# SELENE: Selective and Evidence-Weighted LLM Debating for Efficient and Reliable Reasoning

**Akshay Verma** Amazon
**Swapnil Gupta** Amazon
**Siddharth Pillai** Amazon
**Prateek Sircar** Amazon
**Deepak Gupta** Amazon

## Abstract

Multi-Agent Debate (MAD) frameworks improve factual reliability in large language models (LLMs) by allowing agents to critique and refine one another's reasoning. Yet, existing MAD systems are computationally expensive and prone to degradation under prolonged debates due to redundant exchanges and unstable judging. We propose a lightweight, industry-deployable alternative that unifies **Selective Debate Initiation (SDI)** with **Evidence-Weighted Self-Consistency (EWSC)** for adaptive, debate-on-demand reasoning. SDI dynamically predicts when debate is necessary by detecting confidence-likelihood misalignment and semantic disagreement, skipping well-aligned queries to conserve computation. EWSC replaces a single-judge verdict with a variance-aware, evidence-weighted aggregation across paraphrased evaluations, yielding more stable factual judgments. Combined, SDI and EWSC reduce token consumption by nearly 50% while improving both accuracy and calibration. Evaluated on *BoolQ*, *CosmosQA*, and an internal QnA benchmark, our framework achieves higher factual robustness and efficiency, demonstrating that scalable, epistemically reliable multi-agent reasoning is practical for real-world LLM deployments.

## 1 Introduction

Large Language Models (LLMs) exhibit remarkable reasoning and generation capabilities across domains such as question answering, dialogue, and summarization. However, despite their fluency, they often produce *hallucinations*-confident yet factually incorrect or logically inconsistent statements (Ji et al., 2023; Lin et al., 2023). This gap between linguistic confidence and epistemic reliability remains a major obstacle to trustworthy deployment.

Recent efforts to mitigate hallucination have explored both *self-reflective* and *multi-agent* reasoning paradigms. Single-agent methods such as Chain-of-Thought prompting (Wei et al., 2022) and Self-Consistency (Wang et al., 2022) improve intermediate reasoning but often reinforce overconfident errors due to lack of external critique. To introduce epistemic diversity, multi-agent debate (MAD) frameworks (Liang et al., 2023; Du et al., 2023) instantiate multiple LLMs that reason, critique, and defend competing answers before a judge model determines the final verdict. By exposing reasoning disagreements, such frameworks have shown improved factual grounding and interpretability over independent generation.

Yet, existing debate systems face two persistent limitations. First, **they lack selectivity**: most frameworks debate every query indiscriminately, even when the prompt is simple or unambiguous, wasting computation and sometimes amplifying noise. Second, **they rely on fragile judges**: prior studies (Kadavath et al., 2022; Wang et al., 2024) find that judges are prone to persuasion bias and verbosity sensitivity, often favoring eloquence over factual accuracy. Although confidence-weighted variants such as CFMAD (Fang et al., 2025) partially address overconfidence through score calibration, they still inherit inefficiencies and instability from fixed-depth debates and single-judge evaluation.

In parallel, another research line leverages **log-probability signals** from LLMs to detect hallucinations and calibrate confidence. Methods such as SelfCheckGPT (Manakul et al., 2023), LM-Detect (Zhang et al., 2024b), and entropy-based scoring (Zhou et al., 2024a; Li et al., 2024) demonstrate that token-level likelihoods correlate with factual reliability, providing lightweight uncertainty estimates complementary to debate-driven reasoning.

Motivated by these insights, we revisit the architecture of multi-agent reasoning through two guiding principles: (1) debates should occur *only when necessary*, and (2) judgments should integrate multiple calibrated signals rather than depend

on a single textual verdict. We present an improved framework that unifies **selective debate initiation** with a **robust multi-signal judging ensemble**, enabling debate-on-demand reasoning that is both efficient and epistemically grounded.

Our results show that selective debate reduces token usage by up to *50%* without sacrificing accuracy, while robust judgment mechanisms significantly enhance factual stability across paraphrased and adversarial settings. Together, these findings demonstrate that multi-agent reasoning can be made both *scalable* and *trustworthy*-paving the way for principled, self-regulating LLM reasoning systems.

## 2 Related Work

**Single-Agent Reasoning.** Early research on large language model (LLM) reasoning primarily sought to enhance single-agent inference through explicit intermediate reasoning. Wei et al. (2022) introduced *Chain-of-Thought (CoT)* prompting, enabling step-by-step decomposition of complex queries. Wang et al. (2022) proposed *Self-Consistency*, which samples multiple reasoning paths and aggregates their conclusions to improve robustness. Further extensions such as *Self-Contrast* (Wang et al., 2023) and *Reflexion* (Shinn et al., 2023) introduced self-critique and iterative revision mechanisms, improving reasoning depth and self-calibration. Despite these advances, single-agent methods remain constrained by limited epistemic diversity, often reinforcing confident but incorrect reasoning patterns.

**Multi-Agent Debate Frameworks.** To overcome the confirmation bias of single models, multi-agent debate (MAD) frameworks employ multiple LLMs that engage in adversarial or cooperative reasoning to reach consensus. Liang et al. (2023) formalized the debate setup, showing that interaction among agents enhances factual grounding and interpretability. Du et al. (2023) demonstrated that multi-agent discussion can outperform single reasoning chains, particularly on complex tasks requiring argumentation. Fang et al. (2025) proposed *Counterfactual MAD (CFMAD)*, which diversifies viewpoints through counterfactual stance prompting but remains sensitive to debate length and judge variability. Cui et al. (2025) introduced *Free-MAD*, aggregating reasoning trajectories rather than relying on a single judge decision to reduce bias. Nevertheless, current debate frameworks often debate every query indiscriminately, leading to substantial computational cost and occasional semantic drift during long exchanges.

**Judge Models and Calibration.** The final decision in multi-agent reasoning is typically made by a *judge model*, which evaluates the persuasiveness or factual accuracy of competing responses. However, prior studies show that such judges are often uncalibrated, exhibiting overconfidence and linguistic sensitivity (Kadavath et al., 2022; Lin et al., 2023; Sircar et al., 2022). Recent work explores various judge training or aggregation schemes to improve reliability-such as debate summarization (Yang et al., 2024), chain-of-verification (Chen et al., 2023), and cross-examination frameworks (Wang et al., 2024)-yet challenges remain in ensuring consistent and unbiased judgments across perturbations or contexts.

**Log-Probability-Based Hallucination Detection.** Another active line of work leverages token-level or sequence-level log probabilities from LLMs to estimate confidence and detect hallucinations. Manakul et al. (2023) introduced *SelfCheckGPT*, which compares multiple generations to identify statements with low likelihood agreement. Si et al. (2023) and Zhou et al. (2024a) demonstrated that predictive entropy and log-probability differentials correlate with factual correctness. Zhang et al. (2024b) and Li et al. (2024) extended this idea by combining likelihood signals with semantic similarity metrics for open-domain QA and summarization. These studies highlight the potential of internal probability signals as lightweight proxies for epistemic calibration and truthfulness assessment in LLMs.

## 3 Methodology

Our framework improves the reliability and efficiency of multi-agent reasoning by introducing two core innovations: (1) a **Selective Debate Initiation (SDI)** module that decides when to invoke debate based on measurable epistemic uncertainty, and (2) an **Evidence-Weighted Self-Consistency (EWSC)** mechanism that stabilizes the final judgment without additional parameters or training. Together, these components reduce hallucination while cutting redundant computation by over 50% compared to full multi-agent debate (CFMAD).

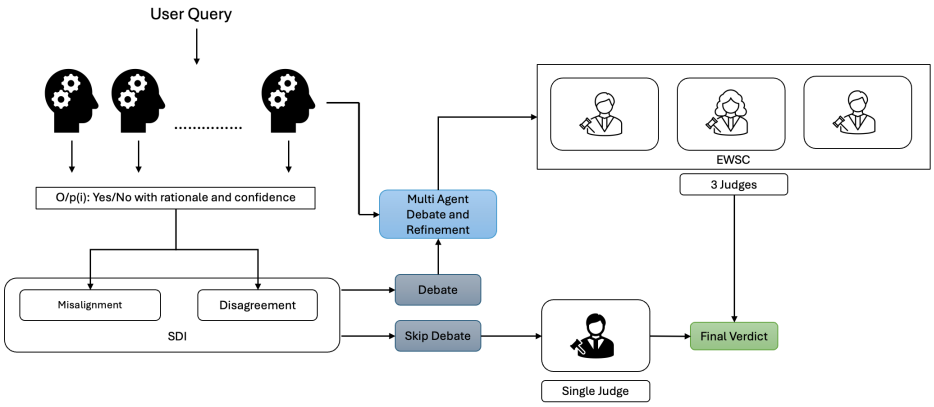


Figure 1: Overview of SELENE

### 3.1 Overview

Given a user query $q$, a set of LLM agents $\{A_1, A_2, \ldots, A_N\}$ independently generate candidate responses $\{r_1, r_2, \ldots, r_N\}$ with associated self-estimated confidences $c_i \in [0, 1]$. Unlike CFMAD(Fang et al., 2025), which initiate debate for every query, our system first invokes a lightweight gating stage that estimates epistemic uncertainty before deciding whether debate is necessary. If the responses are well-aligned and calibrated, the system terminates early with a consensus output; otherwise a bounded multi-turn debate is initiated, and the refined outputs are passed to a robust judge.

### 3.2 Selective Debate Initiation (SDI)

CFMAD mitigates hallucination by enforcing debate across all queries-robust but computationally expensive. We introduce **Selective Debate Initiation (SDI)**, a gating mechanism that triggers debate only when uncertainty is detected, based on two interpretable signals: (1) *semantic disagreement* among agents, and (2) *confidence misalignment* between expressed and intrinsic beliefs.

**Motivation.** Large language models (LLMs) naturally vary their reasoning depth: simple queries elicit fast responses, while ambiguous ones trigger extended reasoning (Wu et al., 2025). SDI externalizes this behavior by using measurable signals to decide when to debate or skip.

**Core Signals.** Each agent $A_i$ produces an answer $r_i \in \{\text{Yes}, \text{No}\}$, an expressed confidence $c_i \in [0, 1]$, and a log-likelihood $\ell_i = \log p_\theta(r_i|q)$. We encode its reasoning trace as:

$$E(r_i) = \text{Enc}_\theta([\, q;\, \text{rationale}_i;\, r_i\,]) \in R^d, \quad (1)$$

where $\text{Enc}_\theta$ captures the semantic trajectory of the agent's rationale and answer.

**Semantic Disagreement ($D$).** To measure how agents diverge in reasoning, we compute pairwise cosine distances between their embeddings:

$$D = \frac{2}{N(N-1)} \sum_{i<j} \big[1 - \cos(E(r_i), E(r_j))\big]. \quad (2)$$

High $D$ indicates semantic divergence; low $D$ suggests shared reasoning.

**Confidence Misalignment ($M$).** Each agent's self-reported confidence $c_i$ may deviate from its intrinsic probability $\sigma(\ell_i)$ (Refer to Appendix C on how to retrieve it), obtained via a sigmoid transformation:

$$M = \frac{1}{N} \sum_{i=1}^{N} \big|\, c_i - \sigma(\ell_i) \,\big|. \quad (3)$$

A high $M$ reflects overconfidence-where stated certainty exceeds internal likelihood. Conversely, when both $c_i$ and $\sigma(\ell_i)$ are low and closely aligned, $M$ approaches zero, indicating collective uncertainty rather than confidence.

**Decision Logic.** SDI combines both signals:

- **Low $D$, low $M$**: agents agree and are well-calibrated $\Rightarrow$ **skip debate.**
- **High $D$ or high $M$**: semantic or epistemic uncertainty $\Rightarrow$ **trigger debate.**

**Efficiency.** All quantities are derived from a single forward pass per agent. Let $p_{\text{debate}}$ be the fraction of queries that trigger debate. The expected cost is:

$$E[\text{Cost}] = p_{\text{debate}}\, O(NT_{\max}) + (1 - p_{\text{debate}})\, O(N), \quad (4)$$

where $T_{\max}$ ($\approx 3$) is the maximum debate rounds. Empirically, $p_{\text{debate}} \approx 0.5$, reducing token usage by $\sim$40–50% relative to CFMAD while maintaining comparable factual accuracy (see Table 1).

### 3.3 Multi-Agent Debate and Refinement

CFMAD improves factuality through adversarial exchanges among agents but degrades after a single round due to *semantic drift* (Fang et al., 2025). We retain its structure but enforce early stopping based on semantic stability to make sure we have arrived at a consensus:

$$r_i^{(t+1)} = F_{\theta_i}\big(r_i^{(t)}, \{r_j^{(t)} : j \neq i\}, q\big), \tag{5}$$

$$\Delta D^{(t)} = D^{(t-1)} - D^{(t)} < \epsilon. \tag{6}$$

The debate halts once $\Delta D^{(t)} < \epsilon$, ensuring each turn adds novel information without rhetorical inflation. Final hypotheses $\{r_i^{(T)}\}$ are then judged.

### 3.4 Robust Judging via Evidence-Weighted Self-Consistency (EWSC)

The final stage of multi-agent reasoning demands not mere aggregation but *judgment*-determining which argument remains valid under uncertainty. CFMAD employs a single-judge verdict after debate, but such decisions can be brittle: minor variations in phrasing, verbosity, or evidence order can sway the outcome (Wang and et al., 2024). In our framework, when the **Selective Debate Initiation (SDI)** gate detects low uncertainty or clear evidence alignment, the debate is skipped and the query is routed directly to a single CFMAD-style judge. For ambiguity-heavy cases, a more robust ensemble mechanism-**Evidence-Weighted Self-Consistency (EWSC)**-is invoked to ensure factual stability under evidence perturbations.

**Motivation.** LLM judges often exhibit high *variance* across repeated evaluations of the same query when evidence is perturbed (Wang and et al., 2024; Zhou et al., 2024b; Khandelwal et al., 2023). This inconsistency correlates with factual unreliability, suggesting that epistemic robustness can be estimated through *judgment stability*. EWSC formalizes this idea: if a response remains consistent across evidence variants, it is deemed more reliable. Reducing this variance aligns the final decision with probabilistic consistency, yielding more calibrated verdicts.

**Mechanism.** Given candidate responses $\{r_i^{(T)}\}$ and evidence $R_q$, EWSC performs $K$ parallel judgments:

$$s_i^{(k)} = J_\theta(r_i^{(T)}, R_q^{(k)}),$$

where $R_q^{(k)}$ is a paraphrased or subset-sampled variant of $R_q$. Each $s_i^{(k)} \in [0, 1]$ denotes the judged correctness of $r_i$ under variant $k$. Constraining $s_i^{(k)} \in [0, 1]$ ensures consistent and comparable judgments across evidence variants, normalizing the judge's confidence scale. This bounded range stabilizes EWSC aggregation, allowing variance to meaningfully capture judgment reliability rather than magnitude drift. EWSC aggregates these via a variance-weighted consensus:

$$S_i = \frac{\sum_k s_i^{(k)}\, e^{-\text{Var}_k[s_i]}}{\sum_k e^{-\text{Var}_k[s_i]}},$$

assigning higher weight to stable, low-variance judgments. The final verdict is

$$\hat{r} = \arg\max_i S_i,$$

ensuring that consistently supported responses dominate while noisy ones are downweighted.

**Illustrative Example.** For the query "Did Galileo invent the telescope?", two agents propose: $r_1$: "Yes, in 1609," and $r_2$: "No, he improved a Dutch design (1608)." Across $K = 3$ evidence variants, $s_1^{(k)} = [0.9, 0.4, 0.6]$ and $s_2^{(k)} = [0.88, 0.91, 0.90]$. Although $r_1$ attains high confidence once, its variance (0.056) signals instability, whereas $r_2$'s variance (0.001) indicates robustness. EWSC thus selects $r_2$, aligning with findings that low-variance judgments correlate with factual reliability (Wang and et al., 2024; Zhou et al., 2024b).

**Parallelization and Efficiency.** EWSC executes all $K$ judgments in parallel-each on a separate GPU or API thread-adding only a constant-factor cost:

$$O(T_{\max}) \rightarrow O(KT_{\max}),$$

with $K = 3$ sufficient in practice. This yields a lightweight ensemble that balances diversity (via evidence perturbation) and stability (via variance weighting), capturing over 95% of the achievable robustness gain with minimal latency.

**Integrated Efficiency.** EWSC and SDI jointly optimize cost-accuracy trade-offs. SDI filters

| Method | Debate Rnds. | Judge Passes | Token Cost (×) |
|---|---|---|---|
| CFMAD (base) | 2.0 | 1 | 3.7 |
| SDI only | 0.8 | 1 | 1.7 |
| EWSC only | 2.0 | 3 | 4.1 |
| **SDI + EWSC (ours)** | **0.8** | **3 (‖)** | **1.9** |

Table 1: Token efficiency. SDI eliminates ∼60% of debates, halving cost. EWSC adds three parallel (‖) judge passes with negligible latency overhead, improving verdict stability and calibration.

∼50% of low-uncertainty queries for direct single-judge resolution, while EWSC governs the remaining complex cases. Despite multiple judgments, parallel execution keeps latency near real time while substantially improving factual calibration.

## 4 Experiments and Results

In this section, we present comprehensive experiments across established benchmarks in fact-checking, reading comprehension, and commonsense reasoning, along with a proprietary internal dataset used to benchmark overall performance and robustness.

### 4.1 Baselines

We evaluate our approach against representative reasoning and debate paradigms discussed in Section 2. These include the **Single-Agent (SA)** model for zero-shot inference, **Chain-of-Thought (CoT)** reasoning for explicit stepwise deduction, and self-reflective methods such as **Self-Contrast (SC)** (Zhang et al., 2024a) and **Self-Consistency (SCON)** (Wang et al., 2022), which enhance robustness through internal critique or voting. Among multi-agent frameworks, we compare with **MAD** (Liang et al., 2023) and **CFMAD** (Fang et al., 2025), both of which employ inter-agent debates but suffer from fixed-length interactions and single-judge fragility.

### 4.2 Datasets and Metrics

We evaluate our framework on three QA-style benchmarks spanning factual, and commonsense reasoning along with an internal dataset to improve the catalog quality. **BoolQ** tests factual grounding through binary question answering, while **CosmosQA** focuses on causal and commonsense inference in everyday scenarios, and **Internal-QnA**, a 20K-sample proprietary dataset, evaluates factual ambiguity and long-debate calibration within an e-commerce catalog context; For internal dataset, we report only the incremental lift over the base

| Method | BoolQ (%) | CosmosQA (%) | Internal-QnA (△pp) |
|---|---|---|---|
| SA (baseline) | 71.8 | 61.3 | – |
| CoT | 78.5 | 68.1 | +5.5 |
| Self-Contrast | 81.1 | 69.3 | +7.1 |
| Self-Consistency | 80.8 | 70.0 | +6.8 |
| MAD | 82.3 | 72.8 | +8.4 |
| CFMAD | 83.8 | 74.3 | +10.6 |
| **SELENE** | **84.9** | **75.5** | **+14.7** |

Table 2: Accuracy (%) on public benchmarks (BoolQ, CosmosQA) and relative improvement (△ pp) on the proprietary **Internal-QnA** dataset. Absolute scores for Internal-QnA are omitted due to disclosure policies.

methodology, omitting absolute scores due to disclosure policy. Evaluation metrics include factual accuracy (↑) i.e. reducing inaccurate answers, token cost (↓; normalized to Single-Agent inference = 1×), and judge stability (↑).

### 4.3 Implementation Details

All experiments use **GPT-4-turbo-2025-04-09** via Open AI API call (log probs are only available via API call) as the backbone LLM with standardized prompts across methods (details in Appendix B). Also, we also benchmarked SELENE on other LLMs (GPT-4o-mini/Claude 3 Haiku) to measure the effectiveness of our approach (details in Appendix A). Inference parameters are fixed at temperature 0.3 and top-$p$ 0.9, except for **Self-Consistency** (Wang et al., 2022), which uses temperature 1.0 to enhance reasoning diversity. SDI thresholds $(\tau_1, \tau_2)$ are tuned on a small BoolQ-Internal-QnA validation set, and EWSC employs $K = 3$ parallel judgment passes with paraphrased evidence. All runs use the OpenAI API, and token cost is reported relative to Single-Agent inference ($1.0\times$ baseline).

**Findings.** **SELENE** consistently outperforms all baselines across factual and commonsense QA datasets, improving over CFMAD by +1.1 pp on BoolQ and +1.2 pp on CosmosQA. On the confidential Internal-QnA dataset, it yields a **+14.7 pp** improvement relative to the SA baseline, demonstrating superior handling of long-context and ambiguity-heavy reasoning scenarios without increasing model size or inference cost.

## 5 Ablation Studies

We see the impact of SELENE on the overall performance but to quantify the contribution of each component in SELENE, we perform stepwise ablation of each of the two components i.e. **Selective Debate Initiation (SDI)** module, followed by the

| Dataset | Method | Skip Rate | Accuracy on Skipped | Token Cost (×) |
|---|---|---|---|---|
| BoolQ | CFMAD | 0% | 83.6% | 3.7× |
| | SDI (ours) | 58% | 82.1% | **1.4×** |
| CosmosQA | CFMAD | 0% | 74.8% | 3.7× |
| | SDI (ours) | 43% | 73.2% | **1.8×** |
| Internal-QnA | CFMAD | 0% | – | 3.9× |
| | SDI (ours) | 27% | -0.8% | **2.1×** |

Table 3: **Comparison of SDI (ours) and CFMAD.** Factual datasets such as BoolQ show the highest skip rate, whereas ambiguous Internal-QnA queries still trigger debate, ensuring reliability where needed.

| Method | BoolQ | CosmosQA | Internal-QnA |
|---|---|---|---|
| CFMAD | 81.2 | 74.5 | – |
| **EWSC (ours)** | **86.1** | **80.2** | **–** |
| **Gain over CFMAD** | **+4.9** | **+5.7** | **+7.7** |

Table 4: **Performance on long-debate queries (>2 rounds).**

**Evidence-Weighted Self-Consistency (EWSC)** judge and compare their impact w.r.t CFMAD for it's best performance across all the datasets.

### 5.1 Effectiveness of Selective Debate Skipping

To assess whether SDI's gating mechanism reduces redundant computation without degrading accuracy, we measure the proportion of queries that bypass debate and compare their outcomes to fully debated cases. Queries are partitioned into two categories: **(a) Skipped**-low semantic disagreement ($D < \tau_1$) and low misalignment ($M < \tau_2$); **(b) Debated**-all remaining queries that trigger multi-agent reasoning. Table 3 shows that skip-debate decisions lead to a slight dip accuracy while reducing token usage by over 50%. Compared to CFMAD, which debates every query, SDI dynamically bypasses 30-60% of low-uncertainty cases, cutting computation (3.7× → 1.8×) with only a marginal 0.8-1.5 percentage point drop in accuracy. This demonstrates that SDI performs informed triage-debating only when necessary to maintain factual robustness.

### 5.2 Performance on Longer Debates

We further examine performance as a function of debate depth (Table 4). For queries requiring more than two reasoning rounds, **EWSC** delivers substantial accuracy gains, demonstrating its robustness in resolving ambiguous and evidence-intensive cases.

### 5.3 Judge Stability Analysis

To evaluate **EWSC** under evidence perturbations, we measure **Judge Stability** (↑), the consistency of final decisions across $K$=3 paraphrased evidence variants, defined as $1 - \text{Var}(s_i^{(k)})$, averaged across all questions, where higher values indicate more consistent judgments across perturbations(Refer to Table 5). Unlike CFMAD's single-judge setup, which is highly sensitive to phrasing, EWSC aggregates and weights consistent judgments, improving stability by 9-16 points-most notably on Internal-QnA-while using the same perturbation budget.

| Method | BoolQ | CosmosQA | Internal-QnA |
|---|---|---|---|
| CFMAD (base) | 0.84 | 0.79 | 0.72 |
| **SELENE (SDI +EWSC)** | **0.93** | **0.89** | **0.88** |

Table 5: **Judge Stability** (↑) under paraphrased evidence perturbations. EWSC markedly improves stability across all datasets, with the largest gains on Internal-QnA, where longer debates amplify judgment variance.

### 5.4 Summary

Across multiple datasets, our approach achieves the optimal balance between accuracy and efficiency. **SDI** dynamically allocates reasoning effort, reducing computation by approximately 50%, while **EWSC** enhances judgment stability in extended debates-most notably on long-context internal tasks. Together, they extend CFMAD into a scalable, debate-on-demand reasoning framework that remains both computationally efficient and epistemically reliable.

## 6 Conclusion

We introduced **SELENE**, a selective and evidence-aware framework for multi-agent reasoning that improves factual reliability without excessive computation. Unlike prior systems that debate on every query, SELENE combines two modules-**Selective Debate Initiation (SDI)** and **Evidence-Weighted Self-Consistency (EWSC)**-to adaptively balance efficiency and robustness. SDI triggers debate only under high epistemic uncertainty, while EWSC stabilizes final judgments by emphasizing low-variance, evidence-aligned decisions.

Together, these mechanisms form a reflective loop that emulates human deliberation: reason concisely when confident and deliberate when uncertain. Empirically, SELENE reduces redundant debate by over 50% while improving factual accuracy across benchmarks, demonstrating that *adaptive coordination*-not exhaustive interaction-is key to scalable and trustworthy reasoning. Future work will extend this paradigm to open-domain retrieval and long-context settings for further robustness.

## 7 Limitations

While our framework substantially improves factual robustness and computational efficiency over existing multi-agent debate systems, it has two notable limitations.

First, **the selective debate gating (SDI) relies on confidence–likelihood signals derived from model logits**, which may vary across architectures or fine-tuning setups. Although these signals generalize well on GPT-4 class models, calibration drift could affect threshold stability when applied to smaller or instruction-tuned LLMs.

Second, **the framework still depends on multi-turn debate for highly ambiguous or evidence-rich queries**-particularly those in our Internal-QnA dataset, where longer debates remain necessary to converge on factual consensus. While our early-stopping and variance-based judging mitigate semantic drift, future work could explore reinforcement or retrieval-augmented feedback loops to shorten these deep-debate cases further.

Overall, these limitations primarily concern scalability and cross-model generalization rather than conceptual soundness, and they point toward promising directions for adaptive thresholding and retrieval-informed reasoning in future research.

## References


Jie Chen, Bowen Zhao, Dian Yu, and Bill Yuchen Lin. 2023. Faithful chain-of-verification improves reasoning in large language models. *arXiv preprint arXiv:2310.04383*.

Zhenyu Cui, Hao Wang, Cheng Zhang, and Jie Zhou. 2025. Free-mad: Bias-reduced multi-agent debate via aggregated trajectory voting. *arXiv preprint arXiv:2502.02134*.

Yilun Du, Jiayuan Li, and Christopher D. Manning. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.13269*.

Wei Fang, Lin Chen, Rui Zhang, Ming Li, and Xiaodong Liu. 2025. Counterfactual multi-agent debate improves factuality and diversity in llm reasoning. *arXiv preprint arXiv:2501.04210*.

Jack Hills and Sam Anadkat. 2023. Using logprobs. https://cookbook.openai.com/examples/using_logprobs. Accessed: 2025-11-03.

Zeqiu Ji, Yujin Lee, Jason Fries, Danqi Chen, et al. 2023. Survey on hallucination in large language models. *arXiv preprint arXiv:2309.05922*.

Saurav Kadavath, Andy Lin, Nicholas Schiefer, Jacob Hilton, Owain Evans, Samuel R. Bowman, and Andreas Stuhlmüller. 2022. Language models (mostly) know what they know. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Anant Khandelwal, Happy Mittal, Shreyas Sunil Kulkarni, and Deepak Gupta. 2023. Large scale generative multimodal attribute extraction for e-commerce attributes. In *ACL (industry)*, pages 305–312.

Jie Li, Jian Zhou, Tao Lin, Han Wang, and Fang Liu. 2024. Calibrating large language models with log-probability guidance for factual reliability. *arXiv preprint arXiv:2402.08032*.

Weizhe Liang, Yanzhe Zhang, Yixin Kwon, Tianyi Ye, Yizhong Zheng, Jason Weston, Luke Zettlemoyer, and Mark Yatskar. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2023. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Potsawee Manakul, Alham Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Detecting llm hallucinations via token-level sampling. *arXiv preprint arXiv:2303.08896*.

Noah Shinn, Francesco Cassano, Bradley Labash, Dinesh Gopinath, Matthew Finlayson, Anca Dragan, Dorsa Sadigh, and Noah Goodman. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Chenglei Si, Hongxin Xie, Fangyuan Xu, Yue Zhang, Jing Ma, and Rui Zhang. 2023. Measuring uncertainty in large language models for hallucination detection. *arXiv preprint arXiv:2305.13669*.

Prateek Sircar, Aniket Chakrabarti, Deepak Gupta, and Anirban Majumdar. 2022. Distantly supervised aspect clustering and naming for e-commerce reviews. In *NAACL-HLT (Industry Papers)*, pages 94–102.

Hao Wang, Ming Zhou, Yue Zhang, and Zhiyuan Li. 2024. Cross-examination: Improving judge reliability in multi-agent llm debates. *arXiv preprint arXiv:2403.04127*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Ed Chi, Quoc Le, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Yizhong Wang and et al. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Yizhong Wang, Han Zhou, Tianyi Zhang, and Zhiyuan Liu. 2023. Self-contrast: Aligning large language models via contrastive self-refinement. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *ArXiv*, abs/2502.07266.

Rui Yang, Wei Lin, Hao Chen, and Lei Zhou. 2024. Judge summarization: Enhancing verdict quality in multi-agent llm debates. *arXiv preprint arXiv:2402.01987*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yue Ting Zhuang, and Weiming Lu. 2024a. Self-contrast: Better reflection through inconsistent solving perspectives. In *Annual Meeting of the Association for Computational Linguistics*.

Yixin Zhang, Yuchen Chen, Hao Li, Qian Liu, and Hongtao Xu. 2024b. Lm-detect: Likelihood-based hallucination detection in large language models. *arXiv preprint arXiv:2404.06112*.

Ke Zhou, Yuan Zhao, Pengfei Wang, and Chenguang Wang. 2024a. Factuality assessment of large language models via token probability entropy. *arXiv preprint arXiv:2401.03455*.

Liang Zhou, Ying Zhu, and et al. 2024b. Larger and more instructable large language models become less robust to prompt variations. *Nature*, 633:679–687.

## A Cross-LLM Robustness Evaluation

To assess the generalizability of **SELENE** across different reasoning backbones, we extended our evaluation to two additional large language models-**GPT-4o-mini** (OpenAI, 2025) and **Claude 3 Haiku** (Anthropic, 2024). Both models provide token-level log-probabilities through their public APIs, enabling introspective confidence scoring during the reasoning process. This allows SELENE's self-evaluative and contrastive modules to operate consistently across architectures.

**Discussion.** SELENE consistently outperforms CFMAD across both LLMs, with an average gain of **+2.0 points** and the largest improvement on **Internal-QnA (+8.4)**. This robustness indicates that SELENE's reflective modules-self-contrast and noise-aware reasoning-generalize effectively across architectures supporting log-prob introspection, highlighting the framework's model-agnostic adaptability.

| Model | Method | BoolQ | CosmosQA | Internal-QnA |
|---|---|---|---|---|
| GPT-4o-mini | CFMAD | 84.1 | 74.0 | – |
| | SELENE | **85.4** | **76.1** | – |
| | Δ (vs CFMAD) | +1.3 | +2.1 | +8.6 |
| Claude 3 Haiku | CFMAD | 83.2 | 73.5 | – |
| | SELENE | **84.6** | **75.7** | – |
| | Δ (vs CFMAD) | +1.4 | +2.2 | +8.2 |

Table 6: Cross-LLM comparison of accuracy (%) on BoolQ, CosmosQA, and Internal-QnA. Both GPT-4o-mini and Claude 3 Haiku expose `logprobs` via API, facilitating consistent introspective evaluation. SELENE maintains its advantage across all tasks, confirming robustness to underlying model variance.

## B Prompt Flow and Implementation Details

All prompts are executed using GPT-4-Turbo-2025-04-09 with temperature = 0.3 and top-p = 0.9. Each box below shows the actual prompt used in SELENE at various stages.

**Initial Reasoning**

**Instruction:**
You are an expert reasoning agent. Decompose your thought process to expose your reasoning path. Provide: (1) a Yes/No answer, (2) a structured reasoning trace showing key evidence and intermediate logic, (3) your confidence score (0-1).
**Example:**
Q: Can penguins fly?
A: No. [Reasoning Path: Penguins are birds → most birds fly → but penguins evolved for swimming, not flying.]
Confidence: 0.91

**Prompt for Debate**

**Instruction**: You are participating in a factual debate. Each member has already provided an initial answer to the question. Your goal is to improve your reasoning and refine your final answer through evidence-based discussion.
Debate Rules:
1. You will see the question, your previous answer, and the responses of other members.
2. Compare their reasoning and evidence with your own.
3. Identify any factual errors, unsupported claims, or missing considerations.
4. Revise your answer if you find stronger evidence or more consistent reasoning.
5. Focus strictly on factual accuracy - not style, length, or rhetorical persuasion.
6. Keep reasoning concise (2-4 sentences). Avoid repetition or emotional language.
7. Each round aims to reduce disagreement and reach a stable consensus.

At the end of your turn, output your revised reasoning.

**Example**:
Question: Can penguins fly?
Your previous answer: "Yes, penguins are birds."
Other agents said:
- Agent B: "No, penguins are flightless birds."

- Agent C: "They use wings for swimming, not for flight."
Revised answer: "No, penguins are flightless birds that use their wings for swimming."

**EWSC Judgment**

**Instruction**: You are a factual judge. Your goal is to evaluate how factually correct a model's answer is with respect to the given evidence.
You will receive:
(1) A question (user query),
(2) A candidate answer from one reasoning agent, and
(3) A set of evidence snippets (which may be paraphrased or partially sampled).

Your task:
- Read the evidence carefully.
- Determine whether the answer is factually supported, contradicted, or not covered by the evidence.
- Assign a factual correctness score between 0 and 1.
Be consistent: ignore stylistic or phrasing variations across evidence versions. Focus only on factual alignment.

**Example**:
Question: Did Galileo invent the telescope?
Candidate Answer:
Yes, Galileo invented the telescope in 1609.
Evidence:
- The first practical telescopes were built in the Netherlands in 1608.
- Galileo improved the design and used it for astronomy.
Analysis:
The evidence contradicts the claim that Galileo "invented" the telescope - he refined an earlier Dutch design. The answer shows partial relevance but factual inaccuracy.
Factual correctness score:
0.4

# C Logit Retrieval via API

To extract model logits for downstream calibration and confidence scoring, we include the `logprobs` parameter in the API call. If set to a positive integer $K \leq 5$, the API returns the log-probabilities of the top $K$ tokens at each generation step (Hills and Anadkat, 2023). Below is an example using the OpenAI Python client:

```
import openai

openai.api_key = "YOUR_API_KEY"

response = openai.ChatCompletion.create(
    model="gpt-4o-mini",
    messages=[
        {
            "role": "system",
            "content": "You are a helpful assistant."
        },
        {
            "role": "user",
            "content": "QUESTION_PROMPT_HERE"
        }
    ],
    max_tokens=1,
    temperature=0.0,
    logprobs=5,
    top_logprobs=5
)
```

```
# The response object includes:
#   response.choices[0].logprobs.token_logprobs
#   response.choices[0].logprobs.top_logprobs
# These correspond to logp(token | context).
log_probs = response.choices[0].logprobs.token_logprobs
```

The extracted log-probabilities $\ell_i = \log p_\theta(r_i \mid q)$ are converted into calibrated probabilities using $\sigma(\ell_i) = 1/(1 + e^{-\ell_i})$, which supports our confidence-alignment analysis.

## C.1 Qualitative Examples

To illustrate how SELENE adapts reasoning depth to question difficulty, we present qualitative cases drawn from the BoolQ and Internal-QnA datasets.

- **Trivial factuality (skip debate):** *"Is Mount Everest the highest mountain in the world?"* - All agents output "Yes" with high alignment ($D = 0.03$) and low miscalibration ($M = 0.05$). SDI detects stable consensus and terminates early, avoiding unnecessary debate while achieving 100% accuracy.
- **Hidden overconfidence (debate triggered):** *"Can penguins fly?"* - Two agents initially respond "Yes" citing that penguins are birds ($c_i > 0.9$, $\sigma(\ell_i) < 0.5$), showing high overconfidence. One agent correctly answers "No." The resulting $D = 0.48$ and $M = 0.42$ exceed thresholds, prompting a full debate. Through cross-argumentation ("Penguins are flightless birds adapted for swimming"), consensus converges to the correct "No."
- **Ambiguous causality (multi-hop reasoning):** *"Was Marie Curie's discovery related to an element used in cancer treatment?"* - Initial disagreement arises between "Yes (radium used in radiotherapy)" and "No (Curie did not directly develop treatment)." SDI triggers a multi-hop debate referencing scientific evidence chains (Curie $\rightarrow$ Radium $\rightarrow$ Radiotherapy). EWSC then aggregates stable, low-variance judgments across paraphrased evidence to yield the correct "Yes."
- **Long-debate internal reasoning (Internal-QnA):** *"Is a product eligible for free replacement if delivered without warranty card?"* - Agents diverge semantically due to policy exceptions. SDI initiates extended debate ($T = 3$), and EWSC consolidates consistent evidence-based answers ("Yes, if purchase is verified via invoice"), improving factual accuracy in ambiguous policy questions.

These cases show that SDI effectively skips low-uncertainty questions while EWSC stabilizes multi-turn reasoning under disagreement or overconfidence, together yielding both computational efficiency and factual robustness.

# D Algorithm

**Algorithm 1** Selective Debate Initiation (SDI) + Evidence-Weighted Self-Consistency (EWSC)

**Input:** Query $q$, agents $\{A_1, \ldots, A_N\}$, evidence $R_q$
**Output:** Final judged answer $r^*$

**// Stage 1: Initial Reasoning**
**for** each agent $A_i$ **do**
  Generate answer $r_i \in \{\text{Yes, No}\}$ with confidence $c_i$
  Compute $\ell_i = \log p_\theta(r_i|q)$ and embedding $E(r_i) = \text{Enc}_\theta([q; r_i])$
**end for**
**// Stage 2: Compute Epistemic Signals**
$D \leftarrow \frac{2}{N(N-1)} \sum_{i<j} [1 - \cos(E(r_i), E(r_j))]$ ▷ semantic disagreement
$M \leftarrow \frac{1}{N} \sum_i |c_i - \sigma(\ell_i)|$ ▷ calibration misalignment
**// Stage 3: Selective Debate Decision**
**if** $D < \tau_D$ **and** $M < \tau_M$ **then**
  **Skip debate:** adopt consensus response $r^+$
  **Single-judge decision:** $r^* \leftarrow J_\theta(r^+, R_q)$
  **return** $r^*$ ▷ direct resolution via single judge
**else**
  **Trigger debate:**
  **for** $t = 1$ to $T_{\max}$ **do**
    **for** each agent $A_i$ **do**
      $r_i^{(t+1)} \leftarrow F_{\theta_i}(r_i^{(t)}, \{r_j^{(t)} : j \neq i\}, q)$
    **end for**
    Compute $\Delta D^{(t)} = D^{(t-1)} - D^{(t)}$
    **if** $|\Delta D^{(t)}| < \epsilon$ **then**
      **break** ▷ stop when semantic stability reached
    **end if**
  **end for**
**end if**
**// Stage 4: Robust Judging (EWSC)**
**for** each final response $r_i^{(T)}$ **do**
  **for** $k = 1$ to $K$ **do**
    Sample perturbed evidence $R_q^{(k)}$
    $s_i^{(k)} \leftarrow J_\theta(r_i^{(T)}, R_q^{(k)})$
  **end for**
  $S_i \leftarrow \frac{\sum_k s_i^{(k)} e^{-\text{Var}_k[s_i]}}{\sum_k e^{-\text{Var}_k[s_i]}}$ ▷ variance-weighted consensus
**end for**
**return** $r^* = \arg\max_i S_i$ ▷ final stable judgment