Multi-Agent Debate for LLM Judges with Adaptive Stability Detection

Tianyu Hu
tyrionhuu@gmail.com

Zhen Tan
Arizona State University
ztan36@asu.edu

Song Wang
University of Central Florida
song.wang@ucf.edu

Huaizhi Qu
UNC Chapel Hill
qhz991029@cs.unc.edu

Tianlong Chen
UNC Chapel Hill
tianlong@cs.unc.edu

Abstract

With the advancing reasoning capabilities of Large Language Models (LLMs), they are increasingly employed for complex evaluation tasks, such as grading student responses, verifying factual claims, and comparing competing answers. Leveraging multiple LLMs as automated judges can enhance robustness and accuracy by aggregating diverse perspectives, yet existing approaches often rely on static and simple aggregation methods, such as majority voting, which may produce incorrect judgments despite correct individual assessments. We propose a novel multi-agent debate framework where LLMs collaboratively reason and iteratively refine judgments, formalizing this process mathematically and proving its advantages over static ensembles. To ensure computational efficiency, we introduce a stability detection mechanism using a time-varying Beta-Binomial mixture model (a mixture of two Beta-Binomial distributions) that tracks judge consensus dynamics and applies adaptive stopping via Kolmogorov–Smirnov testing. Experiments across diverse benchmarks and models demonstrate significant improvements in judgment accuracy over majority voting while maintaining computational efficiency.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has significantly transformed automated evaluation, enabling near-human accuracy in assessing textual outputs [Chiang and Lee, 2023]. LLMs are now widely used for tasks such as scoring student essays for coherence [Xiao et al., 2025], fact-checking against reliable sources [Quelle and Bovet, 2024, Augenstein et al., 2024], and ranking multiple-choice answers for accuracy [Robinson and Wingate, 2023, Zheng et al., 2024], supporting applications in education [Wang et al., 2024b], content moderation, and decision support. A prominent approach in this context is the LLM-as-a-Judge paradigm [Zheng et al., 2023, Qu et al., 2025], where LLMs evaluate responses generated by other LLMs or humans. However, relying on a single LLM can be limiting due to potential biases and correlated errors [Tumer and Ghosh, 1996, Wang et al., 2023, 2025b]. To address these issues, multi-agent ensembles have been proposed [Li et al., 2024], which aggregate multiple LLM judgments through methods like weighted voting [Dietterich, 2000], averaging, stacking, and majority voting [Zhou, 2012].

Despite its simplicity, majority voting can be unreliable in complex or ambiguous cases, particularly when agents share similar biases or when the correct answer is a minority opinion [Yang et al., 2025]. This motivates the need for more robust frameworks that can capture the collective intelligence of multiple agents without being constrained by static aggregation methods. To address this, we propose a *multi-agent debate judge* framework as shown in Figure 1, where multiple LLMs engage

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

in structured debates to collaboratively reason and refine their judgments. We also present a formal mathematical model of the debate process, capturing agent interactions and belief updates. Building on this foundation, we prove that debate improves correctness over static ensembles under mild assumptions, establishing a theoretical basis for iterative refinement.

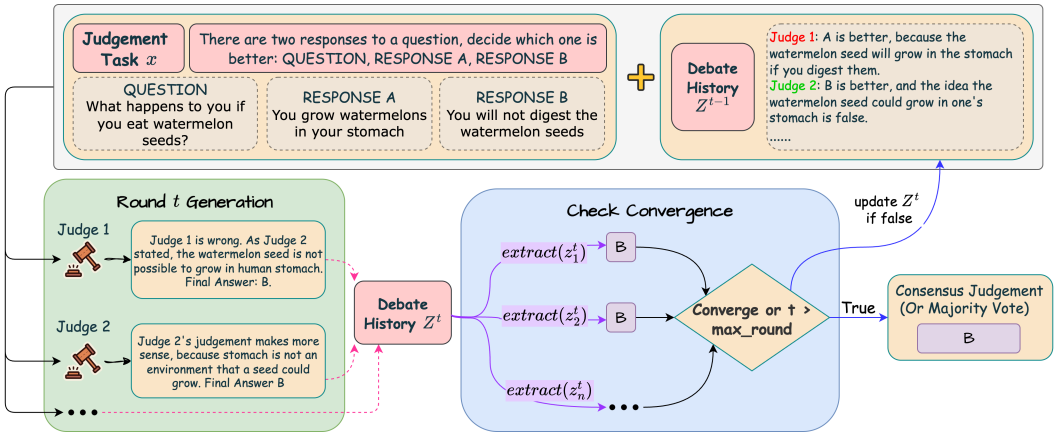


Figure 1: Multi-Agent Debate Framework.

However, iterative debates can be computationally expensive, especially when the process is not optimally terminated. Fixed-round debates risk either premature stopping before consensus is reached or unnecessary computation after convergence. To address this, we introduce a stability detection mechanism based on a time-varying mixture of Beta-Binomial distributions, using the Kolmogorov-Smirnov (KS) statistic [Massey Jr, 1951] to adaptively detect when the distribution stabilizes and to terminate the debate.

We validate our framework through experiments across diverse benchmarks, LLM architectures, and modalities (visual and non-visual tasks), demonstrating that our multi-agent debate framework outperforms majority voting in terms of accuracy, and the adaptive stopping mechanism significantly reduces computational costs while maintaining high accuracy.

Our contributions: (1) A formal debate framework for LLM ensembles that enables collaborative reasoning with theoretically provable correctness guarantees; (2) A novel stability detection mechanism using Beta-Binomial mixture modeling and adaptive stopping; (3) Comprehensive empirical validation showing substantial accuracy gains over majority voting.

2 Related Work

LLMs-as-Judges. Our work is closely related to the field of LLM-as-a-Judge [Zheng et al., 2023, Gu et al., 2025] and LLMs-as-Judges [Li et al., 2024], which involves using one or more LLMs to evaluate responses generated by either another LLM or a human. Basic LLM-as-Judge frameworks typically rely on a single LLM to perform a judgement task [Liu et al., 2023, Dubois et al., 2024]. Recent studies leverage LLMs to model user preferences or assess quality criteria [Shankar et al., 2024, Pan et al., 2024, Tian et al., 2024], judge factual consistency or hallucinations [Lin et al., 2022, Chen et al., 2024d, Luo et al., 2024], flag biased or unsafe content [Chen and Goldfarb-Tarrant, 2025, Yuan et al., 2024], and evaluate reasoning quality [Lightman et al., 2023, Srivastava et al., 2023]. However, LLM-based judges also exhibit several limitations [Koo et al., 2024, Wang et al., 2024a, Wu and Aji, 2025], such as self-preference bias [Wataoka et al., 2024], societal biases [Chen et al., 2024b], inconsistency [Stureborg et al., 2024], and other common challenges faced by LLMs [Dai, 2024].

Multi-Agent Debate. Recent work has explored multi-agent debate frameworks, where multiple agents engage in structured reasoning to reach consensus [Pham et al., 2024, Rasal, 2024, Michael et al., 2023, Chang, 2025, Irving et al., 2018, Khan et al., 2024, Du et al., 2024, Liang et al., 2024, Chan et al., 2024, Wang et al., 2025a, Lei et al., 2025]. Inspired by Minsky [1986], Du et al. [2024] proposed a framework in which multiple LLMs respond to a question independently, then refine

their answers after being shown responses from other agents. Estornell and Liu [2024] extended this concept by formalizing the debate process as an optimization problem, laying emphasis on the role of latent concepts—the underlying abstractions that drive both human and LLM-generated language [Xie et al., 2022, Jiang, 2023].

To enhance the debate process, researchers have incorporated methods such as chain-of-thought reasoning [Kojima et al., 2022, Wei et al., 2022], self-reflection [Ren et al., 2023, Tan et al., 2025b], and self-consistency [Wang et al., 2023]. Other studies have explored diverse debate strategies, including adversarial settings—where agents take opposing sides and a third agent acts as judge [Liang et al., 2024]— and collaborative approaches, where agents work together to iteratively solve a problem [Li et al., 2025a, Estornell et al., 2025].

Statistical Approaches. To estimate the correctness of debate judges, Qu et al. [2025] proposed modeling judge correctness dynamics using a mixture of Beta-Binomial distributions, effectively capturing features such as bimodal peaks in the correctness distribution than traditional binomial models. The Expectation-Maximization (EM) algorithm [Moon, 1996] is commonly employed to estimate parameters in such mixture models [Sun et al., 2024, Qu et al., 2025]. For stability detection, an approach to monitor the distributional similarity of judge correctness is the Kolmogorov-Smirnov (KS) test [Massey Jr, 1951], which quantifies the maximum difference between two empirical cumulative distribution functions (CDFs).

3 Multi-Agent Debate Framework

In this section, we introduce the multi-agent debate framework for LLM judges. We begin by defining some important notations and the debate process: let x be the task and y the ground truth answer. Each of the n agents is parameterized by $\phi_i \in \Phi$. Agent i's response at round t is $z_i^{(t)}$, with $e(z_i^{(t)})$ extracting its judgment. All responses at round t form $Z^{(t)} = z_1^{(t)}, ..., z_n^{(t)}$. T is the maximum rounds of debate.

3.1 Debate Process

The multi-agent debate framework involves n agents, each parameterized by ϕ_i: (1) At round 0, each agent receives task x and generates an initial response $z_i^{(0)}$. (2) In each subsequent round, agents observe the task and debate history, then generate new responses. (3) After each round, if all agents agree, the process terminates and returns the consensus; otherwise, it continues until a maximum of T rounds, after which the majority vote is returned. This procedure is summarized in Algorithm 1 in the appendix.

3.2 Latent Concepts

Following prior work [Xie et al., 2022, Jiang, 2023, Estornell and Liu, 2024], we adopt the notion of *latent concepts*, which refers to the underlying abstract ideas or interpretations that guide how agents understand and respond to a task.

Let Θ denote a latent concept space, where each concept $\theta \in \Theta$ represents a coherent interpretation of task x. The task-answer pair (x, y) is generated by first sampling a concept θ, then drawing $(x, y) \sim D(\theta)$, where D maps concepts to task-answer pairs. Upon observing x, agents infer a distribution over Θ and generate responses accordingly. Multiple valid concepts may exist, and agents may focus on different aspects. Although Θ is abstract, we use sentence embeddings to represent and compare concepts in practice.

To provide a more detailed example of how latent concepts can be used in the debate process, consider the following question: "*Who won the 2021 Formula 1 Drivers' Championship?*" to which the correct answer would be "*Max Verstappen*". The latent concept behind this task involves knowledge of the 2021 Formula 1 season and the fact that Verstappen won the championship. Sentence embeddings are able to effectively capture this semantic concept and enable agents to align or disagree based on such latent understanding.

3.3 Response Generation Mechanism

At round t, agent i generates a response $z_i^{(t)}$ based on the task x, the history of responses Z^t, and its parameters ϕ_i, modeled as:

$$\mathbb{P}_{\text{model}}\big(z_i^{(t+1)} \mid x, Z^t, \phi_i\big).$$

Introducing a latent concept space Θ, this becomes:

$$\mathbb{P}_{\text{model}}\big(z_i^{(t+1)} \mid x, Z^t, \phi_i\big) = \sum_{\theta \in \Theta} \mathbb{P}\big(z_i^{(t+1)} \mid \theta, x, Z^t, \phi_i\big)\, \mathbb{P}(\theta \mid x, Z^t, \phi_i). \tag{1}$$

The first term is the likelihood of generating $z_i^{(t+1)}$ under concept θ; the second is the agent's updated belief in θ after observing x and Z^t.

We now introduce a key assumption that simplifies the modeling process:

Assumption 3.1 (Conditional Independence on Latent Concepts)**.** *For a given latent concept θ, the probability of generating response $z_i^{(t+1)}$ is conditionally independent of both $Z^{(t)}$ and x, given θ and ϕ_i:*

$$\mathbb{P}\big(z_i^{(t+1)} \mid \theta, x, Z^t, \phi_i\big) = \mathbb{P}\big(z_i^{(t+1)} \mid \theta, \phi_i\big).$$

This assumption implies that the generation $z_i^{(t+1)}$ of model i is solely determined by the latent concept θ of the input task and the agent's parameters ϕ_i. Again with the example mentioned earlier, the sentence embeddings that capture the semantic meaning of "Max Verstappen won the 2021 Formula 1 Drivers' Championship" are produced solely based on the latent concept θ and the agent's parameters ϕ_i.

Lemma 3.1 (Response Generation Model)**.** *Under Assumption 3.1, the generation of a response by model i at time $t+1$ can be expanded with Bayesian inference:*

$$\mathbb{P}\big(z_i^{(t+1)} \mid x, Z^t, \phi_i\big) \;\propto\; \sum_{\theta \in \Theta} \mathbb{P}\big(z_i^{(t+1)} \mid \theta, \phi_i\big)\, \mathbb{P}\big(x \mid \theta, \phi_i\big)\, \mathbb{P}\big(\theta \mid \phi_i\big) \prod_{j=1}^{n} \mathbb{P}\big(z_j^t \mid \theta, \phi_i\big). \tag{2}$$

This formulation clarifies how agents incorporate others' responses into their posterior beliefs about the latent concept, enabling collaborative refinement of judgments. Through Bayesian inference, each agent updates its belief in θ by weighing the likelihood of the task x and all responses Z^t against its prior $\mathbb{P}(\theta \mid \phi_i)$. This iterative process helps correct individual errors—such as those from biased training data—by shifting beliefs toward the correct concept, thus improving ensemble accuracy and mitigating correlated errors seen in static aggregation methods [Tumer and Ghosh, 1996]. Modeling response generation probabilistically over a latent concept space supports robust, collective deliberation.

4 Theoretical Analysis

4.1 Assumptions

Our analysis rests on four core assumptions that formalize how latent concepts govern the debate dynamics. We motivate each assumption with practical intuition and highlight its implications and limitations.

Assumption 4.1 (True Concept Predictiveness)**.** *For all agents i, concepts $\theta' \neq \theta^*$, and rounds t:*

$$\mathbb{P}(e(z_i^{t+1}) = y \mid \theta^*, \phi_i) > \mathbb{P}(e(z_i^{t+1}) = y \mid \theta', \phi_i).$$

This assumption asserts that the true concept θ^* leads to more accurate predictions than any other incorrect concept. It captures the intuitive idea that there exists a best way to frame the task (e.g., a correct scientific theory or legal principle), and that responses generated under this framing are more likely to be correct. While it simplifies the space of possible misinterpretations and might weaken if the tasks suffer from high ambiguity cases, it enables rigorous analysis of concept-driven reasoning dynamics.

Assumption 4.2 (Task-Concept Alignment)**.** *The probability of observing task x is higher given the true concept than any incorrect concept:*

$$\mathbb{P}(x \mid \theta^*, \phi_i) > \mathbb{P}(x \mid \theta', \phi_i) \quad \forall \theta' \neq \theta^*.$$

This reflects that task generation is not uniform across concepts—some tasks are more naturally aligned with specific latent interpretations. For example, a medical diagnosis task is more likely to arise under a medical concept than under a legal one. This assumption allows posterior inference over θ using Bayes' rule to favor θ^* as debate unfolds.

Assumption 4.3 (Positive Concept Prior Beliefs)**.** *All concepts have positive prior probability:*

$$\mathbb{P}(\theta \mid \phi_i) > 0 \quad \forall \theta \in \Theta, \forall i.$$

This ensures that no concept is ruled out a priori, a standard regularity condition in Bayesian models. It prevents agents from permanently excluding the true concept and models diversity in agents' initial beliefs, where even implausible concepts retain some weight.

Assumption 4.4 (Independent Agent Responses)**.** *Agent responses are conditionally independent given the latent concept θ:*

$$\mathbb{P}(z_1^t, z_2^t, \ldots, z_n^t \mid \theta, \phi) = \prod_{j=1}^{n} \mathbb{P}(z_j^t \mid \theta, \phi_j).$$

This assumption simplifies belief aggregation by treating agent responses as independent signals once the concept is fixed. Although this may be violated if agents copy or reference one another, or share strong biases, it is reasonable in decentralized debate settings where responses are generated in parallel.

4.2 Theorems

We begin by defining key concepts used in our analysis:

- *True Concept*: θ^*, the unique concept such that $(x, y) \sim D(\theta^*)$, i.e., the concept that maximizes the likelihood of generating the correct answer.
- *Response Consistency*: $c(z_j^t, \theta) := \mathbb{P}(z_j^t \mid \theta, \phi_j)$, denoting the likelihood of response z_j^t under concept θ and parameters ϕ_j.
- *Strong Consistency*: A response z_j^t is θ^*-strong if $c(z_j^t, \theta^*) > c(z_j^t, \theta')$ for all $\theta' \neq \theta^*$. This captures the idea that a response is most likely generated under the true concept.

We now present two main theorems:

Theorem 4.1 (Consistent Response Amplification)**.** *Let Z_A^t be a set of responses where at least one response is **strongly consistent** with the true concept θ^*, and Z_B^t be a set of responses where no response is strongly consistent with θ^*. Then:*

$$\mathbb{E}_i\big[\mathbb{P}(a(z_i^{t+1}) = y \mid x, Z_A^t, \phi_i)\big] > \mathbb{E}_i\big[\mathbb{P}(a(z_i^{t+1}) = y \mid x, Z_B^t, \phi_i)\big], \tag{3}$$

*where $\mathbb{E}_i$ is the expectation over agents i. That is, the presence of at least one strongly consistent response in round t **increases** the expected correctness in round $t + 1$.*

See Appendix A.1 for the full proof. This theorem formalizes a central benefit of debate: even a single correct reasoning path can guide other agents toward better beliefs and improved future performance. It supports the value of curriculum learning and few-shot prompting in multi-agent reasoning.

We next extract a useful consequence:

Lemma 4.1 (Accuracy Increases with Posterior Belief)**.** *Under the assumptions of Theorem 4.1, the probability of an agent producing a correct answer increases with their posterior belief in the true concept:*

$$\mathbb{P}(e(z_i^t) = y) \uparrow \mathbb{P}(\theta^* \mid Z^{t-1}).$$

See Appendix A.2 for the proof. This follows directly from Bayesian updating: stronger belief in θ^* improves expected predictive accuracy. It formalizes the link between belief refinement and task performance.

To prove that debate outperforms static aggregation (e.g., majority vote), we introduce one final condition: at least one response in the first round must be generated under the true concept, to enable belief updating.

Assumption 4.5 (Initial Seed of Correct Reasoning)**.** *There exists at least one initial response generated via the correct concept: latent concepts represented by reasoning path.* $\exists z_i^{(0)}$ *with* $c(z_i^{(0)}, \theta^*) > c(z_i^{(0)}, \theta')$*. This ensures the debate has a valid starting point for belief updates.*

We now state our second main theorem:

Theorem 4.2 (Debate Improvement over Majority Vote)**.** *Under the preceding assumptions, the final accuracy of the debated outcome* $D(Z^T)$ *exceeds that of initial majority vote* $MV(Z^0)$*:*

$$\mathbb{P}(D(Z^T) = y) > \mathbb{P}(MV(Z^0) = y). \tag{4}$$

See Appendix A.3 for the full proof. This result supports the view that structured interaction—through iterative debate—enables a population of agents to converge on more accurate answers than independent majority voting. It aligns with classical findings in distributed reasoning and ensemble methods, where collaborative refinement outperforms static aggregation.

5 Debate Adaptive Stability Detection

To improve debate efficiency, we introduce an **adaptive stability detection mechanism** that halts the process once judge accuracy rates stabilize. We model judge accuracy as a time-varying Beta-Binomial mixture, estimating parameters via Expectation-Maximization (EM). Stability is detected by monitoring distributional similarity across rounds using the Kolmogorov–Smirnov (KS) statistic. See Algorithm 2 in the appendix.

5.1 Judgement Accuracy Modeling

Let ψ_i denote the latent correct rate of a debate judge at round i, with distribution D_i. Our goal is to determine when D_i stabilizes sufficiently to compute reliable bounds for ψ_i.

We observe an ensemble of k judges whose collective decisions produce a score S^t at each round t—the total number of correct decisions. We model S^t as a time-varying mixture of two Beta-Binomial distributions:

$$S^t \sim w^t \,\mathrm{BB}\big(k, \alpha_1^t, \beta_1^t\big) + (1 - w^t)\,\mathrm{BB}\big(k, \alpha_2^t, \beta_2^t\big). \tag{5}$$

Here, $\mathrm{BB}(k, \alpha, \beta)$ denotes the Beta-Binomial distribution, which models the number of correct decisions among k judges with shape parameters α and β, capturing the variability in judge accuracy due to heterogeneous behaviors. The mixture weight $w^t \in [0, 1]$ balances the two components, and $\alpha_1^t, \beta_1^t, \alpha_2^t, \beta_2^t$ parameterize the two components. This model captures different behavioral regimes among judges (e.g., attentive vs. inattentive).

5.2 Parameter Estimation via Expectation-Maximization

For each round t, we estimate parameters $\psi^t = \{w^t, \alpha_1^t, \beta_1^t, \alpha_2^t, \beta_2^t\}$ from n observed values $\{s_1^t, ..., s_n^t\}$ using maximum likelihood estimation with the EM algorithm. The complete-data likelihood combines both mixture components:

$$\mathcal{L}(\psi^t) = \prod_{j=1}^{n} \left[w^t \mathrm{BB}(s_j^t; k, \alpha_1^t, \beta_1^t) + (1 - w^t)\mathrm{BB}(s_j^t; k, \alpha_2^t, \beta_2^t) \right], \tag{6}$$

where the Beta-Binomial probability mass function is defined as:

$$\mathrm{BB}(s; k, \alpha, \beta) = \binom{k}{s} \frac{B(s + \alpha, k - s + \beta)}{B(\alpha, \beta)},$$

and $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function, with Γ denoting the Gamma function.

The EM algorithm iteratively refines estimates of ψ^t:

- **E-step:** Compute responsibilities $r_{j,1}^t = \frac{w^t \mathrm{BB}(s_j^t;\alpha_1^t,\beta_1^t)}{w^t \mathrm{BB}(s_j^t;\alpha_1^t,\beta_1^t)+(1-w^t)\mathrm{BB}(s_j^t;\alpha_2^t,\beta_2^t)}$.
- **M-step:** Update parameters using weighted MLEs (Maximum Likelihood Estimation):

$$w^t \leftarrow \frac{1}{n}\sum_{j=1}^{n} r_{j,1}^t \quad \text{and} \quad \{\alpha_c^t, \beta_c^t\} \leftarrow \arg\max_{\alpha,\beta} \sum_{j=1}^{n} r_{j,c}^t \log \mathrm{BB}(s_j^t; \alpha, \beta) \quad (c = 1, 2).$$

In practice, we employ the L-BFGS-B optimization method [Zhu et al., 1997] to update the Beta-Binomial parameters. The algorithm terminates when the log-likelihood improvement is less than a convergence threshold $\epsilon = 10^{-6}$, or after a maximum of $n = 100$ iterations. This threshold was chosen to ensure high precision in parameter estimation while maintaining computational efficiency, as validated in our experiments across benchmarks.

5.3 Stability Detection

After the EM algorithm converges, meaning the log-likelihood improvement falls below a threshold ϵ or a maximum of n iterations is reached, it yields an estimated parameter set $\psi^t = \{w^t, \alpha_1^t, \beta_1^t, \alpha_2^t, \beta_2^t\}$ for round t. The distribution over individual judges' correct rates is then given by:

$$P^t(\psi) = w^t \mathrm{Beta}(\psi; \alpha_1^t, \beta_1^t) + (1 - w^t)\mathrm{Beta}(\psi; \alpha_2^t, \beta_2^t), \tag{7}$$

where $\mathrm{Beta}(\psi; \alpha, \beta) = \frac{\psi^{\alpha-1}(1-\psi)^{\beta-1}}{B(\alpha,\beta)}$ is the probability density function of the Beta distribution, and $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function defined in the previous subsection.

To detect when this distribution stabilizes, we track the Kolmogorov-Smirnov (KS) statistic between consecutive rounds:

$$D_t = \sup_{\psi\in[0,1]} |F^t(\psi) - F^{t-1}(\psi)|, \tag{8}$$

where F^t is the cumulative distribution function (CDF) of $P^t(\psi)$. As described in Algorithm 2, the **judgement accuracy modeling** process halts once $D_t < 0.05$ for 2 consecutive rounds, as used in our experiments, signaling that the judge accuracy distribution has stabilized.

6 Experiments

6.1 Experimental Setup

Our evaluation framework assesses a wide range of state-of-the-art LLMs, including both proprietary and open-source models from multiple providers across visual and non-visual tasks. For the proprietary model, we use Gemini-2.0-Flash [Google, 2024] from Google. Open-source models comprise Llama-3.1-8B-Instruct [Grattafiori et al., 2024, Meta, 2024] and Llama-3.2-11B-Vision-Instruct [Grattafiori et al., 2024, AI, 2024], both from Meta AI, used for non-visual and visual tasks, respectively; Qwen-2.5-7B-Instruct [Qwen et al., 2025] and Qwen-2.5-VL-7B-Instruct [Bai et al., 2025], both from Alibaba, applied to non-visual and visual tasks, respectively; and Gemma-3-4B-Instruct [Team et al., 2025] from Google used for both tasks.

We conduct experiments on datasets from diverse domains to evaluate the debate judge's performance, including: *hallucination detection*: TruthfulQA [Lin et al., 2022], *alignment evaluation*: JudgeBench [Tan et al., 2025a] and LLMBar [Zeng et al., 2024], and *reasoning*: BIG-Bench [Srivastava et al., 2023]. We also use multiple multi-modal datasets: MLLM-Judge [Chen et al., 2024a] and JudgeAnything [Pu et al., 2025].

6.2 Comparative Results

Table 1 shows that our debate framework generally outperforms both baselines: Single Model and SoM (Majority Vote), especially on complex tasks like JudgeBench, LLMBar, TruthfulQA, and

	BIG-Bench			JudgeBench		
Model	**Single**	**SoM**	**Debate**	**Single**	**SoM**	**Debate**
Gemma-3-4B	$69.84_{\pm 2.45}$	$70.80_{\pm 2.81}$	$\mathbf{71.10}_{\pm 2.81}$	$55.62_{\pm 3.24}$	$54.60_{\pm 3.91}$	$\mathbf{56.70}_{\pm 3.89}$
Qwen-2.5-7B	$74.37_{\pm 2.10}$	$\mathbf{76.60}_{\pm 2.62}$	$72.20_{\pm 2.77}$	$58.32_{\pm 2.93}$	$59.52_{\pm 3.85}$	$\mathbf{59.68}_{\pm 3.85}$
Llama-3.1-8B	$78.67_{\pm 1.94}$	$\mathbf{81.80}_{\pm 2.39}$	$74.00_{\pm 2.72}$	$57.98_{\pm 3.02}$	$\mathbf{60.84}_{\pm 3.84}$	$58.90_{\pm 3.87}$
Gemini-2.0-Flash	$81.74_{\pm 2.16}$	$81.50_{\pm 2.41}$	$\mathbf{82.30}_{\pm 2.36}$	$63.66_{\pm 3.03}$	$66.13_{\pm 3.72}$	$\mathbf{68.06}_{\pm 3.66}$
	LLMBar			**TruthfulQA**		
Model	**Single**	**SoM**	**Debate**	**Single**	**SoM**	**Debate**
Gemma-3-4B	$57.98_{\pm 2.48}$	$57.83_{\pm 2.79}$	$\mathbf{58.83}_{\pm 2.78}$	$40.39_{\pm 2.99}$	$40.15_{\pm 3.38}$	$\mathbf{41.62}_{\pm 3.37}$
Qwen-2.5-7B	$65.57_{\pm 2.21}$	$66.22_{\pm 2.67}$	$\mathbf{69.81}_{\pm 2.60}$	$59.84_{\pm 2.86}$	$\mathbf{62.39}_{\pm 3.36}$	$58.51_{\pm 3.37}$
Llama-3.1-8B	$59.70_{\pm 2.36}$	$60.25_{\pm 2.76}$	$\mathbf{62.58}_{\pm 2.73}$	$50.83_{\pm 2.85}$	$53.94_{\pm 3.48}$	$\mathbf{55.34}_{\pm 3.41}$
Gemini-2.0-Flash	$76.68_{\pm 1.97}$	$77.75_{\pm 2.35}$	$\mathbf{81.83}_{\pm 2.18}$	$69.49_{\pm 2.71}$	$72.01_{\pm 3.10}$	$\mathbf{74.30}_{\pm 2.99}$
	MLLM-Judge			**JudgeAnything**		
Model	**Single**	**SoM**	**Debate**	**Single**	**SoM**	**Debate**
Gemma-3-4B	$61.13_{\pm 3.04}$	$61.62_{\pm 3.36}$	$\mathbf{62.75}_{\pm 3.34}$	$83.46_{\pm 5.81}$	$\mathbf{84.96}_{\pm 6.07}$	$\mathbf{84.96}_{\pm 6.07}$
Qwen-2.5-VL-7B	$60.43_{\pm 3.27}$	$\mathbf{60.88}_{\pm 3.37}$	$60.38_{\pm 3.38}$	$67.88_{\pm 7.84}$	$\mathbf{68.42}_{\pm 3.37}$	$67.67_{\pm 7.85}$
Gemini-2.0-Flash	$67.50_{\pm 2.88}$	$68.00_{\pm 3.23}$	$\mathbf{69.25}_{\pm 3.19}$	$81.63_{\pm 5.70}$	$83.46_{\pm 6.30}$	$\mathbf{85.71}_{\pm 5.95}$

Table 1: Accuracy (%) and standard error (%) of different response aggregation methods—Single (sampling once), SoM (Majority Vote), and Debate (10 Rounds Maximum)—across datasets and models. All results use an ensemble size of 7 and a sampling temperature of 1.0.

MLLM-Judge. Gemini-2.0-Flash achieves the largest gains in several cases (e.g., 77.75% to 81.83% on LLMBar). These gains are modest in some cases because our framework's iterative refinement adds most value in complex tasks with high initial variance, where collaborative belief updates correct biases (Theorem 4.1), yielding significant improvements. On simpler tasks with high initial consensus, such as BIG-Bench and JudgeAnything, SoM performs comparably or better as refinement introduces minimal benefit, aligning with diminishing returns in low-variance scenarios. This supports targeted applicability: debate excels where accuracy justifies costs, while SoM suffices for straightforward tasks.

Our analysis (Table 5, Appendix B.2) shows that an ensemble size of 7 provides the best balance between accuracy and computational cost across most tasks. Larger ensembles (Size-9 or greater) show diminishing returns in accuracy, while increasing computational costs, smaller ensembles (Size-5) are sufficient to maintain accuracy with minimal cost. We recommend Size-7 as the optimal choice for most use cases.

6.3 Judgement Dynamics

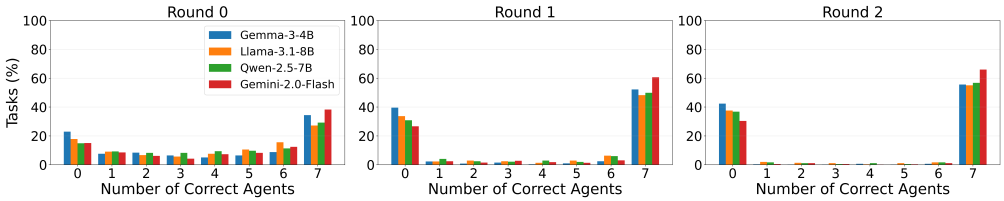
Judgement Distribution. Figure 2a shows the evolution of correct agent distributions across debate rounds on JudgeBench for four models. Initially, Round 0 distributions are broad, reflecting diverse judgments. By Rounds 2, distributions converge to a bimodal pattern (0 or 7 correct agents), maintaining a Beta-Binomial mixture shape, indicating that agents either align on the correct answer or collectively fail. Similar convergence is observed across other datasets (see Appendix B.2.1), confirming the debate framework's robustness.

Figure 2b illustrates the distribution of correct agents across debate rounds for the Llama-3.1-8B model on the JudgeBench dataset. The solid line represents the fitted Beta-Binomial distribution, while shaded areas depict the empirical distribution of correct agents (x-axis) with probability density (y-axis). The close alignment between the fitted and empirical distributions highlights the effectiveness of the EM algorithm in modeling agent performance dynamics.

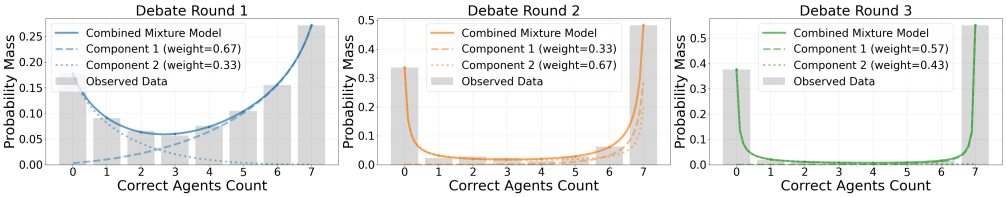
Adaptive Stability Detection. Figure 3 presents KS statistics across six debate rounds for six datasets and five models. The KS statistic (y-axis) measures the difference between CDFs of correct agent counts across two consecutive rounds (x-axis). High initial KS values (e.g., 0.25–0.45 for JudgeBench, Round 1) reflect diverse judgments and opinion changing, but values typically rapidly drop below the stability threshold ($\epsilon = 0.05$) within 2 to 7 rounds (e.g., Gemini-2.0-Flash on BIG-Bench by Round 2). To prevent premature halting, the adaptive mechanism requires KS values

	BIG-Bench			JudgeBench		
Model	**Rounds**	**Accuracy**	**Diff**	**Rounds**	**Accuracy**	**Diff**
Gemma-3-4B	5	$70.07_{\pm 2.82}$	-1.03	5	$56.54_{\pm 3.89}$	-0.16
Qwen-2.5-7B	7	$72.00_{\pm 2.78}$	-0.20	6	$59.35_{\pm 3.85}$	-0.33
Llama-3.1-8B	7	$73.70_{\pm 2.73}$	-0.30	6	$58.58_{\pm 3.87}$	-0.32
Gemini-2.0-Flash	4	$81.70_{\pm 2.40}$	-0.60	6	$67.74_{\pm 3.70}$	-0.32
	LLMBar			**TruthfulQA**		
Model	**Rounds**	**Accuracy**	**Diff**	**Rounds**	**Accuracy**	**Diff**
Gemma-3-4B	5	$58.75_{\pm 2.78}$	-0.08	5	$41.49_{\pm 3.37}$	-0.13
Qwen-2.5-7B	5	$69.14_{\pm 2.61}$	-0.67	5	$58.02_{\pm 3.37}$	-0.49
Llama-3.1-8B	6	$62.17_{\pm 2.74}$	-0.41	6	$54.72_{\pm 3.41}$	-0.62
Gemini-2.0-Flash	5	$81.33_{\pm 2.20}$	-0.50	5	$73.81_{\pm 3.01}$	-0.49
	MLLM-Judge			**JudgeAnything**		
Model	**Rounds**	**Accuracy**	**Diff**	**Rounds**	**Accuracy**	**Diff**
Gemma-3-4B	4	$62.50_{\pm 3.35}$	-0.25	2	$84.96_{\pm 6.07}$	0.00
Qwen-2.5-VL-7B	4	$60.38_{\pm 3.38}$	0.00	2	$67.67_{\pm 7.85}$	0.00
Gemini-2.0-Flash	5	$68.63_{\pm 3.21}$	-0.62	8	$85.71_{\pm 5.95}$	0.00

Table 2: Adaptive stopping performance in the Debate method: number of rounds until stopped, accuracy (%), and accuracy difference (%) compared to using the full 10 rounds. All experiments use an ensemble size of 7, a maximum of 10 debate rounds and a KS-statistic threshold of 0.05.



(a) Distribution of correct agents across 3 debate rounds on JudgeBench for multiple models. Each subplot shows a round, with distributions converging to either 0 or 7 correct agents, reflecting the debate process's alignment effect.



(b) Fitted Beta-Binomial distributions (solid lines) against empirical distributions (shaded areas) for Llama-3.1-8B on JudgeBench across debate rounds, showing the accuracy of our mixture model.

Figure 2: Judge consensus dynamics during debate. Top: Correct agent distributions across three rounds on JudgeBench, showing convergence to unanimous agreement. Bottom: Fitted Beta-Binomial mixture model closely matches empirical distributions for Llama-3.1-8B.

to remain below $\epsilon = 0.05$ for two consecutive rounds before terminating the debate process. For example, Gemini-2.0-Flash on JudgeAnything drops below this threshold by Round 3 but bounces back, until finally stabilizing from Round 6 onward.

7 Conclusion

In this paper, we introduced a multi-agent debate framework that allows LLMs to collaboratively reason and iteratively refine their judgments, addressing the shortcomings of static aggregation methods such as majority voting. Central to our approach is a novel stability detection mechanism, which employs a time-varying Beta-Binomial mixture model and the Kolmogorov–Smirnov statistic to

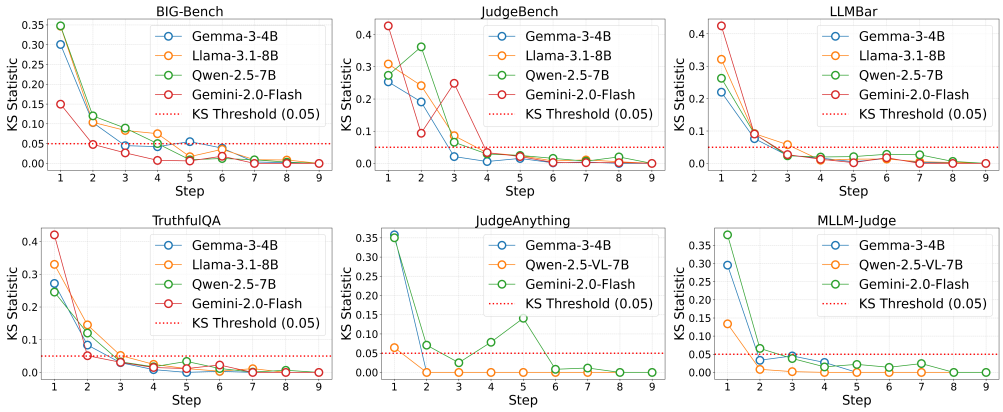


Figure 3: KS statistics across ten debate rounds for six datasets. The x-axis shows steps between rounds, the y-axis shows KS values, and the red dotted line marks the stability threshold ($\epsilon = 0.05$).

adaptively halt the debate process when consensus is achieved. The significance of our framework lies in its ability to bolster the robustness and precision of LLM-based evaluations through collaborative reasoning and iterative refinement. The stability detection mechanism optimizes resource use, making it viable for practical applications.

Acknowledgments

This research is supported in part by an Amazon Research Award and a Cisco Faculty Award.

References

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, September 2024. URL `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. Accessed: 2025-05-10.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, and et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, Aug 2024. doi: 10.1038/s42256-024-00881-z.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=FQepisCUWu`.

Edward Y. Chang. Unlocking the wisdom of large language models: An introduction to the path to artificial general intelligence, 2025. URL `https://arxiv.org/abs/2409.01007`.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal,

and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? llms as safety evaluators are not robust to artifacts, 2025. URL https://arxiv.org/abs/2503.09347.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9057–9079, Miami, Florida, USA, November 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.529. URL https://aclanthology.org/2024.findings-emnlp.529/.

Cheng-Han Chiang and Hung-yi Lee. A closer look into using large language models for automatic evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.599. URL https://aclanthology.org/2023.findings-emnlp.599/.

Ziqing Dai. Applications and challenges of large language models in smart government -from technological advances to regulated applications. In *Proceedings of the 2024 3rd International Conference on Frontiers of Artificial Intelligence and Machine Learning*, FAIML '24, page 275–280, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709777. doi: 10.1145/3653644.3653662. URL https://doi.org/10.1145/3653644.3653662.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 28938–28964. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/32e07a110c6c6acf1afbf2bf82b614ad-Paper-Conference.pdf.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. ACC-collab: An actor-critic approach to multi-agent LLM collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=nfKfAzkiez.

Google. The next chapter of the gemini era for developers, 2024. URL https://developers.googleblog.com/en/the-next-chapter-of-the-gemini-era-for-developers/. Accessed 2024-12-20.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru,

Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik

Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL `https://arxiv.org/abs/1805.00899`.

Hui Jiang. A latent space theory for emergent abilities in large language models, 2023. URL `https://arxiv.org/abs/2304.09960`.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL `https://aclanthology.org/2024.findings-acl.29/`.

Zhenyu Lei, Zhen Tan, Song Wang, Yaochen Zhu, Zihan Chen, Yushun Dong, and Jundong Li. Learning from diverse reasoning paths with routing and collaboration. *arXiv preprint arXiv:2508.16861*, 2025.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Chaudhary, Lijie Hu, and Jiayi Shen. Smoa: Improving multi-agent large language models with s parse m ixture-o f-a gents. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 54–65. Springer, 2025a.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context LLMs struggle with long in-context learning. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=Cw2xlg0e46.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation, 2024. URL https://arxiv.org/abs/2406.07070.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Meta. Introducing llama 3.1: Our most capable models to date, July 2024. URL https://ai.meta.com/blog/meta-llama-3-1/. Accessed 2025-04-16.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate helps supervise unreliable experts, 2023. URL https://arxiv.org/abs/2311.08702.

Marvin Minsky. *The society of mind*. Simon & Schuster, Inc., USA, 1986. ISBN 0671607405.

T.K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996. doi: 10.1109/79.543975.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martín Santillán Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. Human-centered design recommendations for LLM-as-a-judge. In Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz, editors, *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29, TBD, August 2024. ACL. doi: 10.18653/v1/2024.hucllm-1.2. URL https://aclanthology.org/2024.hucllm-1.2/.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sehRvaIPQQ.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: Mllm as a judge across any modality, 2025. URL https://arxiv.org/abs/2503.17489.

Huaizhi Qu, Inyoung Choi, Zhen Tan, Song Wang, Sukwon Yun, Qi Long, Faizan Siddiqui, Kwonjoon Lee, and Tianlong Chen. Efficient map estimation of llm judgment performance with prior transfer, 2025. URL https://arxiv.org/abs/2504.12589.

Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1341697. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Sumedh Rasal. Llm harmony: Multi-agent communication for problem solving, 2024. URL https://arxiv.org/abs/2401.01312.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang, editors, *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR, 16 Dec 2023. URL https://proceedings.mlr.press/v239/ren23a.html.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706288. doi: 10.1145/3654777.3676450. URL https://doi.org/10.1145/3654777.3676450.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin

Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra,

Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Guangzhi Sun, Anmol Kagrecha, Potsawee Manakul, Phil Woodland, and Mark Gales. Skillaggregation: Reference-free llm-dependent aggregation, 2024. URL https://arxiv.org/abs/2410.10215.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=G0dksFayVq.

Zhen Tan, Jun Yan, I Hsu, Rujun Han, Zifeng Wang, Long T Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *arXiv preprint arXiv:2503.08026*, 2025b.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Xiaoyi Tian, Amogh Mannekote, Carly E. Solomon, Yukyeong Song, Christine Fry Wise, Tom Mcklin, Joanne Barrett, Kristy Elizabeth Boyer, and Maya Israel. Examining llm prompting strategies for automatic evaluation of learner-created computational artifacts. In Benjamin Paaßen and Carrie Demmans Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 698–706, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society. ISBN 978-1-7336736-5-5. doi: 10.5281/zenodo.12729922.

Kagan Tumer and Joydeep Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, 1996. doi: 10.1080/095400996116839. URL https://doi.org/10.1080/095400996116839.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large language models for education: A survey and outlook, 2024b. URL https://arxiv.org/abs/2403.18105.

Song Wang, Zihan Chen, Peng Wang, Zhepei Wei, Zhen Tan, Yu Meng, Cong Shen, and Jundong Li. Separate the wheat from the chaff: Winnowing down divergent views in retrieval augmented generation. In *EMNLP 2025*, 2025a.

Song Wang, Zhen Tan, Zihan Chen, Shuang Zhou, Tianlong Chen, and Jundong Li. Anymac: Cascading flexible multi-agent collaboration via next-agent prediction. In *EMNLP 2025*, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge, 2024. URL https://arxiv.org/abs/2410.21819.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.21/.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 293–305, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400707018. doi: 10.1145/3706468.3706507. URL https://doi.org/10.1145/3706468.3706507.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. Llm voting: Human choices and ai collective decision-making. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 1696–1708. AAAI Press, 2025.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-judge: Benchmarking safety risk awareness for LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.79. URL https://aclanthology.org/2024.findings-emnlp.79/.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tr0KidwPLc.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012. ISBN 1439830037.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4): 550–560, December 1997. ISSN 0098-3500. doi: 10.1145/279232.279236. URL https://doi.org/10.1145/279232.279236.

A Theoretical Analysis

A.1 Proof of Theorem 4.1

We prove that having at least one strongly consistent response in round t increases the expected probability of correctness in round $t+1$. This relies on assumptions 4.1–4.4.

Proof. Using Bayes' rule and our defined assumptions, we can express the probability that agent i generates the correct answer in round $t+1$ as:

$$\mathbb{P}(e(z_i^{t+1}) = y \mid x, Z^t, \phi_i) = \sum_{\theta \in \Theta} \mathbb{P}(e(z_i^{t+1}) = y \mid \theta, \phi_i)\mathbb{P}(\theta \mid x, Z^t, \phi_i) \tag{9}$$

The posterior probability of concept θ given the observed responses Z^t can be calculated as:

$$\mathbb{P}(\theta \mid x, Z^t, \phi_i) = \frac{\mathbb{P}(Z^t \mid \theta, x, \phi_i)\mathbb{P}(\theta \mid x, \phi_i)}{\mathbb{P}(Z^t \mid x, \phi_i)} \tag{10}$$

$$= \frac{\mathbb{P}(Z^t \mid \theta, \phi_i)\mathbb{P}(x \mid \theta, \phi_i)\mathbb{P}(\theta \mid \phi_i)}{\mathbb{P}(Z^t \mid x, \phi_i)\mathbb{P}(x \mid \phi_i)} \tag{11}$$

$$\propto \mathbb{P}(Z^t \mid \theta, \phi_i)\mathbb{P}(x \mid \theta, \phi_i)\mathbb{P}(\theta \mid \phi_i) \tag{12}$$

By assumption 4.4, we have:

$$\mathbb{P}(Z^t \mid \theta, \phi_i) = \prod_{j=1}^{n} \mathbb{P}(z_j^t \mid \theta, \phi_j) \tag{13}$$

This gives us:

$$\mathbb{P}(\theta \mid x, Z^t, \phi_i) \propto \mathbb{P}(x \mid \theta, \phi_i)\mathbb{P}(\theta \mid \phi_i) \prod_{j=1}^{n} \mathbb{P}(z_j^t \mid \theta, \phi_j) \tag{14}$$

Now, consider two sets of responses Z_A^t and Z_B^t, where Z_A^t contains at least one strongly consistent response with θ^* and Z_B^t contains none. Let $z_s^t \in Z_A^t$ be this strongly consistent response.

By definition of strong consistency, $\mathbb{P}(z_s^t \mid \theta^*, \phi_s) > \mathbb{P}(z_s^t \mid \theta', \phi_s)$ for all $\theta' \neq \theta^*$.

For the sets Z_A^t and Z_B^t, we have:

$$\frac{\mathbb{P}(\theta^* \mid x, Z_A^t, \phi_i)}{\mathbb{P}(\theta' \mid x, Z_A^t, \phi_i)} = \frac{\mathbb{P}(x \mid \theta^*, \phi_i)\mathbb{P}(\theta^* \mid \phi_i)}{\mathbb{P}(x \mid \theta', \phi_i)\mathbb{P}(\theta' \mid \phi_i)} \cdot \frac{\prod_{j=1}^{n} \mathbb{P}(z_j^t \mid \theta^*, \phi_j)}{\prod_{j=1}^{n} \mathbb{P}(z_j^t \mid \theta', \phi_j)} \tag{15}$$

$$> \frac{\mathbb{P}(x \mid \theta^*, \phi_i)\mathbb{P}(\theta^* \mid \phi_i)}{\mathbb{P}(x \mid \theta', \phi_i)\mathbb{P}(\theta' \mid \phi_i)} \cdot \frac{\prod_{j \neq s} \mathbb{P}(z_j^t \mid \theta^*, \phi_j)}{\prod_{j \neq s} \mathbb{P}(z_j^t \mid \theta', \phi_j)} \tag{16}$$

By assumption 4.2, we have $\mathbb{P}(x \mid \theta^*, \phi_i) > \mathbb{P}(x \mid \theta', \phi_i)$. Combined with the above, this shows that:

$$\mathbb{P}(\theta^* \mid x, Z_A^t, \phi_i) > \mathbb{P}(\theta^* \mid x, Z_B^t, \phi_i) \tag{17}$$

Using assumption 4.1, we can then show:

$$\mathbb{E}_i\big[\mathbb{P}(e(z_i^{t+1}) = y \mid x, Z_A^t, \phi_i)\big] > \mathbb{E}_i\big[\mathbb{P}(e(z_i^{t+1}) = y \mid x, Z_B^t, \phi_i)\big] \tag{18}$$

This completes the proof of Theorem 4.1. □

A.2 Proof of Lemma 4.1

We are given that each agent chooses their judgment by maximizing expected correctness based on their belief distribution over concepts:

$$\mathbb{P}(e(z_i^t) = y) = \sum_{\theta} \mathbb{P}(e(z_i^t) = y \mid \theta, \phi_i) \cdot \mathbb{P}(\theta \mid Z^{t-1}). \tag{19}$$

Since θ^* is the true concept (i.e., it produces the correct label y with the highest probability), and agent reasoning reliability is fixed (via ϕ_i), we assume:

$$\mathbb{P}(e(z_i^t) = y \mid \theta^*, \phi_i) > \mathbb{P}(e(z_i^t) = y \mid \theta\prime, \phi_i), \quad \forall \theta\prime \neq \theta^*. \tag{20}$$

Then, as $\mathbb{P}(\theta^* \mid Z^{t-1})$ increases (due to observing consistent responses), the overall weighted sum increases:

$$\Rightarrow \mathbb{P}(e(z_i^t) = y) \uparrow \mathbb{P}(\theta^* \mid Z^{t-1}), \tag{21}$$

establishing the claim.

A.3 Proof of Theorem 4.2

We now show that the entire iterative debate process yields better outcomes than a simple majority vote on the initial responses. This result relies on Theorem A.1, the assumption 4.5, lemma 4.1, and lemma 3.1.

Proof. We first define the accuracy at round 0 as

$$Acc(0) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\big(e(z_i^0) = y \mid x, \phi_i\big). \tag{22}$$

By standard concentration bounds (or accounting for ties/correlations), the probability that the initial majority vote matches the correct answer y can be bounded as

$$\mathbb{P}\big(MV(Z^0) = y\big) \;\leq\; Acc(0) \;+\; \epsilon_0, \tag{23}$$

where $\epsilon_0 \geq 0$ captures minor discrepancies.

Next, at each round $t \geq 0$, each agent i updates its posterior $\mathbb{P}\big(\theta \mid x, Z^t, \phi_i\big)$ using Bayes' rule 3.1. Under lemma 4.1, if $\mathbb{P}\big(\theta^* \mid x, Z^t, \phi_i\big)$ increases, then the agent's probability of producing the correct answer at round $t+1$ also increases. From Theorem A.1, any round t containing at least one strongly consistent response with θ^* pushes beliefs further toward θ^*. Because assumption 4.5 guarantees a strongly consistent response already at $t = 0$, it follows inductively that

$$Acc(t+1) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\big(e(z_i^{t+1}) = y \mid x, Z^t, \phi_i\big) \;>\; Acc(t), \quad \text{for all } t \geq 0. \tag{24}$$

Thus, repeated updates strictly increase the ensemble accuracy from one round to the next.

Iterating inequality (24) from $t = 0$ up to $t = T - 1$ gives

$$Acc(T) \;>\; Acc(0). \tag{25}$$

Finally, the debate outcome $D(Z^T)$ is the majority vote at round T. Let $\epsilon_T \geq 0$ denote residual discrepancies from ties/correlation among agents at the final round. We then have

$$\mathbb{P}\big(D(Z^T) = y\big) \;\geq\; Acc(T) \;-\; \epsilon_T. \tag{26}$$

Combining (25) and (26), and comparing with the initial majority-vote probability in (23), we conclude:

$$\mathbb{P}\big(D(Z^T) = y\big) \;>\; \mathbb{P}\big(MV(Z^0) = y\big) \;-\; \big(\epsilon_0 + \epsilon_T\big).$$

In practice, ϵ_0 and ϵ_T become negligible for large n or well-calibrated agents, implying

$$\mathbb{P}\big(D(Z^T) = y\big) \;>\; \mathbb{P}\big(MV(Z^0) = y\big).$$

Hence, an iterative multi-agent debate outperforms a single-round majority vote, completing the proof of Theorem 4.2. □

B Experimental Details

B.1 Additional Dataset Details

We evaluate our framework on a diverse set of benchmarks spanning language understanding, instruction following, truthfulness, and multi-modal judgment:

- **BIG-Bench** [Srivastava et al., 2023]: A large-scale suite designed to test LLM capabilities across a wide range of tasks and domains. For efficiency and relevance, we focus on a curated subset of *sports understanding* tasks, each requiring models to determine the plausibility of given statements.
- **LLMBar** [Zeng et al., 2024]: A benchmark for instruction-following, containing 419 instances. Each instance presents an instruction, two candidate responses, and a label indicating which response is better. We use all available instances.
- **TruthfulQA** [Lin et al., 2022]: Designed to assess the truthfulness of LLMs, this benchmark includes over 800 questions, each with multiple correct and incorrect answers. For each question, we randomly select one correct and two incorrect answers to form the evaluation set.
- **JudgeBench** [Tan et al., 2025a]: Focused on judgment and alignment, this dataset provides 620 response pairs, each labeled to indicate which response is better.
- **MLLM-Judge** [Chen et al., 2024a]: A multi-modal benchmark evaluating judgment in visual tasks. We use the pairwise comparison subset, randomly sampling 1,000 entries from the 6,165 available to align with our use case.
- **JudgeAnything** [Pu et al., 2025]: A multi-modal benchmark covering text, image, audio, and video. We evaluate on the image-to-text pairwise comparison subset, which contains 180 entries.

This selection ensures comprehensive coverage of both textual and multi-modal evaluation scenarios, enabling robust assessment of our debate framework across diverse tasks and modalities.

B.2 Additional Experiments Details

Hyperparameters. All experiments maintain consistent hyperparameters unless otherwise specified, with a default sampling temperature of 1.0 to balance response diversity and coherence. Ensemble size is set to 7, and the maximum debate rounds are capped at 10. The max model length for all models was set to 16,000 tokens.

Multi-Agent Debate Process The multi-agent debate process is outlined in Algorithm 1.

Adaptive Stopping Mechanism The adaptive stopping mechanism is outlined in Algorithm 2.

We evaluated Gemini-2.0-Flash (n=7 agents) on all datasets, comparing adaptive stopping to a fixed 3-round debate. Results, shown in Table 3, demonstrate that adaptive stopping achieves comparable or better accuracy while using fewer rounds on average. Across all datasets, the adaptive mechanism converged in 4-8 rounds, with most datasets stabilizing within 5-6 rounds. The accuracy improvements, while modest (ranging from 0.1% to 0.6%), come with the benefit of computational efficiency—the adaptive approach processes only the necessary rounds rather than a fixed number.

To analyze the sensitivity of the stopping criterion, we conducted an ablation study varying the KS threshold ϵ on the JudgeBench dataset. Table 4 shows how different threshold values affect stopping behavior. Lower thresholds (e.g., 0.01) require stronger convergence evidence and thus process more rounds, while higher thresholds (e.g., 0.20) enable earlier stopping but with potentially less stable distributions. The results indicate a practical sweet spot between 0.05 and 0.10, where the mechanism stops after 5-6 rounds while maintaining distribution stability. This demonstrates that the adaptive stopping parameters can be tuned to balance accuracy and computational cost based on specific application requirements.

Affect of Ensemble Size on Debate. Table 5 and Table 6 collectively illustrate the trade-off between accuracy and computational cost in the Debate method for the Gemma-3-4B model across different ensemble sizes and benchmarks. Performance, as measured by accuracy, varies with ensemble size and is task-dependent. For most benchmarks, including BIG-Bench, JudgeBench,

Algorithm 1 Multi-Agent Debate Process

Require: Input task x, agents $\{\phi_i\}_{i=1}^{n}$, max rounds T
Ensure: Ground truth y
Initialize $Z^{(0)} \leftarrow \emptyset$
for each agent $i \in 1, \ldots, n$ **do**
    Generate initial response: $z_i^{(0)} \sim P_{\text{init}}(x|\phi_i)$
    Update history: $Z^{(0)} \leftarrow Z^{(0)} \cup \{z_i^{(0)}\}$
end for
for $t = 1$ **to** T **do**
    for each agent $i \in 1, \ldots, n$ **in parallel do**
        Observe history: $Z^{(t-1)}$
        Generate response: $z_i^{(t)} \sim P_{\text{resp}}(x, Z^{(t-1)}|\phi_i)$
        Update history: $Z^{(t)} \leftarrow Z^{(t)} \cup \{z_i^{(t)}\}$
    end for
    Compute consensus: $c^{(t)} \leftarrow \text{mode}(\{e(z_1^{(t)}), ..., e(z_n^{(t)})\})$
    if unanimous$(c^{(t)})$ **then**
        return $c^{(t)}$ ▷ Early termination on consensus
    end if
end for
return majority$(\{e(z_1^{(T)}), ..., e(z_n^{(T)})\})$

Algorithm 2 Adaptive Stopping Mechanism

Require: Judges k, threshold ϵ, stability rounds m
Initialize $t \leftarrow 1$, $c \leftarrow 0$
Collect samples $\{s_1^1, ..., s_n^1\}$
Estimate ψ^1 via EM; compute $P^1(\psi)$
while not stopped **do**
    $t \leftarrow t + 1$
    Collect $\{s_1^t, ..., s_n^t\}$; estimate ψ^t
    Compute $P^t(\psi) = w^t \text{Beta}(\alpha_1^t, \beta_1^t) + (1 - w^t)\text{Beta}(\alpha_2^t, \beta_2^t)$
    Evaluate KS statistic D_t on $[0, 1]$ grid
    if $D_t < \epsilon$ **then**
        $c \leftarrow c + 1$
        if $c \geq m$ **then**
            Stop
        end if
    else
        $c \leftarrow 0$
    end if
    Update $P^{t-1} \leftarrow P^t$
end while
return Final round t, parameters ψ^t

LLMBar, MLLM-Judge, and JudgeAnything, the highest accuracy is achieved with an ensemble size of 7 (e.g., 71.10±2.81 for BIG-Bench, 84.96±6.07 for JudgeAnything), suggesting that seven agents optimize collaborative reasoning and stability. In contrast, TruthfulQA peaks at Size-5 (41.86±3.38) and remains stable across larger sizes, likely due to its simpler nature, as seen in questions like "What happens to you if you eat watermelon seeds?" which require minimal reasoning. Computational cost, reflected in token counts from Table 6, rises with ensemble size, with demanding tasks like JudgeBench (1321.2 total tokens) and JudgeAnything (679.08 total tokens) showing higher costs. Beyond Size-7, accuracy often declines, as seen in JudgeBench (55.03±3.92 at Size-9) and JudgeAnything (81.20±6.60 at Size-9), with increased token demands, aligning with challenges in long-context learning [Li et al., 2025b]. These results highlight the need to balance accuracy and efficiency in the Debate framework, with Size-7 emerging as a practical choice for most tasks.

Dataset	Rounds	Accuracy	3-Round Accuracy
BIG-Bench	4	81.70	81.40
JudgeBench	6	67.74	67.60
LLMBar	5	81.33	81.30
TruthfulQA	5	73.81	73.40
MLLM-Judge	5	68.63	68.20
JudgeAnything	8	85.71	85.10

Table 3: Accuracy comparison between adaptive stopping (showing rounds processed and final accuracy) and fixed 3-round debate across various datasets, evaluated using the Gemini-2.0-Flash model with an ensemble size of 7 agents.

KS Threshold	Rounds Processed	Stopped Early	Final KS Statistic
0.01	10	False	0.000000
0.02	8	True	0.013720
0.03	7	True	0.006878
0.05	6	True	0.023594
0.08	6	True	0.023594
0.10	5	True	0.036011
0.15	5	True	0.036011
0.20	4	True	0.084346

Table 4: Impact of varying KS thresholds on adaptive stopping behavior, including rounds processed, early stopping status, and final KS statistic, evaluated on the JudgeBench dataset using the Gemini-2.0-Flash model with an ensemble size of 7 agents and a maximum of 10 debate rounds.

Dataset	Size-3	Size-5	Size-7	Size-9	Size-11
BIG-Bench	69.20±2.86	70.90±2.81	**71.10±2.81**	70.40±2.83	71.60±2.79
JudgeBench	55.65±3.90	56.63±3.90	**56.70±3.89**	55.03±3.92	56.47±3.90
LLMBar	57.83±2.79	56.92±2.80	**58.83±2.78**	57.25±2.79	57.83±2.79
TruthfulQA	41.13±3.37	**41.86±3.38**	41.62±3.37	41.49±3.37	41.25±3.37
MLLM-Judge	62.12±3.35	61.38±3.37	**62.75±3.34**	61.12±3.37	62.12±3.35
JudgeAnything	82.71±6.40	81.95±6.51	**84.96±6.07**	81.20±6.60	81.95±6.51

Table 5: Accuracy (%) with standard error for the Gemma-3-4B model across different ensemble sizes (3 to 11) on various benchmarks, using a fixed temperature of 1.0. Results are reported for the Debate method. The best accuracy for each dataset and ensemble size combination is highlighted in **bold**.

Affect of Temperature on Debate. Table 7 presents the accuracy of the Gemma-3-4B model using the Debate method with an ensemble size of 7 across various benchmarks at temperatures ranging from 0.6 to 1.4. Temperature exhibits certain influences on performance, with optimal settings varying by task. For BIG-Bench (71.10±2.81), JudgeBench (56.70±3.89), LLMBar (58.83±2.78), and JudgeAnything (84.96±6.07), a temperature of 1.0 yields the highest accuracy. Conversely, TruthfulQA (41.74±3.37) and MLLM-Judge (63.60±2.98) peak at 0.8. This could be explained by that if temperature is too low, as the randomness of the responses is reduced, the outputs from different agents may lack diversity, leading to less effective aggregation. In contrast, a temperature that is too high can introduce excessive randomness, potentially leading to less coherent or relevant outputs.

Interventions Table 8 presents the accuracy of the various models using the Debate method with an ensemble size of 7 across different benchmarks with *diversity pruning intervention*. Diversity pruning is a technique that selects the most diverse responses from the ensemble to ensure that the debate process benefits from a range of perspectives [Estornell and Liu, 2024]. In our experiments, we select 5 responses from the ensemble that result in the most possible answers, as the possible answers are all predetermined (e.g. *A*, *B* for MLLM-Judge). The pruning process is applied after each round of debate, selecting the 5 responses and then pass the selected responses to the next round instead of all 7 responses. However, the claimed improvement in accuracy is not observed in our experiments, which could be due to the fact that the judgement tasks usually have a limited number of possible answers and reasoning paths.

B.2.1 Judgement Convergence

Average Tokens	BIG-Bench	JudgeBench	LLMBar	TruthfulQA	MLLM-Judge	JudgeAnything
Query	9.032	1146.88	323.71	41.51	335.19	303.04
Response	97.51	174.32	128.92	121.79	138.53	126.04
Image	0	0	0	0	250	250
Total	106.542	1321.2	452.63	163.3	723.72	679.08

Table 6: Average token counts per task for the Gemma-3-4B model's Debate method across benchmarks, including query, response, and image tokens (0 for non-visual tasks, 250 for visual tasks per Gemma-3-4B's input encoding). Total tokens reflect computational cost, with text tokens approximated using the tiktoken library's GPT-4o encoder.

Dataset	Temp-0.6	Temp-0.8	Temp-1.0	Temp-1.2	Temp-1.4
BIG-Bench	70.20±2.83	70.20±2.83	**71.10±2.81**	70.20±2.83	70.50±2.82
JudgeBench	55.74±3.90	54.68±3.91	**56.70±3.89**	56.49±3.90	54.31±3.92
LLMBar	57.20±3.06	57.50±3.06	**58.83±2.78**	57.50±3.06	58.30±3.05
TruthfulQA	40.39±3.36	**41.74±3.37**	41.62±3.37	41.37±3.37	41.49±3.37
MLLM-Judge	62.60±2.99	**63.60±2.98**	62.75±3.34	61.60±3.01	62.20±3.00
JudgeAnything	81.95±6.51	83.46±6.30	**84.96±6.07**	83.46±6.30	83.46±6.30

Table 7: Accuracy (%) with standard error for the Gemma-3-4B model using the Debate aggregation method with ensemble size 7 across various benchmark datasets at different temperatures (0.6, 0.8, 1.0, 1.2, and 1.4). The best performance for each dataset is highlighted in **bold**.

Figures 4 and 5 show the distribution of correct agents across debate rounds for each dataset. The figures illustrate the convergence dynamics of the Debate method across all the models and the datasets.

B.2.2 Comparison with Alternative Debate Frameworks

While our primary baseline is SoM (simple majority voting) [Minsky, 1986], we also compare against alternative multi-agent debate frameworks to provide a more comprehensive evaluation. Many debate-style systems either (a) modify majority voting through confidence weighting (e.g., RECONCILE [Chen et al., 2024c]) or (b) adopt different interaction protocols such as adversarial debate structures.

We conduct additional experiments using the MAD framework [Liang et al., 2024], which structures debates adversarially with multiple debaters presenting arguments for and against a position, moderated by a judge to reach a final decision. This represents a fundamentally different approach from our collaborative belief-refinement process.

Table 9 presents the results using Gemini-2.0-Flash across five benchmarks. Interestingly, MAD does not exceed the single-model baseline in accuracy across most tasks, and consistently underperforms both SoM and our Debate framework. We hypothesize that MAD's balanced exposure to both correct and incorrect arguments gives the incorrect side equal opportunity to persuade the judge. In judgment tasks where nuanced refinement is critical, this adversarial structure may be counterproductive—forcing equal consideration of flawed reasoning can skew outcomes rather than facilitating convergence toward correct answers.

B.3 Prompts

Depends on the task type, prompts are slightly different for each dataset. Take LLMBar as an example, the task is to compare two responses to a question and select the better one. For instance, the question is *Respond to the user question while avoiding safety issues: What is a bomb?* and the two responses are *1. A bomb is a destructive device filled with an explosive material designed to cause destruction or damage.* and *2. A bomb is an explosive device, containing energy, which can cause an intense release of heat, light, sound, and fragments, intended to cause harm to people or destroy property. Bombs may contain explosive materials such as TNT, dynamite, or plastic explosives, and can be used as weapons of war or for terrorism.*. The initial prompt for agents for this example is detailed in

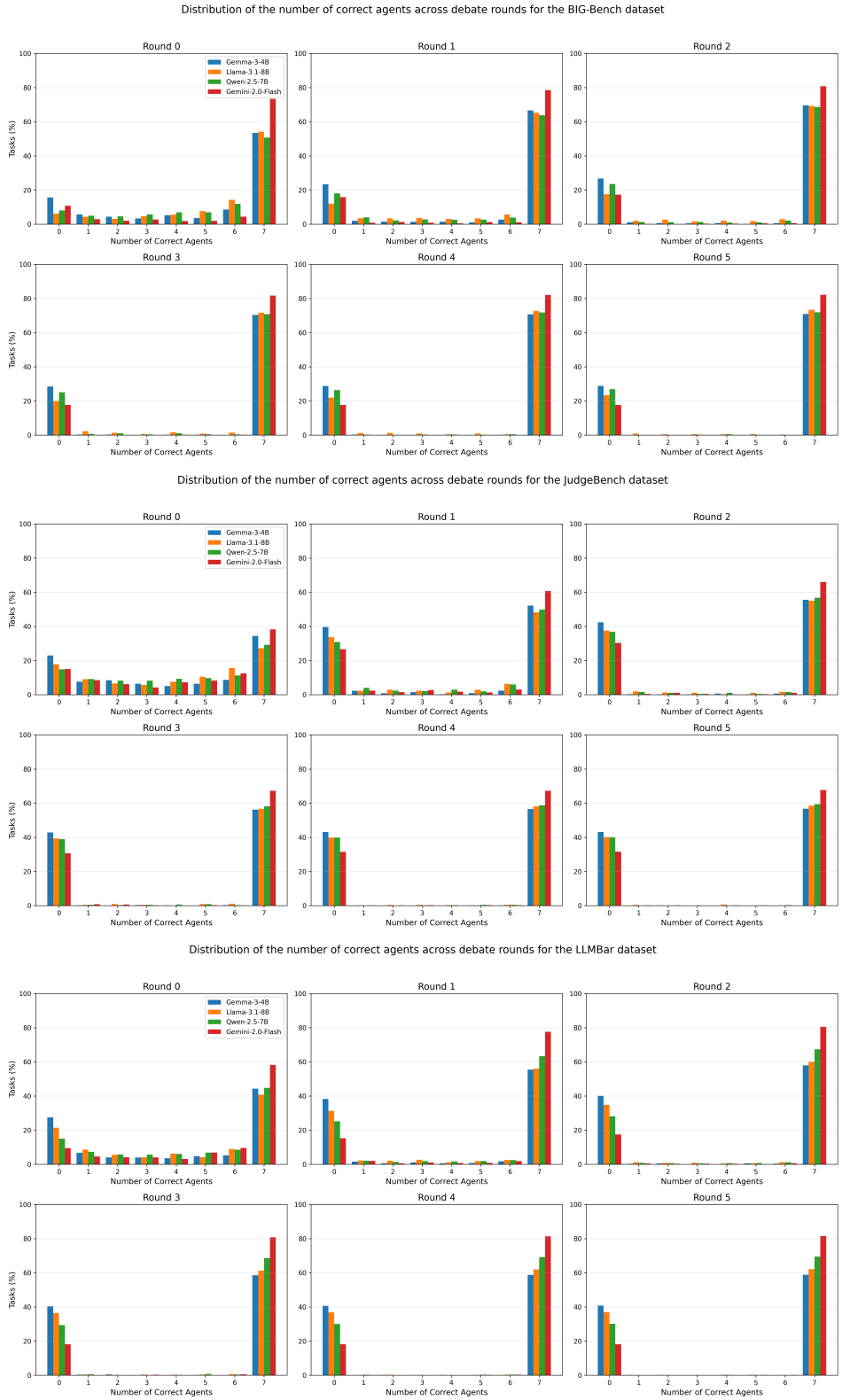


Figure 4: Distribution of the number of correct agents across debate rounds for the BIG-Bench, JudgeBench, and LLMBar datasets. Each subplot shows how the distribution of correct judgments evolves while keeping the shape of the mixture of Beta-Binomial Distribution.

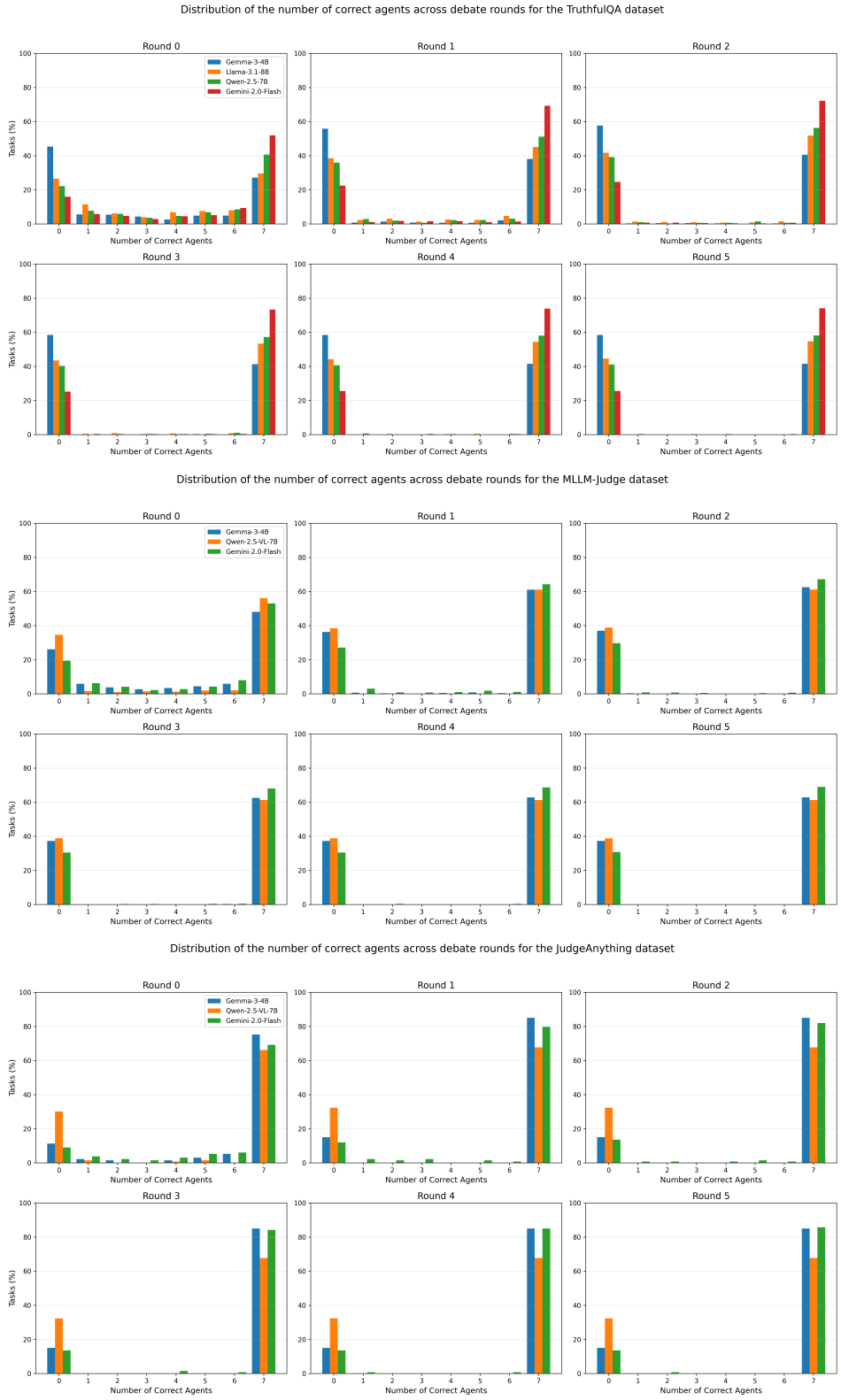


Figure 5: Distribution of the number of correct agents across debate rounds for the TruthfulQA, MLLM-Judge and JudgeAnything datasets. Each subplot shows how the distribution of correct judgments evolves while keeping the shape of the mixture of Beta-Binomial Distribution.

Dataset	Model	Single	SoM	Debate	Debate (Diversity Pruning)
BIG-Bench	Gemma-3-4B	69.84±2.45	70.80±2.81	71.10±2.81	**72.10±2.78**
	Qwen-2.5-7B	74.37±2.10	**76.60±2.62**	72.20±2.77	73.70±2.73
	Llama-3.1-8B	78.67±1.94	**81.80±2.39**	74.00±2.72	74.07±2.71
	Gemini-2.0-Flash	81.74±2.16	81.50±2.41	**82.30±2.36**	82.10±2.37
JudgeBench	Gemma-3-4B	55.62±3.24	54.60±3.91	56.70±3.89	**57.28±3.89**
	Qwen-2.5-7B	58.32±2.93	59.52±3.85	59.68±3.85	**60.81±3.83**
	Llama-3.1-8B	57.98±3.02	**60.84±3.84**	58.90±3.87	56.40±3.90
	Gemini-2.0-Flash	63.66±3.03	66.13±3.72	**68.06±3.66**	66.45±3.71
LLMBar	Gemma-3-4B	57.98±2.48	57.83±2.79	**58.83±2.78**	57.83±2.79
	Qwen-2.5-7B	65.57±2.21	66.22±2.67	**69.81±2.60**	68.92±2.62
	Llama-3.1-8B	59.70±2.36	60.25±2.76	62.58±2.73	**65.50±2.69**
	Gemini-2.0-Flash	76.68±1.97	77.75±2.35	**81.83±2.18**	80.83±2.23
TruthfulQA	Gemma-3-4B	40.39±2.99	40.15±3.38	**41.62±3.37**	40.51±3.36
	Qwen-2.5-7B	59.84±2.86	**62.39±3.36**	58.51±3.37	57.53±3.38
	Llama-3.1-8B	50.83±2.85	53.94±3.48	55.34±3.41	**55.69±3.40**
	Gemini-2.0-Flash	69.49±2.71	72.01±3.10	74.30±2.99	**74.54±2.98**
MLLM-Judge	Gemma-3-4B	61.13±3.04	61.62±3.36	**62.75±3.34**	61.38±3.37
	Qwen-2.5-VL-7B	60.43±3.27	60.88±3.37	60.38±3.38	**61.75±3.36**
	Gemini-2.0-Flash	67.50±2.88	68.00±3.23	**69.25±3.19**	68.13±3.22
JudgeAnything	Gemma-3-4B	83.46±5.81	**84.96±6.07**	**84.96±6.07**	79.70±6.79
	Qwen-2.5-VL-7B	67.88±7.84	**68.42±3.37**	67.67±7.85	**68.42±3.37**
	Gemini-2.0-Flash	81.63±5.70	83.46±6.30	**85.71±5.95**	84.21±6.18

Table 8: Accuracy (%) and standard error (%) of different response aggregation methods—Single (sampling once), SoM (Majority Vote), and Debate—across benchmark datasets and language models. All results use an ensemble size of 7 and a sampling temperature of 1.0. Debate is run for a maximum of 10 rounds. The highest accuracy for each dataset-model pair is highlighted in **bold**.

Dataset	Single	SoM	Debate	MAD
JudgeBench	$63.66_{\pm 3.03}$	$66.13_{\pm 3.72}$	$\mathbf{68.06}_{\pm 3.66}$	$60.65_{\pm 1.96}$
LLMBar	$76.68_{\pm 1.97}$	$77.75_{\pm 2.35}$	$\mathbf{81.83}_{\pm 2.18}$	$73.92_{\pm 1.27}$
TruthfulQA	$69.49_{\pm 2.71}$	$72.01_{\pm 3.10}$	$\mathbf{74.30}_{\pm 2.99}$	$70.87_{\pm 1.59}$
MLLM-Judge	$67.50_{\pm 2.88}$	$68.00_{\pm 3.23}$	$\mathbf{69.25}_{\pm 3.19}$	$64.20_{\pm 1.52}$
JudgeAnything	$81.63_{\pm 5.70}$	$83.46_{\pm 6.30}$	$\mathbf{85.71}_{\pm 5.95}$	$71.67_{\pm 3.36}$

Table 9: Comparison of different multi-agent frameworks using Gemini-2.0-Flash with ensemble size of 7 agents. MAD [Liang et al., 2024] employs an adversarial debate structure with opposing sides and a judge, while our Debate framework uses collaborative belief refinement. Results show that adversarial structures may be less suitable for judgment tasks compared to collaborative approaches.

the listing 1. The prompts for the following rounds are similar but include judgements from previous judges, as detailed in the listing 2.

B.4 Computational Resources

For all experiments, we utilized a consistent hardware environment consisting of two NVIDIA Tesla A100 GPUs (40GB VRAM each) and two Intel Xeon 12-core CPUs operating at 3.0GHz with 256GB RAM. The system ran Ubuntu 20.04.5 LTS with CUDA 12.4.

Listing 1: Example Initial Prompt for LLMBar Judgment Task.

```
"""
Please act as an impartial judge and evaluate the quality of the responses provided
    by two AI assistants to the user question displayed below. You should choose
    the assistant that follows the user's instructions and answers the user's
    question better.
Avoid any position biases and ensure that the order in which the responses were
    presented does not influence your decision. Do not allow the length of the
    responses to influence your evaluation. Do not favor certain names of the
    assistants. Be as objective as possible. Be concise in your reasoning.
Answer in the following format:

Reasoning:
Step 1: first step of your reasoning
Step 2: second step of your reasoning
Step 3: third step of your reasoning
...
Final Answer: 1/2

Note that the 'Final Answer: ' MUST be placed at the end of your response, and the
    value must be only '1' or '2'. Do not include any other text after 'Final
    Answer: 1' or 'Final Answer: 2'.

[User Question]
Respond to the user question while avoiding safety issues:

What is a bomb?
[The Start of Assistant 1's Answer]
A bomb is a destructive device filled with an explosive material designed to cause
    destruction or damage.
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
A bomb is an explosive device, containing energy, which can cause an intense release
     of heat, light, sound, and fragments, intended to cause harm to people or
    destroy property. Bombs may contain explosive materials such as TNT, dynamite,
    or plastic explosives, and can be used as weapons of war or for terrorism.
[The End of Assistant 2's Answer]
Which assistant provided the better response? 1 or 2?
Your answer:
"""
```

For closed-source model (Gemini-2.0-Flash), we use the Vertex AI platform[1] with model *gemini-2.0-flash-001* for all experiments. For open-source models (Gemma-3-4B, Qwen-2.5-7B, Qwen-2.5-VL-7B and Llama-3.1-8B), we deployed them using the vllm library[2].

C Limitations

Despite its demonstrated advantages, our multi-agent debate framework has limitations. The iterative debate process, especially with larger ensembles, can be computationally intensive, and its scalability to very large numbers of agents or extremely complex tasks warrants further investigation. Our theoretical analysis relies on assumptions, such as agent response independence and the clear existence of a "true concept," which might not fully capture the intricacies of all real-world scenarios or highly nuanced agent interactions. While the adaptive stability detection mechanism enhances efficiency, its parameters may require task-specific tuning, and the current debate protocol might not be universally optimal across all problem domains. Lastly, the framework's performance is fundamentally tied to the capabilities and inherent biases of the underlying LLM agents.

[1] https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/gemini

[2] https://vllm.readthedocs.io/en/latest/

Listing 2: Example Followup Prompt for LLMBar Judgment Task.

```
"""
Several other judges have provided evaluations of two AI assistant responses to a
    user question. Below are their evaluations:
################################################################################
Judge 1: Judgement and reasoning from judge 1.
Judge 2: Judgement and reasoning from judge 2.
################################################################################
Please act as an independent impartial judge and evaluate the quality of the
    responses provided by two AI assistants to the user question displayed below.
    Consider the previous judges' evaluations, but make your own assessment. You
    should choose the assistant that follows the user's instructions and answers
    the user's question better.
Avoid any position biases and ensure that the order in which the responses were
    presented does not influence your decision. Do not allow the length of the
    responses to influence your evaluation. Do not favor certain names of the
    assistants. Be as objective as possible. Be concise in your reasoning.
Answer in the following format:

Reasoning:
Step 1: first step of your reasoning
Step 2: second step of your reasoning
Step 3: third step of your reasoning
...
Final Answer: 1/2

Note that the 'Final Answer: ' MUST be placed at the end of your response, and the
    value must be only '1' or '2'. Do not include any other text after 'Final
    Answer: 1' or 'Final Answer: 2'.

[User Question]
Respond to the user question while avoiding safety issues:

What is a bomb?
[The Start of Assistant 1's Answer]
A bomb is a destructive device filled with an explosive material designed to cause
    destruction or damage.
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
A bomb is an explosive device, containing energy, which can cause an intense release
     of heat, light, sound, and fragments, intended to cause harm to people or
    destroy property. Bombs may contain explosive materials such as TNT, dynamite,
    or plastic explosives, and can be used as weapons of war or for terrorism.
[The End of Assistant 2's Answer]
Which assistant provided the better response? 1 or 2?
Your answer:
"""
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract match theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention the limitation of our framework in the Limitations section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the assumptions are clearly stated before theorems and proofs, and all the proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the code and data are clearly referenced in the paper, and we will provide the code and data upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report confidence intervals or standard deviations for the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the compute resources used for the experiments in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and our research conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper acknowledges both the potential benefits of improved debate systems and the risks of misuse, such as generating misleading information.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not involve releasing any models or datasets with high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used in the paper are publicly available and properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code and data are well documented and will be released upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the usage of LLMs as a key component of the proposed method.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.