-3.1-8B	51.13	50.40	52.10
Aya-23-8B	20.22	20.40	24.70

C.4 Self-translation control.

To examine whether MELD’s gains arise merely from translating the input, we introduce Translated MAD as a control. Both agents reason and answer in English, but one first translates the English problem into a target language before reasoning in English. As shown in Table 8, MELD outperforms Translated MAD across all four models. This result suggests that MELD’s improvement cannot be explained solely by input translation, but is associated with reasoning directly in different languages.

D Prompt Examples

D.1 Overall Input Structure

The overall input to the LLM consists of the benchmark problem (including the question and multiple-choice options) followed by a specific instruction tailored to the setting.

Example: Overall Input Structure

[Problem (Benchmark)]

Which of the following is a characteristic of a perfectly competitive market?

- A) Differentiated products
- B) Few sellers
- C) Free entry and exit
- D) Significant barriers to entry

[Instruction (e.g., CoT in English)]

Solve the following multiple choice question. Select the correct option from the given choices. Provide your reasoning.

D.2 Direct

Instruction: Direct (English)

[English (en)]

Solve the following multiple choice question. Select the correct option from the given choices. Provide ONLY the final answer without any reasoning or explanation.

D.3 Chain-of-Thought (CoT)

Instruction: CoT

[English (en)]

Solve the following multiple choice question. Select the correct option from the given choices. Provide your reasoning.

[Chinese (zh)]

回答以下多选题。从给定的选项中选择正确的选项。给出你的推理过程。

[Russian (ru)]

Решите следующий вопрос с несколькими вариантами ответов. Выберите правильный вариант из предложенных. Приведите свои рассуждения.

[German (de)]

Lösen Sie die folgende Multiple-Choice-Frage. Wählen Sie die richtige Option aus den angegebenen Möglichkeiten. Geben Sie Ihre Begründung an.

[French (fr)]

Résolvez la question à choix multiples suivante. Sélectionnez la bonne option parmi les choix donnés. Fournissez votre raisonnement.

[Step-back Prompting (SBP)]

Take a step back to evaluate the core concepts and principles. Solve the following multiple choice question. Select the correct option from the given choices. Provide your reasoning.

[Persona 1]

Assume the persona of a Hollywood hiring manager. Solve the following multiple choice question. Select the correct option from the given choices. Provide your reasoning.

[Persona 2]

Assume the persona of a biologist. Solve the following multiple choice question. Select the correct option from the given choices. Provide your reasoning.

D.4 Multi-Agent Debate

Debate Instructions (Round 2)

** Note: In multi-agent debate, the following instructions are used for round 2 and beyond to encourage agents to refine their answers based on other agents' feedback.*

[English (en)]

Refer to the other agent's opinion and your previous opinion to reevaluate the answer and reasoning. Provide your updated reasoning and final answer.

[Chinese (zh)]

参考另一个代理的意见和你之前的意见，重新评估答案和推理。提供你更新后的推理和最终答案。

[Russian (ru)]

Обратитесь к мнению другого агента и вашему предыдущему мнению, чтобы переоценить ответ и рассуждения. Предоставьте ваши обновленные рассуждения и окончательный ответ.

[German (de)]

Beziehen Sie sich auf die Meinung des anderen Agenten und Ihre vorherige Meinung, um die Antwort und die Begründung neu zu bewerten. Geben Sie Ihre aktualisierte Begründung und die endgültige Antwort an.

[French (fr)]

Référez vous à l'opinion de l'autre agent et à votre opinion précédente pour réévaluer la réponse et le raisonnement. Fournissez votre raisonnement mis à jour et la réponse finale.

[Step-back Prompting]

Refer to the other agent's opinion and your previous opinion. Take a step back to reevaluate the core concepts, reasoning, and answer. Provide your updated reasoning and final answer.

[Persona 1]

Refer to the other agent's opinion and your previous opinion to reevaluate the answer and reasoning. Maintain your Hollywood hiring manager persona. Provide your updated reasoning and final answer.

[Persona 2]

Refer to the other agent's opinion and your previous opinion to reevaluate the answer and reasoning. Maintain your biologist persona. Provide your updated reasoning and final answer.