iMAD: Intelligent Multi-Agent Debate for Efficient and Accurate LLM Inference

Wei Fan, JinYi Yoon, Bo Ji

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
{fanwei, jinyiyoon, boji}@vt.edu

Abstract

Large Language Model (LLM) agent systems have advanced rapidly, driven by their strong generalization in zero-shot settings. To further enhance reasoning and accuracy on complex tasks, Multi-Agent Debate (MAD) has emerged as a promising framework that engages multiple LLM agents in structured debates to encourage diverse reasoning. However, triggering MAD for every query is inefficient, as it incurs substantial computational (token) cost and may even degrade accuracy by overturning correct answers from single-agent. To address these limitations, we propose intelligent Multi-Agent Debate (iMAD), a token-efficient framework that selectively triggers MAD only when it is likely to be beneficial (i.e., correcting an initially wrong answer). To achieve this goal, iMAD learns generalizable model behaviors to make accurate debate decisions. Specifically, iMAD first prompts a single agent to produce a structured self-critique response, from which we extract 41 interpretable linguistic and semantic features capturing hesitation cues. Then, iMAD uses a lightweight debate decision classifier, trained using our proposed FocusCal loss without test-dataset-specific tuning, to make robust zero-shot debate decisions. Through extensive experiments using six (visual) question answering datasets against five competitive baselines, we show that iMAD significantly reduces token usage (by up to 92%) while also improving final answer accuracy (by up to 13.5%).

Code — https://github.com/Fanwei100/iMAD
Technical Report — http://arxiv.org/abs/2511.11306

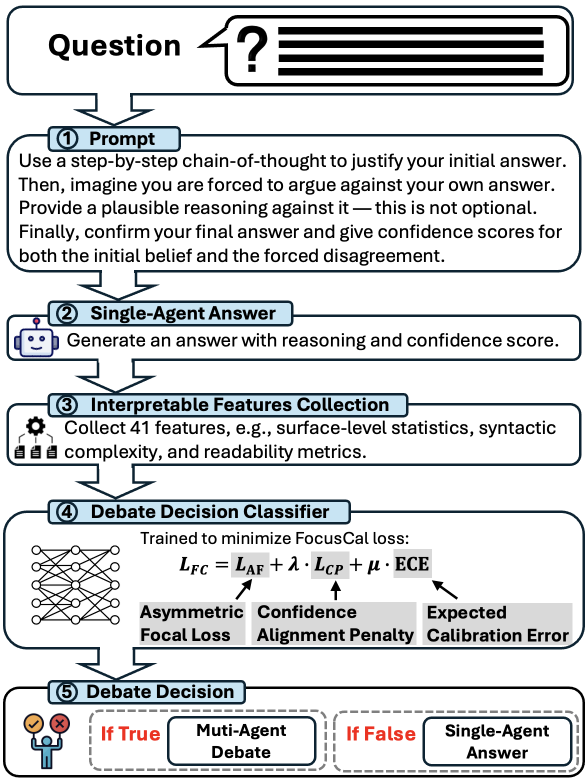

Figure 1: Overall workflow of iMAD.

1 Introduction

With the rapid progress in Large Language Models (LLMs), agent systems have shown impressive zero-shot reasoning capabilities across tasks such as (visual) question answering, problem solving, or code generation. This ability in zero-shot settings, without access to evaluation data, makes LLM agents appealing for real-world applications by enabling fast and scalable deployment across diverse domains (Wan et al. 2023; Yang, Tsai, and Yamada 2025). These agent systems typically rely on a single LLM agent to generate step-by-step reasoning, using methods like Chain-of-Thought (CoT) (Wei et al. 2022) or Self-Consistency (Wang et al. 2023). However, these approaches often suffer from limited diversity and may overlook alternative reasoning paths as they rely on a single agent's perspective.

To address this limitation, recent studies have explored multi-agent systems that enable collaboration or debate among multiple agents to enhance reasoning and decision-making (Liang et al. 2024; Chen, Saha, and Bansal 2024). Among these approaches, *Multi-Agent Debate (MAD)*, inspired by the role of structured discourse in human cognition, has emerged as a particularly promising framework (Wu et al. 2024). In MAD, multiple agents independently reason over a query and critique each other's output through structured interactions, stimulating adversarial dialogue and iterative refinement (Liang et al. 2024; Tillmann

2025). Such interactions encourage diverse reasoning paths and perspective shifts, enabling agents to recover from faulty initial answers and thus often outperform single-agent systems (Liang et al. 2024; Hsu et al. 2025).

Despite this benefit, MAD frameworks face two critical limitations that hinder their practical deployment. First, *due to iterative LLM queries, MAD incurs significantly higher computational costs*–measured in total token usage, which includes both input tokens (prompts to each agent) and output tokens (generated responses). Most MAD systems consume three to five times more tokens than single-agent baselines (Liu et al. 2024), making them costly to scale (see Insight 1). Second, perhaps more counterintuitively, *MAD does not consistently improve response quality*. Both prior work (Zhang et al. 2025) and our own empirical analysis show that in many cases, the single-agent output is already correct, making it redundant to trigger MAD. In other cases, the error in the single-agent response cannot be corrected by MAD, or even worse, triggering MAD may override a correct answer with an incorrect one, resulting in degraded accuracy (see Insight 2). These observations suggest that while MAD can be beneficial, applying it to every query may be inefficient. Not only does it incur substantial computational costs, but it can also degrade accuracy. Therefore, we need an intelligent and principled mechanism that can selectively trigger MAD only when it is likely to be beneficial. This raises a key question: *When should a debate be performed to preserve the benefits of MAD while avoiding unnecessary token costs and potential accuracy degradation?*

A straightforward approach is to use the confidence score (typically computed as the average log-probability of the output tokens) to estimate single-agent answer correctness. Leveraging this idea, a concurrent work recently proposed a selective MAD framework, called DOWN, which attempts to trigger MAD when the confidence score falls below a threshold (Eo et al. 2025). However, it requires a threshold tuned on a subset of the evaluation data, which violates the core zero-shot setting assumption. Moreover, even with a fixed threshold, our empirical analysis reveals that confidence scores alone are not reliable indicators of whether MAD is necessary (see Insight 3). We observe that confidence scores could be high even for incorrect answers, revealing the model's overconfidence. Also, confidence scores are often misaligned with reasoning uncertainty: responses that contain hesitation cues (such as hedging, contradictions, or shallow reasoning) may still receive a high score. This misalignment causes two undesirable situations: skipping the necessary MAD or triggering it unnecessarily.

While existing methods like repeated sampling (i.e., querying the same agent multiple times) or invoking more LLM agents can estimate whether MAD is necessary, they incur high token costs and thus are not scalable (Wang et al. 2023). This highlights the need for token-efficient mechanisms that can make informed decisions about when to trigger MAD based on the initial single-agent output. There are two key challenges to achieving this goal:

(C1) *How to design an effective prompt that guides the single-agent response to expose richer features for making a debate decision?* This decision must rely on features embedded in the initial single-agent output. These features need to be informative and can be efficiently extracted (i.e., without requiring repeated sampling or additional LLM queries).

(C2) *How to design a mechanism that intelligently decides when to trigger MAD for efficient and accurate inference in the zero-shot setting?* One needs to identify when MAD is more likely to recover incorrect answers without relying on the evaluation dataset. This requires learning generalizable model behaviors, which involves addressing the aforementioned issues of LLM agents: overconfidence in incorrect answers and misalignment between confidence scores and semantic cues of hesitation in the response.

To address these challenges, we propose **i**ntelligent **M**ulti-**A**gent **D**ebate (iMAD), a lightweight debate-triggering framework for token-efficient MAD in the zero-shot setting. iMAD selectively triggers MAD only when it is more likely to improve the final answer. The overall workflow of iMAD is illustrated in Fig. 1. We summarize our main contributions along with the key components of iMAD as follows:

- To address Challenge (C1), we propose the iMAD framework with a structured self-critique prompt (Steps ①-② in Fig. 1). This prompt directs a single agent to produce (i) an initial CoT justification, (ii) a required self-critique that argues for a plausible alternative, and (iii) confidence scores for both perspectives. This prompt stimulates a lightweight internal mini-debate without adding input tokens and incurs only minimal additional output tokens. This design offers rich semantic and uncertainty cues, enabling accurate and token-efficient debate decisions.
- To address Challenge (C2), we formulate MAD triggering as a classification problem. From each structured single-agent response, we extract 41 interpretable linguistic and semantic features along with the confidence score (Step ③), which will be fed into a lightweight debate decision classifier of a multi-layer perceptron (MLP) (Step ④). To enable accurate decisions in zero-shot settings, the classifier learns generalizable model behaviors using a proposed Confidence-Calibrated *FocusCal* loss that integrates: (i) *Asymmetric Focal loss* (L_{AF}) to penalize overconfident errors and emphasize incorrect cases; (ii) *Confidence Penalty* (L_{CP}) to penalize misalignment between confidence scores and semantic uncertainty in the response; and (iii) *Expected Calibration Error* (ECE) to encourage the predicted debate-triggering score to align with empirical correctness. This enables the classifier to prioritize debatable cases (i.e., recoverable errors) while reducing unnecessary MAD (Step ⑤).
- We evaluate iMAD on three question answering (QA) and three visual question answering (VQA) datasets against five competitive baselines (including two single-agent and three multi-agent frameworks). iMAD reduces token usage by up to 92% while improving accuracy by up to 13.5% through selectively skipping debates that are unnecessary or detrimental. Notably, we train the classifier solely on two representative datasets selected to capture diverse model reasoning behaviors, allowing the classifier to learn generalizable model behaviors and perform effectively across six held-out datasets.

2 Related Work

We categorize highly related works into three groups.

Single-Agent and Multi-Agent LLMs. LLMs have demonstrated strong reasoning capabilities in single-agent settings, where a single LLM agent performs all reasoning steps independently. Foundational methods like CoT prompt models to generate intermediate reasoning steps, enabling better handling of complex tasks (Wei et al. 2022). Recently, Self-Consistency further improves accuracy over CoT by sampling multiple outputs and selecting the most frequent answer (Wang et al. 2023, 2024a; Li et al. 2024b). However, single-agent reasoning relies solely on internal sampling, lacking perspective diversity and explicit self-correction. To address this, multi-agent LLM frameworks leverage multiple agents to reason independently or coordinate through structured interaction, thus improving accuracy over single-agent approaches. Some methods generate multiple outputs for joint evaluation (e.g., CoMM (Chen, Han, and Zhang 2024)), while others assign hierarchical agent roles (e.g., Mixture-of-Agents (MoA) to progressively refine reasoning (Wang et al. 2025; Chen et al. 2024; Li et al. 2024a)). While these methods stimulate collaborative reasoning among agents, they often incur high computational costs and deliver inconsistent improvements over single-agent baselines (Zhang et al. 2025; Pan et al. 2025).

MAD Frameworks. MAD frameworks represent a structured subclass of multi-agent LLM where agents engage in explicit argumentative exchanges (e.g., critiques, rebuttals, or deliberation) to refine initial outputs (Tillmann 2025). Existing MAD methods include role-based debates with assigned affirmative, negative, and moderator roles (Liang et al. 2024; Wang et al. 2024b), implicit debate via input perturbation and aggregation (e.g., Reconcile (Chen, Saha, and Bansal 2024)), and intra-agent self-refinement (Srivastava et al. 2025; Zhang et al. 2024). Building on these MAD designs, GroupDebate extends MAD by coordinating subgroup debate (Liu et al. 2024). While MAD enhances interpretability and error correction, recent studies show that it can also introduce noise or overturn correct single-agent answers, thus degrading accuracy (Zhang et al. 2025).

Confidence-based Selective Debate. The concurrent work DOWN aims to reduce MAD token costs by using LLM-generated confidence scores to decide when to trigger debate (Eo et al. 2025). However, selecting an appropriate confidence threshold requires labeled evaluation data, which violates the zero-shot setting assumption normally upheld by current single-agent and MAD baselines and limits real-world applicability. Moreover, confidence scores alone are unreliable in determining whether triggering a debate will improve the answer (see Insight 3). Although DOWN saves tokens by skipping debate on high-confidence responses, it often fails to identify cases where debate is beneficial and misses opportunities to correct recoverable errors.

3 Key Insights

While MAD shows promise in enhancing LLM reasoning, its deployment remains limited, as modest accuracy gains often come with substantial computational cost. In this section, we first quantify the token overhead of MAD compared to single-agent CoT. Then, we analyze when MAD is beneficial and reveal that the positive impact exists only for a subset of instances. Finally, we examine whether standard uncertainty heuristics (e.g., based on confidence score) can effectively guide debate triggering.

Dataset	Single-Agent (CoT)		Multi-Agent Debate (MAD)	
	Acc (%)	# Token	Acc (%)	# Token
MEDQA	76.6	653	81.9	4,034
MMLU	86.8	764	89.5	3,348
GSM8K	71.3	618	76.4	3,446
OKVQA	88.3	1,945	89.8	7,803
VQA-v2	77.5	2,245	81.0	8,796
ScienceQA	86.0	1,720	89.4	6,777

Table 1: Comparison of accuracy (Acc) and average token costs (# Token) per question between single-agent CoT and MAD frameworks across QA and VQA datasets.

(Insight 1) MAD achieves a higher accuracy at the cost of a substantial token overhead. We quantify the trade-off between accuracy and token costs by comparing CoT (Wei et al. 2022) and MAD (Liang et al. 2024) across six QA and VQA datasets in Table 1. Consistent with prior findings (Liu et al. 2024), we observe that MAD achieves a higher accuracy than CoT, with gains ranging from 1.5% (on OKVQA) to 5.3% (on MEDQA). However, MAD consumes 3–5 times more tokens than CoT, mainly due to routing the same query to multiple agents, each requiring separate input prompts and generating individual responses (Chen et al. 2025; Eo et al. 2025). This cost is more pronounced in VQA tasks, where visual inputs further increase token usage. These observations indicate that the accuracy gains come at a high token cost, making MAD impractical to deploy at scale.

(Insight 2) The accuracy gains in MAD are primarily driven by a subset of cases. While a concurrent work has noted a similar observation (Zhang et al. 2025), we provide a systematic breakdown to quantify the specific sources of MAD's accuracy gains. Specifically, we categorize each input instance into four cases: (i) incorrect in single-agent but correct in MAD (✗→✓); (ii) correct in single-agent but incorrect in MAD (✓→✗); (iii) correct in both (✓→✓); and (iv) incorrect in both (✗→✗). As shown in Table 2, the ideal scenario where MAD makes corrections (✗→✓) accounts for a small portion (e.g., 4.9% in OKVQA to 19.1% in GSM8K). In contrast, many debates are either redundant (i.e., single-agent answers are already correct: ✓→✓), ineffective (i.e., unresolved single-agent errors: ✗→✗), or harmful (i.e., flipping correct single-agent answers to incorrect: ✓→✗). This shows that while MAD improves the overall accuracy, the benefit is limited to a small portion of cases. Thus, indiscriminately applying MAD to all cases could waste computational resources and even degrade accuracy.

Dataset	✗→✓ (%)	✓→✗ (%)	✓→✓ (%)	✗→✗ (%)
MEDQA	11.9	6.6	70.0	11.5
MMLU	6.9	4.2	82.6	6.3
GSM8K	19.1	14.0	57.3	9.6
OKVQA	4.9	3.4	84.9	6.8
VQA-v2	9.2	5.7	71.8	13.3
ScienceQA	7.9	4.5	81.5	6.1

Table 2: Breakdown of MAD outcomes across datasets: percentage of cases where the MAD flipped the answer correctly (✗→✓) or wrong (✓→✗) from single-agent CoT answer, and percentage of cases where both MAD and single-agent CoT are correct (✓→✓) or wrong (✗→✗).

Figure 2: Cumulative Density Function (CDF) of confidence scores for correct and incorrect single-agent answers.

(Insight 3) Confidence scores are unreliable indicators of whether debate is beneficial or not. We investigate whether LLM-generated confidence scores can effectively indicate when MAD is likely to improve answers (Yang et al. 2024; Eo et al. 2025). A natural hypothesis is that answers with a high confidence score may not need MAD; only answers with a low confidence score could benefit from MAD. To test this, we consider confidence scores from various prompting strategies (Yang et al. 2024; Eo et al. 2025) and analyze their alignment with cases where MAD corrects initially incorrect single-agent answers. In Fig. 2, we observe that the Cumulative Density Function (CDF) of confidence scores is highly right-skewed and poorly aligned with answer correctness or debate effectiveness. Notably, incorrect answers often receive high confidence scores, sometimes even exceeding correct answers, indicating a strong bias of overconfidence. Moreover, even many hesitant or shallow responses still receive inflated confidence scores. This misalignment between the confidence score and the uncertainty in reasoning undermines the effectiveness of heuristics that make debate decisions based on confidence scores only.

4 Our Design: iMAD

In this section, we present iMAD, a token-efficient framework that selectively triggers MAD only when it is likely to correct an initially incorrect single-agent answer. Leveraging the aforementioned insights, iMAD aims to substantially reduce the token overhead of MAD while retaining or even improving the accuracy (Insight 1). We begin with an overview of the iMAD framework (Section 4.1), which integrates structured self-critique prompting with a lightweight classifier to decide whether debate should be triggered (Insight 2). The classifier leverages interpretable features extracted from the single-agent response to assess the need for debate, enabling robust debate decisions in zero-shot settings without dataset-specific tuning. To train this classifier effectively, we propose FocusCal loss (L_{FC}) (Section 4.2) to address the aforementioned issues of LLM: overconfidence and misalignment between confidence scores and semantic uncertainty (Insight 3). This design enables iMAD to make token-efficient and accurate debate decisions based on single-agent responses in zero-shot settings.

4.1 Framework Overview

As shown in Fig. 1, iMAD comprises three core stages: (i) generating a structured response from a single-agent LLM using a self-critique prompt (Steps ①-②); (ii) extracting interpretable features from the generated output (Step ③), and (iii) applying a well-trained debate decision classifier to determine whether MAD should be triggered (Steps ④-⑤).

(i) Generating Structured Self-Critique Single-Agent Response (Steps ①-②). Given an input question, the system first prompts the LLM to generate a structured response with three key components: (i) an initial CoT justification supporting the original answer; (ii) a required self-critique presenting a counterargument; and (iii) a final reflection including the chosen answer and explicit confidence scores for both initial reasoning and the self-critique perspectives. This structure stimulates a mini-debate: if both perspectives provide similarly strong or weak reasoning with comparable confidence scores, the model is likely to show internal hesitation, suggesting that MAD could be beneficial. Conversely, if one side presents a clear and well-supported argument while the other is weak, the answer is likely already determined, either confidently correct or confidently incorrect. In the latter case, the internally coherent but flawed reasoning makes it difficult to correct through debate.

(ii) Extracting Interpretable Features (Step ③). We extract 41 interpretable linguistic and semantic features from the structured single-agent output, drawing from the question, initial reasoning, and self-critique. These interpretable features capture human-understandable cues to reasoning quality and internal hesitation, including surface-level statistics, readability scores, part-of-speech counts (e.g., nouns, verbs, and adjectives), question-type indicators, and lexical cues of uncertainty, such as hedging and contrast. The feature set enables fine-grained detection of subtle uncertainty cues that are often not reflected in the model's raw confidence scores. These interpretable features help the classifier identify when MAD is likely to be beneficial by capturing uncertainty cues that are not well aligned with confidence scores. Rather than relying on a subset of features, we combine complementary semantic, syntactic, and pragmatic signals to form a holistic view of model behavior. This is crucial for generalization in zero-shot settings without access to an evaluation dataset. A detailed list of all 41 features is provided in the online technical report.

(iii) Making Debate Decisions via the Classifier (Steps ④-⑤). We develop a lightweight MLP-based classifier to decide whether to trigger MAD based on features extracted from a structured single-agent response. The input feature vector $\mathbf{z} \in \mathbb{R}^d$, a d-dimensional real-valued vector space, where d is the total number of extracted features. The vector $\mathbf{z}$ includes LLM-generated confidence score p_{LLM}, along with semantic and linguistic features from the question, answer, and self-critique. The classifier outputs a scalar $p \in (0, 1)$, indicating the likelihood that the single-agent answer is either correct or unrecoverably incorrect (where a debate is considered unnecessary).

During training, we run the single-agent pipeline on each instance to generate an answer and self-critique, and then assign a binary label y where $y = 1$ if the answer matches the ground truth and $y = 0$ otherwise. The classifier is trained with these binary correctness labels, but its goal is not to replicate these labels. Instead, it aims to identify debatable cases: answers likely to be wrong but potentially correctable through MAD. This distinction is crucial because confidently incorrect answers, often resulting from coherent but flawed reasoning, are unlikely recoverable via MAD.

During inference, we apply a decision threshold $\tau \in (0, 1)$, tuned on a validation set, to the predicted score p: if $p < \tau$, MAD is triggered; otherwise, the original single-agent answer is retained. A high p indicates triggering MAD is unnecessary, while a low p flags uncertain and potentially recoverable errors that warrant MAD. This selective mechanism enables iMAD to devote computational resources to the most uncertain and error-prone cases, reducing token costs while improving accuracy. The classifier is trained using our proposed FocusCal loss, detailed in Section 4.2.

4.2 Debate Decision Classifier with FocusCal Loss

To train the classifier, we propose FocusCal loss (L_{FC}), a composite objective addressing overconfidence and uncertainty in single-agent confidence scores (Insight 3). It combines three components: (i) Asymmetric Focal Loss (L_{AF}) that targets the overconfidence issue by penalizing confidently incorrect predictions more than correct ones, encouraging the model to remain cautious on borderline cases; (ii) Confidence Penalty (L_{CP}) that aligns the predicted score p with an auxiliary uncertainty score $u \in (0, 1)$, derived from semantic hesitation features via an MLP, penalizing overconfident predictions with uncertain reasoning; (iii) Expected Calibration Error (ECE) that regularizes predicted scores for empirical calibration (Nixon et al. 2019). Together, these components help the classifier detect recoverable errors that merit MAD and avoid unnecessary debate.

To realize this design, the classifier passes the input feature vector $\mathbf{z}$ through a shared MLP feature encoder $f_{\text{e}}(\cdot)$ to produce a high-level representation, which is then fed into two separate output heads: a correctness head $f_p(\cdot)$ and a hesitation head $f_u(\cdot)$, producing two scalar logits ℓ_{p} and ℓ_{u}:

$$\ell_{\text{p}} := f_p(f_{\text{e}}(\mathbf{z})) \text{ and } \ell_{\text{u}} := f_u(f_{\text{e}}(\mathbf{z})). \tag{1}$$

To integrate the scalar LLM confidence score p_{LLM} with the MLP-produced logit ℓ_{p} and keep them mathematically consistent, we convert p_{LLM} into ℓ_{LLM} in the logit space:

$$\ell_{\text{LLM}} := \log\left(\frac{p_{\text{LLM}}}{1 - p_{\text{LLM}}}\right). \tag{2}$$

The predicted score p and the uncertainty score u are then calculated as:

$$p := \sigma(w_1 \cdot \ell_{\text{LLM}} + w_2 \cdot \ell_{\text{p}} + \epsilon), \tag{3}$$

$$u := \sigma(\ell_{\text{u}}), \tag{4}$$

where $w_1, w_2, \epsilon \in \mathbb{R}$ are learnable parameters and $\sigma(\cdot)$ is the sigmoid function. This two-headed MLP separates the predicted score p, used for debate decisions, from the auxiliary uncertainty score u, which captures hesitation cues in the reasoning path and is supervised via L_{CP}. This design enables the classifier to better identify hesitant and potentially recoverable errors that warrant debate, while skipping MAD for confidently correct answers or unrecoverable errors.

To optimize the classifier for accurate and calibrated debate triggering, we train it using the FocusCal loss:

$$L_{\text{FC}}(y, p, u) := L_{\text{AF}}(y, p) + \lambda \cdot L_{\text{CP}}(y, p, u) + \mu \cdot \text{ECE}(y, p). \tag{5}$$

The weights λ and μ are non-negative coefficients tuned via grid search on a held-out validation set to balance the contributions of uncertainty alignment and calibration. We discuss the details of each term in L_{FC} below.

Asymmetric Focal Loss (L_{AF}). This term places the strongest penalty on the cases where the classifier assigns a high predicted score p to an answer that is factually incorrect (i.e., labeled $y = 0$). These are exactly the instances where MAD should be triggered but was mistakenly skipped. To address this, we adopt an asymmetric focal loss defined as:

$$L_{\text{AF}}(y, p) := \begin{cases} -\alpha_1 (1 - p)^\gamma \log(p), & \text{if } y = 1, \\ -\alpha_0\, p^\gamma \log(1 - p), & \text{if } y = 0, \end{cases} \tag{6}$$

where $\gamma > 0$ is a focusing parameter that down-weighs well-classified examples (i.e., when the predicted score p is close to the ground-truth label), and emphasizes incorrect or harder cases. The class-specific weights $\alpha_1, \alpha_0 > 0$ control the relative emphasis on each class. We typically set $\alpha_0 > \alpha_1$ to attribute a large penalty to cases where the model assigns a high p to incorrect cases ($y = 0$), reflecting overconfident decisions that skip needed debate. This loss formulation directly addresses overconfidence by encouraging the classifier to assign a low predicted score p to incorrect single-agent answers. The asymmetric focal loss L_{AF} is thus designed to emphasize wrong predictions, many of which can be recovered through MAD, thus helping the classifier better identify cases that warrant debate.

Confidence Penalty (L_{CP}). To further align the model's predicted score p with its uncertainty score u, we introduce a regularization loss term that penalizes misalignment between them. This uncertainty score u reflects the model's internal hesitation: a high value of u indicates distributed uncertainty across semantic interpretations, while a low value of u indicates peaked, confident predictions. Although p expresses the overall correctness belief, u offers a complementary perspective by capturing the uncertainty or hesitation

embedded in the response features. To reconcile these two signals, we define Confidence Penalty as follows:

$$L_{\text{CP}}(y, p, u) := \begin{cases} u^2, & \text{if } y = 0 \text{ and } p > \tau, \\ (1-u)^2, & \text{if } y = 1 \text{ and } p < \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

This confidence penalty term penalizes under-confidence for correct answers (high uncertainty occurs with a low p) and overconfidence for incorrect ones (low uncertainty occurs with a high p). By enforcing consistency between the predicted score p and the uncertainty score u, this loss mitigates the misalignment between p and the semantic hesitation signals, leading to more reliable debate triggering decisions.

Expected Calibration Error (ECE). To improve the reliability of p as a debate-triggering score, we incorporate ECE to encourage its alignment with empirical correctness. Suppose the dataset contains N instances indexed by $i \in \{1, \ldots, N\}$, where each instance i has a predicted score $p^{(i)} \in [0, 1]$ and a binary ground-truth label $y^{(i)} \in \{0, 1\}$. Based on $p^{(i)}$, We divide the interval $[0, 1]$ into B equal-width bins and assign each instance i to a bin $b \in \{1, \ldots, B\}$. Let $\mathcal{I}_b$ denote the set of indices i whose predicted scores fall into bin b. We then compute ECE as:

$$\text{ECE}(y, p) := \sum_{b=1}^{B} \frac{1}{N} \left| \sum_{i \in \mathcal{I}_b} p^{(i)} - \sum_{i \in \mathcal{I}_b} y^{(i)} \right|. \quad (8)$$

For each bin, we measure the average absolute difference between predicted scores and ground-truth labels. Minimizing ECE aligns p with empirical correctness, leading to more reliable decision boundaries for triggering MAD.

5 Evaluation

We report token efficiency (i.e., token usage per question) and accuracy (i.e., final answer correctness) across six datasets. We also analyze iMAD's debate decisions on whether to trigger or skip MAD, and how often these decisions lead to beneficial outcomes. In the online technical report, we further provide additional analysis of token efficiency and inference time, and present additional results, including ablation studies on structured self-critique prompting and FocusCal loss, as well as cross-LLM evaluation.

Datasets. We evaluate on six datasets across textual QA and image-text VQA. The QA benchmarks include (i) MedQA: USMLE-style medical QA (Jin et al. 2021); (ii) MMLU: professional exam questions (Hendrycks et al. 2021); and (iii) GSM8K: grade-school math word problems (Cobbe et al. 2021). For VQA, we use (iv) OKVQA: emphasizes visual questions that require external knowledge (Marino et al. 2019); (v) VQA-v2: natural image QA with reduced bias (Goyal et al. 2017); and (vi) ScienceQA: science questions from school curricula (Lu et al. 2022).

Baselines. We compare iMAD with five strong baselines across single-agent, full-debate MAD, and selective MAD approaches. For single-agent methods, we include: (i) CoT (Wei et al. 2022); and (ii) SC (Wang et al. 2023), which runs CoT five times and selects the answer via majority voting. For full-debate MAD approaches that trigger debate for all instances, we consider: (iii) MAD (Liang et al. 2024), using agents with distinct personas in a three-agent setup; and (iv) GD (Liu et al. 2024), which clusters 5 agents into subgroups for parallel discussion, followed by 3 rounds of inter-group consensus voting. For selective MAD triggering, we evaluate: (v) DOWN (Eo et al. 2025). Unlike other baselines, DOWN requires labeled evaluation data to tune its threshold. For fair comparison, we use a threshold of 0.8 as reported to be most effective in the original paper.

Metrics. We evaluate each method using four key metrics: (i) total (input + output) token usage per data instance and (ii) answer accuracy, measuring answer correctness. In the technical report, we evaluate (iii) accuracy per 100k tokens (ApT), measuring token efficiency and (iv) per-question inference time, measuring computational efficiency.

Classifier Configurations. We use a lightweight MLP with six fully connected layers of 200 hidden units each, batch normalization, ReLU activations, and a dropout rate of 0.2. We train the classifier for 50 epochs using the Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.001 on standardized features via StandardScaler (de Amorim, Cavalcanti, and Cruz 2023). To support task-agnostic generalization, we use PubMedQA (Jin et al. 2019) and GQA (Hudson and Manning 2019) datasets for training, which are not included in the evaluation. We set the FocusCal loss hyperparameters as $\alpha_0 = 2.0$, $\alpha_1 = 1.0$, $\gamma = 2$, $\lambda = 6$, $\mu = 5$, and $B = 15$ bins for ECE. The debate threshold is set to $\tau = 0.7$. We chose these hyperparameters via grid search on a held-out validation set to ensure stable training and strong performance across all datasets.

Implementation Details. For LLM-based prompting, we use Gemini 2.0 Flash as the LLM agents for the primary results and also use GPT-5 nano and Qwen 3.0 (see online technical report for details and results). By default, we use a temperature of 0.0 to ensure deterministic outputs and a maximum of 512 tokens. We present the detailed prompt templates in the technical report. We conducted MLP training and inference using a single NVIDIA RTX 4090 GPU.

5.1 Main Results

We compare the token efficiency and accuracy of iMAD against single-agent, full-debate MAD, and selective MAD.

Advantage over Single-Agent and Full-Debate MAD. As shown in Table 3, iMAD achieves superior token efficiency while maintaining or improving accuracy over all single-agent and full-debate MAD baselines. While CoT uses the fewest tokens, iMAD achieves up to 13.5% higher accuracy. Compared to SC, iMAD drastically reduces token usage with consistently higher accuracy. For example, on MEDQA, iMAD reduces token cost by 62.7% while improving accuracy by 4.7%. Against full-debate MAD baselines, iMAD reduces token usage significantly while achieving comparable or higher accuracy. On MEDQA, iMAD uses 68% fewer tokens than MAD and 92% fewer than GD, while achieving the highest accuracy. A key to iMAD's

Dataset	Single-Agent				Full-Debate MAD				Selective MAD			
	CoT		SC		MAD		GD		DOWN		iMAD	
	Acc (%)	# Token	Acc (%)	# Token	Acc (%)	# Token	Acc (%)	# Token	Acc (%)	# Token	Acc (%)	# Token
MEDQA	76.6	653	77.3	3,482	<u>81.9</u>	4,034	80.2	16,832	79.2	1,161	**82.0**	1,300
MMLU	86.8	764	88.2	3,772	**89.5**	3,348	82.6	13,216	88.3	901	<u>89.2</u>	1,010
GSM8K	71.3	618	74.5	3,622	<u>76.4</u>	3,446	73.4	15,321	72.6	812	**84.8**	1,025
OKVQA	88.3	1,945	89.2	11,031	<u>89.8</u>	7,803	87.3	33,932	88.1	2,344	**90.3**	2,601
VQA-v2	77.5	2,245	77.6	14,013	81.0	8,796	**81.3**	36,091	78.6	3,262	**81.3**	3,489
ScienceQA	86.0	1,720	86.2	9,833	<u>89.4</u>	6,777	87.4	26,924	87.0	2,519	**90.8**	2,893

Table 3: Accuracy (Acc) and average token cost (# Token) per question comparison of iMAD and baselines across datasets. **Bold** values indicate the best result in each row, and <u>underlined</u> values indicate the second best.

Dataset	Skipped (%)				Triggered (%)			
	Good			Bad	Bad			Good
	✗→✗	✓→✓	✓→✗	✗→✓	✗→✗	✓→✓	✓→✗	✗→✓
MEDQA	10.5	66.3	4.9	4.8	1.0	3.7	1.7	7.1
MMLU	5.8	80.0	3.2	3.5	0.5	2.5	1.1	3.4
GSM8K	6.8	57.1	11.3	2.9	2.7	0.3	2.7	16.2
OKVQA	6.8	83.7	2.8	2.3	0.0	1.2	0.6	2.6
VQA-v2	12.9	69.7	5.3	4.9	0.4	2.1	0.4	4.3
ScienceQA	6.1	78.0	3.6	2.2	0.0	3.5	0.9	5.7

Table 4: Breakdown of iMAD debate decisions by whether MAD was skipped or triggered, and whether the decision was beneficial (Good) or harmful (Bad).

zero-shot efficiency is the classifier's strong generalization: we trained it on two representative datasets to capture diverse model reasoning behaviors, enabling it to learn generalizable model behaviors. iMAD's advantage over the full-debate strategy is especially evident on GSM8K, where iMAD outperforms MAD by 8.4% in accuracy. The only exception is MMLU, where MAD performs slightly better. In Table 4, the classifier skips debate in 3.5% of questions where it would help and 3.4% where a triggered debate fixes an error. This is because many MMLU questions are short and factual across diverse domains, wrong single-agent answers often sound fluent and confident, giving few hesitation cues and causing the classifier to miss some needed debates.

Advantage over Confidence-Based Selective MAD Triggering. As shown in Table 3, DOWN and iMAD incur comparable token costs, with DOWN using slightly fewer tokens by omitting self-critique and skipping some needed debates. However, both iMAD and MAD consistently achieve higher accuracy than DOWN, since DOWN cannot tune its confidence thresholds in the zero-shot setting. In contrast, iMAD incurs slightly more tokens due to self-critique prompting and more necessary debates, but this cost is justified by improved identification of recoverable errors. For example, on OKVQA, DOWN's accuracy remains near the single-agent baseline, revealing its inability to identify when debate is truly needed. This limitation arises because DOWN learns data-specific model behavior rather than generalizable behavior. In contrast, iMAD captures hesitation cues through a classifier trained on diverse features, enabling more accurate debate decisions and achieving higher accuracy with fewer tokens in zero-shot settings.

5.2 Breakdown of iMAD Debate Decisions

To evaluate the effectiveness of iMAD's selective debate triggering, we analyze its decision outcomes across all datasets in Table 4. For each instance, we precompute the single-agent and MAD outputs to establish whether MAD improves the answer, serving as the ground truth for evaluating decisions. We then measure how often iMAD matches the beneficial outcomes. Overall, up to 95.9% of iMAD's decisions are beneficial. When skipping debate, iMAD preserves correct answers (✓→✓) in 65-80% of cases and avoids wasted computation on unrecoverable errors (✗→✗) by up to 13%. When triggering debate, iMAD often recover incorrect answers (✗→✓). For example, iMAD successfully flips 16.2% of cases on GSM8K and 7.1% on MEDQA, approaching their respective upper bounds of 19.1% and 11.9% (see Table 2). Crucially, harmful decisions, such as overturning correct answers (✓→✗) or incurring unnecessary debate overhead (✗→✗ and ✓→✓), remain consistently low (around 5-10%). These results highlight iMAD's ability to selectively trigger MAD only when it is likely to improve accuracy, while avoiding unnecessary token costs.

6 Conclusion

We presented iMAD, a token-efficient MAD framework that triggers debates only when it is likely to improve outcomes. It combines structured self-critique prompting with a lightweight debate decision classifier trained with FocusCal loss, enabling effective debate decisions based on single-agent responses without using evaluation data. Compared with five baselines, iMAD reduces token usage by up to 92% while improving accuracy by up to 13.5%. These results show that iMAD is a practical and scalable solution for collaborative reasoning in agentic LLM systems. Future work includes exploring adaptive or online learning approaches to reduce labeling costs during classifier training and further improve generalization, as discussed in the technical report.

Acknowledgments

This research was supported in part by NSF grant CNS-2315851, the Commonwealth Cyber Initiative (CCI), and a Virginia Tech Presidential Postdoctoral Fellowship.

References

Chen, J.; Saha, S.; and Bansal, M. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7066–7085. Bangkok, Thailand: Association for Computational Linguistics.

Chen, P.; Han, B.; and Zhang, S. 2024. CoMM: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving. In *NAACL-HLT*.

Chen, S.; Zeng, L.; Raghunathan, A.; Huang, F.; and Kim, T. C. 2024. Moa is all you need: Building llm research team using mixture of agents. *arXiv preprint arXiv:2409.07487*.

Chen, W.; Yuan, J.; Qian, C.; Yang, C.; Liu, Z.; and Sun, M. 2025. Optima: Optimizing Effectiveness and Efficiency for LLM-Based Multi-Agent System. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 11534–11557. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

de Amorim, L. B.; Cavalcanti, G. D.; and Cruz, R. M. 2023. The choice of scaling technique matters for classification performance. *Applied Soft Computing*, 133: 109924.

Eo, S.; Moon, H.; Zi, E. H.; Park, C.; and Lim, H. 2025. Debate Only When Necessary: Adaptive Multiagent Collaboration for Efficient LLM Reasoning. arXiv:2504.05047.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hsu, C.-J.; Buffelli, D.; McGowan, J.; Liao, F.-T.; Chen, Y.-C.; Vakili, S.; and shan Shiu, D. 2025. Group Think: Multiple Concurrent Reasoning Agents Collaborating at Token Level Granularity. arXiv:2505.11107.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6693–6702.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, W.; Lin, Y.; Xia, M.; and Jin, C. 2024a. Rethinking Mixture-of-Agents: Is Mixing Different Large Language Models Beneficial? In *Language Gamification - NeurIPS 2024 Workshop*.

Li, Y.; Yuan, P.; Feng, S.; Pan, B.; Wang, X.; Sun, B.; Wang, H.; and Li, K. 2024b. Escape Sky-high Cost: Early-stopping Self-Consistency for Multi-step Reasoning. In *The Twelfth International Conference on Learning Representations*.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904. Miami, Florida, USA: Association for Computational Linguistics.

Liu, T.; Wang, X.; Huang, W.; Xu, W.; Zeng, Y.; Jiang, L.; Yang, H.; and Li, J. 2024. GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion. arXiv:2409.14051.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Pan, M. Z.; Cemri, M.; Agrawal, L. A.; Yang, S.; Chopra, B.; Tiwari, R.; Keutzer, K.; Parameswaran, A.; Ramchandran, K.; Klein, D.; Gonzalez, J. E.; Zaharia, M.; and Stoica, I. 2025. Why Do Multiagent Systems Fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Srivastava, G.; Bi, Z.; Lu, M.; and Wang, X. 2025. DEBATE, TRAIN, EVOLVE: Self-Evolution of Language Model Reasoning. In Christodoulopoulos, C.; Chakraborty, T.; Rose,

C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32752–32798. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Tillmann, A. 2025. Literature Review Of Multi-Agent Debate For Problem-Solving. arXiv:2506.00066.

Wan, X.; Sun, R.; Dai, H.; Arik, S.; and Pfister, T. 2023. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 3493–3514. Toronto, Canada: Association for Computational Linguistics.

Wang, H.; Prasad, A.; Stengel-Eskin, E.; and Bansal, M. 2024a. Soft Self-Consistency Improves Language Models Agents. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 287–301. Bangkok, Thailand: Association for Computational Linguistics.

Wang, J.; WANG, J.; Athiwaratkun, B.; Zhang, C.; and Zou, J. 2025. Mixture-of-Agents Enhances Large Language Model Capabilities. In *The Thirteenth International Conference on Learning Representations*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2024b. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 257–279. Mexico City, Mexico: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E. E.; Jiang, L.; Zhang, X.; Zhang, S.; Awadallah, A.; White, R. W.; Burger, D.; and Wang, C. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *COLM 2024*.

Yang, D.; Tsai, Y.-H. H.; and Yamada, M. 2025. On Verbalized Confidence Scores for LLMs. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Yang, R.; Rajagopal, D.; Hayati, S. A.; Hu, B.; and Kang, D. 2024. Confidence Calibration and Rationalization for LLMs via Multi-Agent Deliberation. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Zhang, H.; Cui, Z.; Chen, J.; Wang, X.; Zhang, Q.; Wang, Z.; Wu, D.; and Hu, S. 2025. Stop Overvaluing Multi-Agent Debate – We Must Rethink Evaluation and Embrace Model Heterogeneity. arXiv:2502.08788.

Zhang, Y.; Yang, X.; Feng, S.; Wang, D.; Zhang, Y.; and Song, K. 2024. Can LLMs Beat Humans in Debating? A Dynamic Multi-agent Framework for Competitive Debate. *CoRR*, abs/2408.04472.