UNDERSTANDING METACOGNITION IN MULTI-AGENT LLMS: ROUTING, NOT REASONING

Mafizur Rahman
CREDIT Center, Department of ECE
Prairie View A&M University
Prairie View, TX 77446, USA

Lijun Qian
CREDIT Center, Department of ECE
Prairie View A&M University
Prairie View, TX 77446, USA

ABSTRACT

Multi-agent reasoning and metacognitive strategies are widely employed to enhance large language model (LLM) performance; however, the functional role of metacognition remains unclear. Most existing approaches implicitly treat metacognition as a mechanism for generating improved or more diverse reasoning. In this work, we argue that metacognition primarily acts as an information routing and compression mechanism rather than a generator of new reasoning content. We introduce MC-MAS, an inference-time framework that separates problem solving from metacognitive arbitration, where independent solvers propose candidate answers and a metacognitive arbiter critiques and consolidates these outputs. Using routing-centric metrics such as semantic novelty, entropy reduction, and overconfidence errors, we analyze information flow across four reasoning benchmarks and multiple model settings. Results show that MC-MAS yields limited novelty but consistently reduces uncertainty and overconfident errors, with accuracy gains that are modest and dataset-dependent. We also find that reliable improvements arise from structured arbitration rather than reflection or increased sampling alone.

1 INTRODUCTION

Large language models (LLMs) are increasingly relying on multi-step inference strategies, such as self-consistency (Wan et al., 2025; Mitchell et al., 2022), majority voting (Chen et al., 2024a), and multi-agent ensembles (Chowdhury et al., 2025), to improve their robustness on reasoning tasks. While these approaches often yield accuracy gains, they implicitly assume that performance improvements arise from generating better reasoning traces or from sampling more diverse answers (Wei et al., 2022; Yao et al., 2022; Zhou et al., 2022; Wang et al., 2022). As a result, it remains unclear what role metacognition actually plays in multi-agent systems: *does it create new information, or does it primarily organize, filter, and route existing reasoning signals?*

Despite the growing adoption of metacognitive and multi-agent reasoning strategies, their functional role remains poorly characterized. Many recent methods report performance improvements without disentangling whether these gains stem from stronger reasoning, increased aggregation of sampled answers, or implicit uncertainty filtering effects (Da et al., 2025; Jin et al., 2025). This ambiguity makes it difficult to determine when metacognitive mechanisms are truly necessary, how they should be designed, or when their additional computational cost is justified. As multi-agent systems are increasingly deployed in settings where reliability and calibration matter as much as raw accuracy, a clearer understanding of what metacognition actually contributes becomes essential.

In this work, we argue that metacognition in LLMs is best understood as an *information routing mechanism* rather than a generator of new reasoning content. To test this hypothesis, we introduce **MC-MAS** (**M**etacognitive **M**ulti-**A**gent **S**ystem), a lightweight inference-time framework that separates problem solving from metacognitive arbitration. MC-MAS consists of two independent solvers that propose candidate solutions, followed by a metacognitive stage that critiques, compares, and arbitrates among these proposals. Crucially, the arbiter does not introduce new task-specific knowledge or external supervision; instead, it primarily restructures and consolidates information already present in solver outputs, without reliably increasing semantic novelty.

Unlike prior ensemble or self-consistency methods (Chowdhury et al., 2025; Wan et al., 2025; Mitchell et al., 2022), MC-MAS is explicitly designed to expose and measure information flow across reasoning stages. We develop a set of routing-centric metrics, including semantic novelty, redundancy, entropy reduction, compression ratio, and overconfidence errors, to quantify how information evolves from solvers to the final decision. This facilitates analysis beyond accuracy-only evaluations and allows us to directly assess whether metacognition synthesizes new evidence or merely redistributes existing signals. Our results indicate that a substantial portion of previously reported "reasoning gain" in multi-agent LLM systems may be explained by uncertainty compression through structured aggregation, rather than enhanced reasoning capability.

Across four diverse reasoning benchmarks: BoolQ (Clark et al., 2019), ARC-Easy (Clark et al., 2018), StrategyQA (Geva et al., 2021), and GSM8K (Cobbe et al., 2021)—and multiple model configurations (Mistral-7B-Instruct-v0.3 and LLaMA-3.1-8B-Instruct, both full-precision and 4-bit), we observe a consistent pattern: MC-MAS rarely increases semantic novelty, but reliably reduces answer entropy and improves confidence calibration. In many cases, accuracy improvements are modest or absent, yet entropy decreases and overconfidence errors are reduced, indicating improved decision reliability even when raw performance remains unchanged. Importantly, these trends persist under aggressive quantization, demonstrating that metacognitive routing remains effective even when generative capacity is constrained. Overall, this work makes three key contributions:

- **Conceptual:** We reframe metacognition in LLMs as an information routing and compression mechanism rather than a source of new reasoning.
- **Methodological:** We introduce MC-MAS, a simple yet interpretable multi-agent framework that cleanly separates solving, reflection, and arbitration.
- **Empirical:** Through detailed routing-level analysis, we demonstrate that metacognitive systems primarily improve confidence calibration, stability, and reliability, with accuracy gains emerging only when solver disagreement is meaningful.

By shifting the focus from how many answers are sampled to how information is structured and consolidated, this work provides a principled foundation for analyzing and designing future metacognitive and multi-agent reasoning systems.

2 Related Works

Multi-Agent and Self-Consistency Reasoning. Recent work has shown that aggregating multiple reasoning trajectories can improve LLM performance on complex tasks (Ge et al., 2023; Yuan et al., 2024; Yang et al., 2025; Wang et al., 2025b). Techniques such as self-consistency (Wan et al., 2025; Tan et al., 2024), majority voting (Chen et al., 2024b; Zhuge et al., 2024), and multi-sample ensembling (Chowdhury et al., 2025) leverage stochastic decoding to lower variance and mitigate individual reasoning errors. Extensions include debate-based systems (Ozaki et al., 2025), multi-agent collaboration (Zhang et al., 2025b), and role-based prompting (Louatouate & Zeriouh, 2025), where agents critique or refine one another's outputs (Yu et al., 2025; Wang et al., 2025c). While these methods often yield accuracy gains, they are typically evaluated only in terms of final performance and rely on implicit assumptions that improvements arise from better reasoning traces or increased diversity Wei et al. (2022); Yao et al. (2022); Zhou et al. (2022); Wang et al. (2022). In contrast, they offer limited insight into how information evolves across reasoning stages or whether additional components genuinely contribute new task-relevant information.

Metacognition, Reflection, and Calibration in LLM. A parallel line of work explores reflection (Zhang et al., 2025c), critique (Lan et al., 2024), and self-evaluation (Qu et al., 2024; Wang et al., 2025a; Mehandru et al., 2024) as mechanisms for improving reasoning quality and robustness. Prior studies show that reflective prompts can increase factual consistency (Laban et al., 2023), surface errors, or enhance calibration (Tyen et al., 2024; Song et al., 2025) under certain settings. However, reflection is often conflated with improved reasoning (Jiang et al., 2025), and its effects are inconsistently linked to accuracy or reliability. Moreover, most techniques treat metacognition as an auxiliary reasoning step rather than an analyzable process (Fan et al., 2025; Qian et al., 2025). Our work differs by explicitly modeling metacognition as an information routing and compression layer and by introducing routing-centric metrics to directly measure its functional role. This allows us to

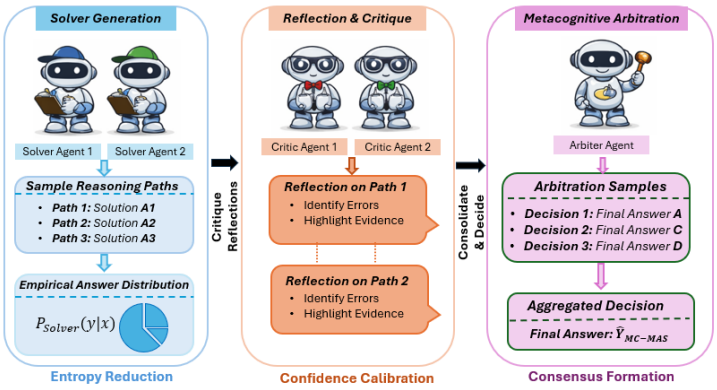

Figure 1: Overview of MC-MAS, illustrating how sampled solver outputs are routed through critique and arbitration stages to transform an empirical answer distribution into a calibrated consensus decision.

disentangle novelty generation from uncertainty reduction and to show that reliable gains emerge primarily from structured arbitration rather than reflection alone.

3 Multi-Agent Metacognitive Framework: MC-MAS

3.1 Problem Setup

Let x denote an input instance (e.g., a question or problem description) drawn from a dataset $\mathcal{D}$, and let $y \in \mathcal{Y}$ be the corresponding ground-truth answer. Given a pretrained large language model (LLM) f_θ, our goal is to improve decision reliability at inference time without modifying model parameters or introducing external supervision.

We consider the setting where multiple reasoning trajectories can be sampled from f_θ due to stochastic decoding. Existing approaches such as self-consistency or majority voting aggregate these samples directly, implicitly assuming that performance gains arise from generating better reasoning traces. In contrast, we explicitly separate *problem solving* from *metacognitive arbitration* and study how information flows between these stages.

3.2 Overview of MC-MAS

The Metacognitive Multi-Agent System (MC-MAS) consists of three stages: (i) independent solver generation, (ii) reflective critique, and (iii) metacognitive arbitration. A schematic overview is shown in Figure 1. Two independent solver agents generate multiple sampled reasoning paths, forming an empirical answer distribution. Critic agents then reflect on solver outputs to identify errors and highlight supporting evidence, without directly selecting an answer. A metacognitive arbiter performs multi-sample arbitration over solver outputs and critiques, consolidating information to produce a final consensus decision. This process primarily enables entropy reduction, confidence calibration, and consensus formation through structured information routing rather than generating new task-specific reasoning. Fundamentally, all components share the same underlying LLM f_θ; MC-MAS introduces no additional task-specific knowledge and operates purely at inference time.

3.3 Solver Generation

Given input x, we instantiate two independent solver agents that generate sets of stochastic reasoning trajectories $\mathcal{S}_i$ from f_θ:

$$\mathcal{S}_i = \{s_i^{(k)}\}_{k=1}^{K_s}, \quad s_i^{(k)} \sim p_\theta(\cdot \mid x), \quad i \in \{1, 2\}, \tag{1}$$

where K_s is the number of solver samples per agent.

Each solver sample produces both a textual reasoning trace and a final answer $a_i^{(k)} \in \mathcal{Y}$. The union of all solver answers induces an empirical answer distribution:

$$p_{\text{solver}}(y \mid x) = \frac{1}{2K_s} \sum_{i=1}^{2} \sum_{k=1}^{K_s} \mathbb{I}[a_i^{(k)} = y]. \tag{2}$$

This distribution captures the initial uncertainty and disagreement among solvers.

3.4 Reflection and Critique

For each solver, we generate a reflective critique that evaluates the solver's reasoning without modifying its answer. Let $\tilde{s}_i$ denote a representative solver trace (e.g., the first sample) from solver i. A reflection is generated as:

$$r_i \sim p_\theta(\cdot \mid x, \tilde{s}_i), \tag{3}$$

where the reflection prompt instructs the model to identify errors, gaps, or alternative interpretations.

Reflections are treated as *auxiliary semantic signals*. Importantly, they do not directly determine the final answer and are not required to agree with the solver. Their role is to expose latent inconsistencies and highlight relevant evidence for downstream arbitration.

3.5 Metacognitive Arbitration

The arbitration stage integrates solver outputs and reflections to produce a final decision. We sample K_a arbitration responses:

$$a^{(m)} \sim p_\theta(\cdot \mid x, \tilde{s}_1, \tilde{s}_2, r_1, r_2), \quad m = 1, \ldots, K_a. \tag{4}$$

Here, $a^{(m)} \in \mathcal{Y}$ denotes the discrete final answer extracted from the m-th arbitration response.

These samples define the arbitration-induced answer distribution:

$$p_{\text{arb}}(y \mid x) = \frac{1}{K_a} \sum_{m=1}^{K_a} \mathbb{I}[a^{(m)} = y]. \tag{5}$$

The final MC-MAS prediction is obtained by majority vote:

$$\hat{y}_{\text{MC-MAS}} = \arg\max_{y \in \mathcal{Y}} p_{\text{arb}}(y \mid x). \tag{6}$$

We analyze the computational implications of increasing K_a, along with scaling behavior and test-time complexity, in Appendix A.4 and Appendix A.5.

3.6 Uncertainty and Information Measures

To analyze information flow, we compute entropy before and after arbitration:

$$H_{\text{solver}}(x) = -\sum_y p_{\text{solver}}(y \mid x) \log p_{\text{solver}}(y \mid x), \tag{7}$$

$$H_{\text{arb}}(x) = -\sum_y p_{\text{arb}}(y \mid x) \log p_{\text{arb}}(y \mid x). \tag{8}$$

Entropy reduction,

$$\Delta H(x) = H_{\text{solver}}(x) - H_{\text{arb}}(x), \tag{9}$$

quantifies uncertainty compression induced by metacognitive arbitration.

We additionally measure semantic novelty and redundancy by embedding solver and arbiter claims and computing pairwise cosine similarity. These metrics allow us to distinguish between *information generation* (high novelty) and *information consolidation* (entropy reduction with low novelty).

3.7 INTERPRETATION

Under this formulation, MC-MAS does not aim to maximize novelty. Instead, it seeks to reduce decision uncertainty by routing, filtering, and consolidating existing solver information. Multi-sample arbitration is essential: as $K_a \to 1$, arbitration becomes unstable, while larger K_a enables consensus formation and reliable entropy compression. This formalization directly motivates our empirical evaluation and ablation studies.

Relation to Majority Voting and Self-Consistency. Although MC-MAS ultimately produces a consensus prediction, it differs fundamentally from majority voting and self-consistency. Majority voting aggregates solver outputs directly by treating all samples as equally informative and lacking an explicit mechanism for uncertainty regulation (Yang et al., 2024). In contrast, MC-MAS performs arbitration over both solver outputs and reflective critiques, allowing selective consolidation based on semantic consistency and confidence. As a result, MC-MAS can reduce entropy and overconfidence even when the majority answer is incorrect, as shown in Table 4. This distinction highlights that MC-MAS operates as an information routing mechanism rather than a simple aggregation strategy.

4 INFORMATION ROUTING AND CALIBRATION METRICS

Although semantic novelty, redundancy, and entropy are not conventional task-level metrics, they are principled information-theoretic quantities used to analyze distributions and uncertainty in probabilistic and neural models (Guo et al., 2017; Farquhar et al., 2024). We use them not as substitutes for accuracy, but to expose how information flows, compresses, and stabilizes across multi-agent reasoning stages—phenomena that accuracy alone cannot capture.

Answer Entropy. Answer entropy calculates the uncertainty of a model's decision based on the empirical distribution of sampled answers. Given a set of answers $\{y_1, \ldots, y_N\}$ produced by solvers or by the arbiter, we define

$$H(Y) = -\sum_{c \in \mathcal{C}} p(c) \log p(c), \tag{10}$$

where $p(c)$ is the fraction of samples predicting class c. Lower entropy indicates higher consensus and confidence. In MC-MAS, entropy reduction reflects effective consolidation of solver disagreement rather than increased sampling.

Entropy Reduction. To quantify uncertainty compression across stages, we compute the entropy decrease between solver outputs and the arbiter decision:

$$\Delta H = H_{\text{solver}} - H_{\text{arb}}. \tag{11}$$

Positive ΔH indicates successful uncertainty compression, while values near zero imply that arbitration does not alter an already confident decision. Negative values correspond to low-uncertainty regimes where further compression is unnecessary.

Semantic Novelty. Semantic novelty measures how much new semantic content is introduced by a metacognitive stage relative to solver outputs. We extract sentence-level claims from model outputs and compute their embedding similarity using a pretrained sentence encoder. Given final-stage claims $\mathcal{F}$ and solver claims $\mathcal{S}$, novelty is defined as

$$\text{Novelty} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{I}\left[\max_{s \in \mathcal{S}} \text{sim}(f, s) < \tau\right], \tag{12}$$

where τ is a similarity threshold. High novelty indicates the introduction of new semantic signals, while low novelty suggests information reuse or reorganization.

Redundancy. Redundancy captures semantic overlap within the combined set of claims produced during inference. It is computed as the average pairwise cosine similarity among all extracted claims. Lower redundancy indicates more diverse information, while higher redundancy reflects repetition or agreement among reasoning paths.

Compression Ratio. We define a compression ratio to summarize the degree of uncertainty consolidation:

$$\text{Compression} = 1 - \frac{H_{\text{arb}}}{H_{\text{solver}}}. \tag{13}$$

Higher values correspond to stronger information compression achieved by metacognitive arbitration.

Overconfident Errors. Overconfident errors measure calibration failures where incorrect predictions are made with high confidence. An error is counted as overconfident if the predicted class probability exceeds a fixed threshold (0.7 in our experiments) while being incorrect. Lower counts indicate improved confidence calibration and reliability.

4.1 Experimental Configuration

All experiments are conducted in a zero-shot setting: models receive only task instructions and inputs, without any in-context examples, fine-tuning, or external supervision. We evaluate publicly available instruction-tuned language models, including Mistral-7B-Instruct-v0.3 and LLaMA-3.1-8B-Instruct, under both full-precision and 4-bit quantized inference using NF4 quantization. Inference is performed locally using HuggingFace Transformers with deterministic decoding seeds fixed across 3 runs. For each input, two independent solver instances of the same backbone model generate 5 stochastic samples each using temperature 0.6, without sharing intermediate states or information. Reflection stages use temperature 0.2, while metacognitive arbitration uses temperature 0.4 with 50 arbitration samples unless otherwise specified. All solver, reflection, and arbitration agents share the same underlying backbone language model. Experiments are conducted on BoolQ, ARC-Easy, StrategyQA, and GSM8K, with 200 randomly selected validation examples per dataset. See Appendix A.3 for dataset and task details. All experiments were conducted on a single NVIDIA A800 GPU with 80GB of memory, enabling local inference for both full-precision and 4-bit quantized models without requiring distributed execution. See the prompt templates and pseudocode in Appendix A.1.

5 Results Analysis

Comparison with the Strongest Individual Solver. Table 1 compares MC-MAS against the strongest individual solver across multiple model families, precision settings, and reasoning benchmarks. On simpler tasks such as BoolQ and ARC-Easy, MC-MAS closely matches the best solver, with performance differences typically within ± 0.1 percentage points and well inside run-to-run variance. This suggests that metacognitive arbitration does not degrade performance when solver agreement is already high and the decision boundary is clear. In contrast, on more reasoning-intensive benchmarks, StrategyQA and GSM8K, MC-MAS consistently yields measurable improvements of approximately $+1.7$ to $+1.9$ points over the best solver. These gains are observed under 4-bit quantization and remain stable across independent runs, indicating that MC-MAS effectively consolidates solver disagreement without relying on increased model capacity or higher-precision representations. Notably, the absence of improvements on simpler datasets, combined with consistent gains on harder benchmarks, suggests that MC-MAS does not function as a generic accuracy booster. Instead, its benefits emerge specifically in regimes where solver uncertainty and disagreement are substantial. All these results support the interpretation that MC-MAS primarily improves decision consolidation and reliability, while preserving strong baseline performance under both full-precision and quantized inference settings.

Information Routing and Compression Analysis. To understand how metacognitive arbitration transforms solver outputs beyond accuracy, we analyze routing-level information metrics. Table 2 reports these metrics for *the same MC-MAS framework instantiated with different backbone language models and precision settings*. Across all configurations, MC-MAS introduces modest semantic novelty while consistently reducing redundancy. Entropy generally decreases, indicating effective consolidation of solver disagreement, particularly on ARC-Easy with Mistral-7B. In a few cases, entropy changes are near zero or slightly negative, reflecting low-uncertainty regimes where further compression is neither necessary nor beneficial. Overall, these trends support the view that

Base Model	Prec.	Dataset	Best Solver Acc (%)	MC-MAS Acc(%)	Δ
Mistral-7B	4-bit	BoolQ	$90.1_{\pm 0.4}$	$\mathbf{90.2}_{\pm \mathbf{0.3}}$	$+0.1$
Mistral-7B	Full	BoolQ	$86.2_{\pm 0.5}$	$\mathbf{86.1}_{\pm \mathbf{0.4}}$	-0.1
Mistral-7B	4-bit	ARC-Easy	$84.3_{\pm 0.6}$	$\mathbf{84.4}_{\pm \mathbf{0.5}}$	$+0.1$
Mistral-7B	Full	ARC-Easy	$80.4_{\pm 0.7}$	$\mathbf{82.1}_{\pm \mathbf{0.6}}$	$+1.7$
LLaMA-3.1-8B	4-bit	BoolQ	$88.2_{\pm 0.4}$	$\mathbf{88.3}_{\pm \mathbf{0.3}}$	$+0.1$
LLaMA-3.1-8B	4-bit	ARC-Easy	$87.9_{\pm 0.6}$	$\mathbf{88.0}_{\pm \mathbf{0.5}}$	$+0.1$
LLaMA-3.1-8B	4-bit	StrategyQA	$66.1_{\pm 0.8}$	$\mathbf{68.0}_{\pm \mathbf{0.7}}$	$+1.9$
LLaMA-3.1-8B	4-bit	GSM8K	$78.3_{\pm 0.7}$	$\mathbf{80.2}_{\pm \mathbf{0.6}}$	$+1.9$

Table 1: Comparison between the strongest individual solver and MC-MAS across base models, precisions, and datasets. MC-MAS consistently matches the best solver on simpler tasks (BoolQ, ARC-Easy), while yielding consistent improvements on reasoning-intensive benchmarks (StrategyQA, GSM8K), without degrading performance under 4-bit quantization.

MC-MAS functions primarily as an information routing and compression mechanism rather than a generator of new reasoning content.

Base Model	Dataset	Novelty(%) ↑	Redundancy(%) ↓	Entropy ↓	Compression
Mistral-7B (4-bit)	BoolQ	$9.7_{\pm 0.6}$	$44.7_{\pm 1.2}$	$0.046_{\pm 0.018}$	$0.271_{\pm 0.031}$
Mistral-7B (Full)	BoolQ	$20.0_{\pm 0.9}$	$43.6_{\pm 1.0}$	$0.017_{\pm 0.021}$	$0.477_{\pm 0.042}$
Mistral-7B (4-bit)	ARC-Easy	$49.1_{\pm 1.3}$	$43.7_{\pm 1.1}$	$0.227_{\pm 0.032}$	$0.548_{\pm 0.038}$
Mistral-7B (Full)	ARC-Easy	$51.2_{\pm 1.5}$	$42.2_{\pm 1.0}$	$-0.007^{\dagger}_{\pm 0.028}$	$0.575_{\pm 0.041}$
LLaMA-3.1-8B (4-bit)	BoolQ	$2.0_{\pm 0.4}$	$38.2_{\pm 0.9}$	$0.036_{\pm 0.019}$	$0.027_{\pm 0.014}$
LLaMA-3.1-8B (4-bit)	ARC-Easy	$12.8_{\pm 0.8}$	$39.6_{\pm 1.1}$	$-0.065^{\dagger}_{\pm 0.034}$	$0.187_{\pm 0.026}$

Table 2: Information routing and compression metrics for MC-MAS. MC-MAS introduces limited semantic novelty while consistently reducing redundancy and entropy. Negative entropy changes (†) indicate low-uncertainty regimes where arbitration does not further compress solver disagreement.

Semantic vs. Symbolic Information Routing. We next examine the relative contribution of semantic and symbolic information routing within MC-MAS across reasoning benchmarks. Table 3 shows that GSM8K shows a balanced reliance on semantic and symbolic signals across both Mistral-7B and LLaMA-3.1-8B, indicating that metacognitive arbitration integrates both numerical structure and contextual reasoning when consolidating solver outputs. In contrast, StrategyQA is overwhelmingly dominated by semantic routing, with negligible symbolic contribution, reflecting the open-ended and knowledge-driven nature of the task. These results suggest that MC-MAS adapts its routing behavior to the structure of the underlying problem, prioritizing symbolic consistency when explicit formal structure is present (GSM8K), and favoring semantic coherence when reasoning is primarily contextual (StrategyQA). Importantly, MC-MAS does not impose a fixed reasoning bias, but instead flexibly reallocates information flow based on task demands.

Base Model	Dataset	Semantic (%)	Symbolic (%)
Mistral-7B (4-bit)	GSM8K	$6.5_{\pm 0.6}$	$9.6_{\pm 0.7}$
Mistral-7B (Full)	GSM8K	$11.2_{\pm 0.8}$	$10.4_{\pm 0.6}$
LLaMA-3.1-8B (4-bit)	GSM8K	$6.5_{\pm 0.5}$	$6.3_{\pm 0.6}$
LLaMA-3.1-8B (4-bit)	StrategyQA	$69.8_{\pm 1.4}$	$0.4_{\pm 0.2}$

Table 3: Semantic versus symbolic contribution of MC-MAS across reasoning benchmarks. Values denote the % of routed claims attributed to semantic or symbolic information. GSM8K exhibits a balanced reliance on semantic and symbolic reasoning, while StrategyQA is dominated by semantic information routing, with minimal symbolic structure exploited by the models.

Overconfidence Analysis Finally, we report the number of overconfident errors, defined as incorrect predictions made with high confidence (Table 4). Across all evaluated datasets, MC-MAS consistently reduces or maintains the number of such errors relative to the strongest individual solver. The reduction is particularly evident on BoolQ and StrategyQA, suggesting that metacognitive arbitration helps suppress confidently incorrect decisions by consolidating conflicting solver evidence.

Base Model	Dataset	Solver ↓	MC-MAS ↓
Mistral-7B (4-bit)	BoolQ	$6.2_{\pm 0.8}$	$\mathbf{5.1}_{\pm \mathbf{0.6}}$
Mistral-7B (4-bit)	ARC-Easy	$8.1_{\pm 1.0}$	$\mathbf{7.9}_{\pm \mathbf{0.9}}$
LLaMA-3.1-8B (4-bit)	GSM8K	$2.3_{\pm 0.5}$	$\mathbf{2.1}_{\pm \mathbf{0.4}}$
LLaMA-3.1-8B (4-bit)	StrategyQA	$15.4_{\pm 1.3}$	$\mathbf{14.9}_{\pm \mathbf{1.1}}$

Table 4: Overconfident error counts (lower is better), measured as incorrect predictions made with high confidence. MC-MAS consistently reduces or maintains overconfidence relative to individual solvers, indicating improved calibration without sacrificing accuracy.

Variant	Accuracy (%)	Novelty ↑	Entropy ↓
Majority Vote	$88.1_{\pm 0.6}$	0.0	–
+ Reflections	$88.3_{\pm 0.5}$	$18.1_{\pm 1.2}$	–
+ Single Arbiter	$86.2_{\pm 0.7}$	$15.3_{\pm 1.0}$	$0.051_{\pm 0.006}$
Full MC-MAS	$\mathbf{90.0}_{\pm \mathbf{0.4}}$	$13.8_{\pm 0.9}$	$\mathbf{0.015}_{\pm \mathbf{0.004}}$

Table 5: The contribution of metacognitive components.

Even in cases where accuracy gains are modest, the lower overconfidence counts suggest improved calibration and decision reliability. These results indicate that MC-MAS enhances robustness not by inflating confidence, but by selectively filtering uncertain or misleading reasoning signals.

Ablation on the Role of Metacognitive Components. Table 5 analyzes how individual metacognitive components influence accuracy and uncertainty regulation. Majority voting provides a strong baseline but, by construction, introduces no semantic novelty and offers no mechanism for uncertainty compression. Adding reflections substantially increases novelty, confirming that critique stages introduce additional semantic signals; however, these signals remain weakly coupled to final decision-making and therefore do not improve accuracy or entropy. Introducing a single-shot arbiter enables entropy estimation but results in degraded accuracy and higher uncertainty, indicating that one-sample arbitration is insufficient to reliably resolve solver disagreement. In contrast, full MC-MAS achieves the highest accuracy while yielding the lowest entropy, demonstrating effective uncertainty compression through multi-sample consensus formation. Notably, novelty decreases from the reflection-only variant to full MC-MAS, indicating that performance gains arise from structured information consolidation rather than from generating additional reasoning content. Overall, this ablation confirms that MC-MAS improvements stem from principled information routing and aggregation across agents, rather than from isolated reflection or increased sampling alone.

Ablation on Arbitration Sample Size. Table 6 shows that increasing the arbitration sample size K_a improves MC-MAS accuracy while reducing entropy, indicating more stable consensus formation. Performance gains saturate beyond $K_a \approx 25$, where entropy stabilizes and accuracy improves marginally, suggesting diminishing returns rather than degraded uncertainty handling. Overconfident errors remain comparable across larger K_a, implying improved confidence estimation rather than induced overconfidence. Overall, moderate arbitration sizes ($K_a \approx 25$) offer the best trade-off between uncertainty compression and computational cost.

5.1 Discussion

While earlier work often attributes performance gains to richer reasoning traces or increased sampling diversity, our results point to a more constrained and mechanistic explanation. As shown in Table 1, MC-MAS yields modest and dataset-dependent accuracy gains, frequently matching rather than exceeding the strongest individual solver on simpler tasks. In contrast, MC-MAS consistently improves reliability-related metrics, including reduced answer entropy (Table 2) and fewer overconfident errors (Table 4). These findings suggest that metacognition primarily reorganizes and consolidates existing solver outputs rather than generating substantial new task-specific knowledge.

A key insight emerges from contrasting novelty-related metrics with downstream decision quality. Table 2 shows that reflection stages increase semantic novelty across models and datasets. However, these gains do not translate into improved accuracy or entropy reduction, as shown by the ablation results in Table 5. While reflection increases intermediate semantic novelty, the full MC-MAS

K_a	#LLM	Acc. (%)	Entropy ↓	OverConf ↓
1	13	$82.0_{\pm 0.8}$	$0.062_{\pm 0.006}$	3
5	17	$86.2_{\pm 0.7}$	$0.036_{\pm 0.005}$	7
25	17	$87.1_{\pm 0.3}$	$0.026_{\pm 0.005}$	6
50	62	$88.3_{\pm 0.6}$	$0.049_{\pm 0.007}$	7

Table 6: Effect of arbiter sample size (K_a) on MC-MAS performance and uncertainty. #LLM denotes the number of LLM forward calls; multiple arbitration samples may be drawn from a single invocation when batching is possible.

pipeline preserves such novelty only when it improves decision reliability, resulting in limited net novelty at the system level.

Entropy and compression analyses further support a routing-based interpretation of metacognition. As reported in Table 2, MC-MAS generally reduces answer entropy and redundancy, especially on ARC-Easy with Mistral-7B, indicating effective consolidation of solver disagreement. In several low-uncertainty regimes, entropy changes are near zero or slightly negative, indicating that the arbiter refrains from unnecessary compression when solvers already agree. This selective behavior contrasts sharply with majority voting, which lacks an explicit uncertainty regulation mechanism and can propagate overconfident errors, as evidenced by the higher error counts in Table 4. Importantly, these trends persist under aggressive quantization. Accuracy comparisons in Table 1 show that 4-bit models maintain performance comparable to their full-precision counterparts. At the same time, routing, entropy, and calibration metrics in Tables 2 and 4 remain largely unchanged under 4-bit precision. This robustness suggests that metacognitive routing relies more on structural comparison and aggregation of solver outputs than on high-fidelity generation, making MC-MAS particularly well suited for resource-constrained deployment scenarios. Overall, our results show that metacognition in LLM systems primarily functions as a mechanism for filtering, weighting, and consolidating information rather than generating new reasoning content. These findings align with the toy information-theoretic view in Appendix A.2, which interprets metacognitive arbitration as an information bottleneck that compresses solver uncertainty while preserving task-relevant signals. This perspective shifts the design focus of multi-agent systems away from increasingly elaborate reasoning traces and toward structured integration of existing solver outputs.

Design Principle: When to Use Metacognitive Routing
MC-MAS is most effective when solver disagreement reflects meaningful epistemic uncertainty rather than random noise or shared bias. When solver entropy is high and disagreement is structured, metacognitive arbitration compresses uncertainty and improves confidence calibration, as measured by reduced overconfident errors. When solver entropy is already low or disagreement is uninformative, arbitration yields limited benefit. This principle motivates adaptive arbitration strategies conditioned on observed solver uncertainty.

Broader Implications for Existing Agentic Methods. Although our analysis centers on MC-MAS, the routing-based interpretation of metacognition extends naturally to a broad class of existing agentic and multi-sample reasoning techniques. Methods such as self-consistency (Wan et al., 2025; Tan et al., 2024; Mitchell et al., 2022), and majority voting (Zhuge et al., 2024; Chen et al., 2024a) primarily act as implicit entropy reducers by aggregating sampled answers. However, they lack explicit mechanisms to distinguish structured disagreement from random noise, which can propagate overconfident errors when solver biases are correlated. In contrast, debate-based (Zhang et al., 2025a; Hu et al., 2025) and reflection-driven (Ge et al., 2025) methods often increase semantic novelty by introducing critiques or alternative reasoning paths; nevertheless, without a principled consolidation stage, this novelty does not reliably translate into improved calibration or decision stability. Therefore, reflection chains may amplify informational diversity without compressing uncertainty. From this perspective, the key distinction across agentic methods is not whether they generate additional reasoning, but whether they perform effective uncertainty-aware routing—i.e., selectively filtering, weighting, and consolidating existing signals into a calibrated consensus. This lens suggests that many reported "reasoning improvements" in agentic systems may be better understood as varying degrees of entropy compression versus novelty expansion, clarifying when such methods enhance reliability and when they risk confident failure.

5.2 Failure Modes: When Metacognitive Routing Fails

In this section, we characterize the regimes in which metacognitive routing provides limited or no gains, and explain why these failures arise from principled constraints rather than implementation shortcomings.

High-Agreement Regime. When solver agents already agree on the predicted answer, the empirical solver distribution $p_{\text{solver}}(y \mid x)$ exhibits low entropy. In this regime, there is little uncertainty to compress, and arbitration cannot substantially improve either accuracy or calibration. Formally, when $H_{\text{solver}}(x)$ is already low, the achievable entropy reduction

$$\Delta H(x) = H_{\text{solver}}(x) - H_{\text{arb}}(x)$$

is necessarily small. This behavior is observed empirically on low-ambiguity benchmarks such as BoolQ, where MC-MAS closely matches the strongest individual solver but does not yield significant gains (Table 1). Importantly, this indicates that MC-MAS does not degrade confident decisions, but also does not introduce unnecessary intervention when solver agreement is already strong.

Shared Systematic Bias. Our MC-MAS relies on routing and consolidating information already present in solver outputs. As a result, when all solvers share the same systematic bias by generating incorrect answers for the same underlying reason—arbitration cannot correct the error. In such cases, arbitration reinforces a biased consensus rather than uncovering alternative evidence. This limitation highlights a fundamental distinction between information routing and error correction: metacognitive arbitration can reduce disagreement, but it cannot introduce external knowledge or compensate for correlated solver failures.

Reflection Noise and Unreliable Novelty. Reflection stages often introduce semantically novel content, as confirmed by increased novelty scores in reflection-only ablations (Table 5). However, novelty alone does not guarantee reliability. When critiques surface speculative, misleading, or low-quality signals, arbitration may amplify noise rather than suppress it. This explains why reflection-only variants increase semantic novelty without improving accuracy or entropy reduction, and why full MC-MAS selectively retains novelty only when it contributes to reliable consensus formation.

Low-Quality or Miscalibrated Arbiter. The effectiveness of metacognitive routing depends critically on the calibration of the arbiter itself. A poorly calibrated arbiter may overweight spurious critiques or fail to distinguish reliable solver evidence, leading to unstable or incorrect consensus decisions. Our overconfidence analysis (Table 4) shows that MC-MAS improves confidence calibration when arbitration is well-behaved, but this benefit is not guaranteed under arbitrary arbiter behavior. Thus, metacognition does not eliminate the need for calibrated decision mechanisms—it merely shifts where calibration matters. These failure modes reinforce our central claim: MC-MAS functions as an information routing and compression mechanism rather than a universal reasoning enhancer. Its benefits arise under meaningful solver disagreement and diminish when such uncertainty is absent. We also discuss the broader impacts of this work and its limitations in Appendix A.6.

6 Conclusion

This work reframes metacognition in multi-agent LLM systems as a mechanism for information routing and compression rather than as a source of new reasoning. Through MC-MAS, we demonstrate that separating solution generation from metacognitive arbitration enables reliable consolidation of solver outputs, leading to reduced uncertainty and improved calibration across diverse reasoning tasks. In summary, our findings suggest that the primary value of metacognition lies in structuring and filtering existing information, rather than expanding reasoning traces, offering a principled direction for designing robust multi-agent inference systems. Future work will explore adaptive arbitration strategies, scalability to larger agent ensembles, and extensions to longer-horizon and more complex reasoning tasks.

Acknowledgment: This research work is supported by US NSF 2428761 and by the US Army Research Office W911NF-24-2-0133.

REFERENCES

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37:45767–45790, 2024a.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*, 2024b.

Jishnu Ray Chowdhury, Jayanth Mohan, Tomas Malik, and Cornelia Caragea. Zero-shot keyphrase generation: Investigating specialized instructions and multi-sample aggregation on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7867–7884, 2025.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. *ArXiv*, abs/2502.17026, 2025.

Fulan Fan, Siyu Wang, Mai Dinuer, Mai Hemuti, Xin Nie, and Laurence T Yang. Enhancing students' metacognition with innovative ia-based metacognitive reflective learning tool. *IEEE Transactions on Learning Technologies*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023.

Yubin Ge, Salvatore Romeo, Jason Cai, Monica Sunkara, and Yi Zhang. Samule: Self-learning agents enhanced by multi-level reflection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16602–16621, 2025.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4689–4703, 2025.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1681–1701. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.84.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. Summedits: Measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9662–9676, 2023.

Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. Training language models to critique with multi-agent feedback. *arXiv preprint arXiv:2410.15287*, 2024.

Houda Louatouate and Mohammed Zeriouh. Role-based prompting technique in generative ai-assisted learning: A student-centered quasi-experimental study. *Journal of Computer Science and Technology Studies*, 7(2):130–145, 2025.

Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1):84, 2024.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1754–1768, 2022.

Taisei Ozaki, Chihiro Nakagawa, Naoya Inoue, Shoichi Naito, and Kenshi Yamaguchi. Llm debate opponent: Counter-argument generation focusing on implicit and critical premises. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 456–465, 2025.

Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. Smart: Self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4604–4621, 2025.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching llm agents how to self-improve. In *ICML 2024 Workshop on Structured Probabilistic Inference $\{\backslash\&\}$ Generative Modeling*, 2024.

Peiyang Song, Pengrui Han, and Noah Goodman. A survey on large language model reasoning failures. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pp. 305–322. Springer, 2024.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, 2024.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3613–3635, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025a.

Zhihu Wang, Shiwan Zhao, Yu Wang, Heyuan Huang, Sitao Xie, Yubo Zhang, Jiaxin Shi, Zhixing Wang, Hongyan Li, and Junchi Yan. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4925–4936, 2025b.

Zizhen Wang, Jiangyu Pan, Duola Jin, Jingao Zhang, Jiacheng Cao, Chao Zhang, Zejian Li, Preben Hansen, Yijun Zhao, Shouqian Sun, et al. Charactercritique: Supporting children's development of critical thinking through multi-agent interaction in story reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2025c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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 AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1696–1708, 2024.

Tingting Yang, Ping Feng, Qixin Guo, Jindi Zhang, Jiahong Ning, Xinghan Wang, and Zhongyang Mao. Autohma-llm: Efficient task coordination and execution in heterogeneous multi-agent systems using hybrid large language models. *IEEE Transactions on Cognitive Communications and Networking*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Peiying Yu, Guoxin Chen, and Jingjing Wang. Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. *arXiv preprint arXiv:2502.11799*, 2025.

Quan Yuan, Mehran Kazemi, Xin Xu, Isaac Noble, Vaiva Imbrasaite, and Deepak Ramachandran. Tasklama: probing the complex task understanding of language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19468–19476, 2024.

Kaiyuan Zhang, Qian Liu, Luyang Zhang, Chaoqun Zheng, Shuaimin Li, Bing Xu, Muyun Yang, Xinxiao Qiao, and Wenpeng Lu. Madawsd: Multi-agent debate framework for adversarial word sense disambiguation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22294–22313, 2025a.

Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. Towards efficient llm grounding for embodied multi-agent collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1663–1699, 2025b.

Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23378–23386, 2025c.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

A Appendix

A.1 Prompt Templates and Pseudocode

This section documents all prompt templates used in MC-MAS. Prompts are task-specific, instruction-only, and fixed across all experimental runs. No external tools, retrieval mechanisms, or auxiliary supervision are employed beyond the task input. All outputs are constrained to explicit answer formats to enable deterministic extraction and reliable uncertainty analysis.

A.1.1 Solver, Reflection, and Arbitration Prompts

All evaluations are conducted in a zero-shot, instruction-only setting: no in-context exemplars are provided. The arbiter and critics only condition on the current instance (input) and the solvers' intermediate outputs for that same instance.

MC-MAS Prompt Templates

Solver Prompt – BoolQ (Yes/No Reading Comprehension)
Read the passage carefully and answer the question.
Passage and Question:
[Passage text]
[Question]
Evaluate the evidence carefully, then conclude with:
Final Answer: true OR Final Answer: false

Solver Prompt – ARC-Easy (Multiple-Choice Science QA)
Answer the following science question by selecting the best option.
Question:
[Question text]
Choices:

- A. *[Option A]*
- B. *[Option B]*
- C. *[Option C]*
- D. *[Option D]*

Analyze each option carefully, then conclude with:
Final Answer: A, B, C, or D

Reflection (Critique) Prompt
Review the proposed solution critically. Identify any reasoning errors, missing evidence, or alternative interpretations.
Question:
[Question text]
Proposed Solution:
[Solver reasoning output]
Provide a concise critique describing the strengths and weaknesses of the reasoning.

Metacognitive Arbitration Prompt – BoolQ
Two solver agents produced different answers, each accompanied by a critique.
Question:
[Question text]
Solution 1:
[Solver 1 output]
Answer 1: *[true/false]*
Review 1: *[Critique 1]*
Solution 2:
[Solver 2 output]

Answer 2: *[true/false]*
Review 2: *[Critique 2]*
Based on the solutions and reviews, re-evaluate carefully and conclude with:
Final Answer: true OR Final Answer: false

Metacognitive Arbitration Prompt – ARC-Easy
Two solver agents produced different answers, each accompanied by a critique.
Question:
[Question text]
Solution 1:
[Solver 1 output]
Answer 1: *[A/B/C/D]*
Review 1: *[Critique 1]*
Solution 2:
[Solver 2 output]
Answer 2: *[A/B/C/D]*
Review 2: *[Critique 2]*
Based on the solutions and reviews, re-analyze carefully and conclude with:
Final Answer: A, B, C, or D

Metacognitive Arbitration Prompt – GSM8K
Two solver agents produced different numerical solutions, each accompanied by a critique.
Question:
[Question text]
Solution 1:
[Solver 1 output]
Answer 1: *[number]*
Review 1: *[Critique 1]*
Solution 2:
[Solver 2 output]
Answer 2: *[number]*
Review 2: *[Critique 2]*
Recompute the solution independently. Do not rely blindly on prior answers. Then output exactly one line:
Final Answer: $< numericalvalue >$

Metacognitive Arbitration Prompt – StrategyQA
Two solver agents produced different answers, each accompanied by a critique.
Question:
[Question text]
Solution 1:
[Solver 1 output]
Answer 1: *[yes/no]*
Review 1: *[Critique 1]*
Solution 2:
[Solver 2 output]
Answer 2: *[yes/no]*
Review 2: *[Critique 2]*
Re-evaluate carefully, then output exactly one word:
Final Answer: yes OR Final Answer: no

MC-MAS Pseudocode. The MC-MAS inference process begins by sampling multiple reasoning trajectories from two independent solver agents using the same base model f_θ. Each solver produces a set of candidate answers, inducing an empirical solver distribution $p_{\text{solver}}(y \mid x)$ that captures initial uncertainty and disagreement (see Algorithm 1). A representative solver trace $\tilde{s}_i$ is then selected for each solver and used to generate reflective critiques r_i, which expose potential errors and supporting evidence without directly determining the final answer. In the arbitration stage, multiple arbitration responses are sampled conditioned on the input, representative solver traces, and critiques, yielding

Algorithm 1 MC-MAS with Information Routing and Uncertainty Compression

Require: Input x, base LLM f_θ, dataset $\mathcal{D}$

Number of solver samples K_s, solver temperature T_s
Number of arbitration samples K_a, arbitration temperature T_a
Initialize solver answer multiset $A_s \leftarrow \emptyset$
Initialize arbitration answer multiset $A_a \leftarrow \emptyset$
Stage 1: Independent Solver Generation
for $i = 1$ to 2 **do**
    for $k = 1$ to K_s **do**
        Sample solver reasoning trace $s_i^{(k)} \sim p_\theta(\cdot \mid x)$
        Extract solver answer $a_i^{(k)} \leftarrow \textsc{ExtractAnswer}(s_i^{(k)})$
        if $a_i^{(k)}$ is valid **then**
            $A_s \leftarrow A_s \cup \{a_i^{(k)}\}$
        end if
    end for
    Select representative solver trace $\tilde{s}_i \leftarrow s_i^{(1)}$
end for
Compute solver-induced distribution $p_{\text{solver}}(y \mid x)$ from A_s
Compute solver entropy $H_{\text{solver}}(x)$
Stage 2: Reflection and Critique
for $i = 1$ to 2 **do**
    Sample reflection $r_i \sim p_\theta(\cdot \mid x, \tilde{s}_i)$
    Extract critique claims $C_i^{\text{ref}} \leftarrow \textsc{ExtractClaims}(r_i)$
end for
Stage 3: Metacognitive Arbitration
for $m = 1$ to K_a **do**
    Sample arbitration response

$$a^{(m)} \sim p_\theta(\cdot \mid x, \tilde{s}_1, \tilde{s}_2, r_1, r_2)$$

    if $a^{(m)}$ is valid **then**
        $A_a \leftarrow A_a \cup \{a^{(m)}\}$
    end if
end for
Compute arbitration-induced distribution $p_{\text{arb}}(y \mid x)$ from A_a
Compute arbitration entropy $H_{\text{arb}}(x)$
Final prediction $\hat{y}_{\text{MC-MAS}} \leftarrow \arg\max_y p_{\text{arb}}(y \mid x)$
Information Routing Metrics
Extract solver claims $C_1, C_2 \leftarrow \textsc{ExtractClaims}(\tilde{s}_1), \textsc{ExtractClaims}(\tilde{s}_2)$
Extract arbitration claims $C_{\text{final}} \leftarrow \textsc{ExtractClaims}(A_a)$
Compute novelty

$$\mathcal{N} \leftarrow \frac{|C_{\text{final}} \setminus (C_1 \cup C_2)|}{|C_{\text{final}}|}$$

Compute redundancy

$$\mathcal{R} \leftarrow \textsc{AvgSemanticOverlap}(C_1, C_2, C_1^{\text{ref}}, C_2^{\text{ref}}, C_{\text{final}})$$

return $\hat{y}_{\text{MC-MAS}}, \{\mathcal{N}, \mathcal{R}, H_{\text{solver}}(x), H_{\text{arb}}(x)\}$

an arbitration-induced distribution $p_{\text{arb}}(y \mid x)$. The final prediction $\hat{y}_{\text{MC-MAS}}$ is obtained by majority vote over arbitration samples. Throughout this process, uncertainty is quantified via entropy before and after arbitration, and information routing behavior is characterized using semantic novelty and redundancy metrics computed from solver, reflection, and arbitration claims. This approach operationalizes metacognitive arbitration as a mechanism for selectively consolidating existing solver information rather than generating new reasoning content.

A.2 A Toy Information-Theoretic View of Metacognitive Routing

We provide a minimal theoretical interpretation of MC-MAS to clarify why metacognitive arbitration improves decision reliability without necessarily increasing semantic novelty or task accuracy. Our goal is not to offer a full formal analysis, but to ground the observed empirical behavior in an information-theoretic perspective that is consistent with our methodology and measurements.

Solvers as High-Entropy Information Sources. Let x denote an input instance and let $y \in \mathcal{Y}$ denote the task label. Due to stochastic decoding, independent solvers induce an empirical solver distribution $p_{\text{solver}}(y \mid x)$, as defined in Section 3. When solvers disagree, this distribution exhibits high entropy,

$$H_{\text{solver}}(x) \;=\; -\sum_{y \in \mathcal{Y}} p_{\text{solver}}(y \mid x) \log p_{\text{solver}}(y \mid x), \tag{14}$$

reflecting epistemic uncertainty arising from inconsistent solver outputs rather than a lack of underlying model knowledge. Majority voting aggregates solver predictions directly and reduces variance only implicitly, without an explicit mechanism to regulate or analyze this uncertainty.

Metacognitive Arbitration as an Information Bottleneck. Let $\mathcal{S} = \{a_i^{(k)} \mid i \in \{1, 2\}, k = 1, \ldots, K_s\}$ denote the multiset of solver outputs produced across agents and samples. MC-MAS introduces a metacognitive arbitration stage that integrates solver outputs and reflective critiques before producing a final decision. From an information bottleneck perspective, arbitration can be viewed as constructing an intermediate representation Z (corresponding to the arbiter's internal decision state) that selectively consolidates solver information while suppressing redundancy and conflict. Conceptually, this process trades off compression of solver signals against preservation of task-relevant information,

$$\min_{Z} \; I(Z; \mathcal{S}) \quad \text{s.t.} \quad I(Z; y) \geq \beta, \tag{15}$$

where $\mathcal{S}$ denotes solver outputs and β controls task fidelity. Importantly, this formulation is interpretive rather than prescriptive and explains why arbitration can reduce entropy and redundancy without introducing substantial new semantic content.

Consensus Formation as Entropy Compression. Through multi-sample arbitration, MC-MAS induces an arbitration distribution $p_{\text{arb}}(y \mid x)$ and corresponding entropy

$$H_{\text{arb}}(x) \;=\; -\sum_{y \in \mathcal{Y}} p_{\text{arb}}(y \mid x) \log p_{\text{arb}}(y \mid x). \tag{16}$$

Empirically, we consistently observe

$$H_{\text{arb}}(x) \;<\; H_{\text{solver}}(x), \tag{17}$$

indicating effective uncertainty compression through arbitration. Notably, this entropy reduction often occurs even when task accuracy changes little, suggesting improved calibration and reliability rather than stronger reasoning capability. Reflection-only stages tend to increase semantic novelty but do not enforce this compression constraint, which explains their limited impact on reliability.

Implications. This toy information-theoretic view predicts three behaviors that align with our empirical findings: (i) semantic novelty may decrease after arbitration due to consolidation of solver information, (ii) entropy reduction is the primary benefit of metacognition, and (iii) multi-sample arbitration is necessary to reliably approximate a low-entropy consensus state. Together, these observations support the interpretation of MC-MAS as an information routing and compression mechanism rather than a generator of new reasoning content.

A.3 Datasets and Tasks

We test MC-MAS on four diverse reasoning benchmarks that span different forms of linguistic and cognitive reasoning: BoolQ, ARC-Easy, StrategyQA, and GSM8K. These datasets are commonly used to evaluate reasoning robustness, uncertainty, and calibration in LLMs, making them suitable for studying metacognitive arbitration behavior.

1. BoolQ. BoolQ is a binary (yes/no) question answering dataset constructed from naturally occurring user queries paired with short passages from Wikipedia. Each question requires determining whether the passage entails the queried statement. BoolQ mainly tests reading comprehension and local logical reasoning with limited compositional depth. Due to its relatively low ambiguity, it acts as a useful setting for analyzing calibration and overconfidence behavior when solver agreement is high.

2. ARC-Easy. ARC-Easy is a subset of the AI2 Reasoning Challenge focused on elementary-level science questions. Questions are multiple-choice and often require combining factual knowledge with simple reasoning or elimination strategies. Compared to BoolQ, ARC-Easy exhibits higher ambiguity and solver disagreement, making it well-suited for analyzing entropy reduction and consensus formation under metacognitive arbitration.

3. StrategyQA. StrategyQA is a challenging yes/no question answering benchmark designed to require implicit multi-hop reasoning and background knowledge. Questions are intentionally constructed such that surface-level pattern matching is insufficient, and correct answers depend on combining multiple facts or commonsense assumptions. This dataset underscores the importance of semantic reasoning over explicit symbolic structure and provides a setting in which reflection often introduces novel but unreliable intermediate content.

4. GSM8K. GSM8K consists of grade-school-level math word problems that require multi-step numerical reasoning. Answers are numerical and typically require explicit symbolic manipulation, arithmetic operations, and intermediate calculations. Unlike the other benchmarks, GSM8K strongly emphasizes symbolic reasoning structure, allowing us to analyze how metacognitive arbitration interacts with solver disagreement in settings where reasoning traces are more explicit.

Task Characteristics: Across these datasets, we cover a range of task properties, including binary versus multi-class outputs, semantic versus symbolic reasoning demands, and low- versus high-uncertainty regimes. This variety allows us to examine whether metacognitive arbitration consistently improves decision reliability across different reasoning styles, rather than being tailored to a specific task format or dataset.

A.4 Scaling Behavior of Metacognitive Routing.

Our results in Table 6 show that the benefits of metacognitive routing exhibit a clear saturation effect with respect to test-time compute. As the arbitration sample size K_a increases, both accuracy and entropy reduction improve initially but plateau beyond moderate values ($K_a \approx 25$), indicating diminishing returns from additional arbitration. This behavior is consistent across benchmarks and persists under aggressive quantization, suggesting that metacognitive routing depends primarily on structured aggregation of solver outputs rather than increased model capacity.

From a formal perspective, this saturation can be understood by examining the entropy reduction

$$\Delta H(K_a; x) = H_{\text{solver}}(x) - H_{\text{arb}}^{(K_a)}(x),$$

where $H_{\text{arb}}^{(K_a)}(x)$ represents the arbitration-induced entropy obtained using K_a arbitration samples. As K_a increases, the empirical arbitration distribution $p_{\text{arb}}(y \mid x)$ converges toward a stable consensus, causing $\Delta H(K_a; x)$ to increase initially but approach a finite limit. Beyond this regime, additional arbitration primarily reduces sampling variance rather than altering the underlying decision distribution, yielding diminishing marginal gains in both entropy reduction and accuracy.

These observations indicate that metacognitive routing exhibits diminishing returns with additional test-time compute, reinforcing its interpretation as an uncertainty compression mechanism rather than a source of unbounded performance gains.

A.5 Test-Time Complexity Analysis

We explore the test-time computational complexity of MC-MAS in terms of LLM forward passes and arbitration sampling. Since MC-MAS operates purely at inference time without parameter updates or external supervision, its cost is dominated by the number of model invocations required to generate solver, reflection, and arbitration outputs.

LLM Invocation Count. For each input instance x, MC-MAS performs:

- $2K_s$ solver forward passes, corresponding to K_s stochastic samples from each of two independent solver agents;
- 2 reflection forward passes, one per solver, conditioned on representative solver traces; and
- K_a arbitration forward passes, conditioned on the input, solver traces, and reflective critiques.

The total number of LLM forward passes per instance is therefore

$$\text{Cost}_{\text{MC-MAS}} \;=\; 2K_s + 2 + K_a. \tag{18}$$

Asymptotic Inference Cost. Let C_{LLM} denote the cost of a single LLM forward pass for a fixed input length. The overall inference cost of MC-MAS scales linearly as

$$\mathcal{O}((2K_s + K_a) \cdot C_{\text{LLM}}), \tag{19}$$

with constant overhead from reflection stages. This linear scaling contrasts with approaches that increase model size or context length, and highlights that MC-MAS trades additional test-time sampling for improved uncertainty regulation.

Saturation and Cost–Benefit Boundary. Empirically, our arbitration ablation (Table 6) shows that both accuracy and entropy reduction improve with increasing K_a only up to moderate values, after which gains saturate. Formally, the marginal benefit of additional arbitration satisfies

$$\frac{\partial \mathbb{E}[\Delta H(x)]}{\partial K_a} \;\to\; 0 \quad \text{as} \quad K_a \;\to\; \infty, \tag{20}$$

while computational cost continues to grow linearly. This defines a practical cost–benefit boundary, beyond which additional arbitration yields diminishing returns without commensurate reliability improvements.

Practical Considerations. In practice, multiple arbitration samples can be efficiently batched within a single LLM invocation when supported by the inference framework, partially amortizing the cost of large K_a. Moreover, since MC-MAS relies on structured aggregation rather than high-fidelity generation, its routing behavior remains effective under aggressive quantization, as demonstrated in our 4-bit experiments. These properties make MC-MAS suitable for deployment in resource-constrained or latency-sensitive settings, provided that arbitration is applied selectively when solver uncertainty is high.

A.6 Limitations and Broader Impact

Limitations. Beyond the principled failure regimes discussed in Section 5.2, this study has several practical and empirical limitations. First, MC-MAS is evaluated on a small set of reasoning benchmarks and medium-scale instruction-tuned models, and the observed routing behavior may not directly generalize to larger frontier models or tasks requiring long-horizon planning. Second, the metacognitive arbiter introduces additional inference cost due to repeated arbitration, which may limit practicality in latency-sensitive settings. Additionally, MC-MAS relies on the quality and calibration of the metacognitive arbiter itself; failures or biases in the arbiter may propagate or amplify errors rather than mitigate them. Finally, while we show that metacognition improves confidence calibration—as measured by reduced overconfident errors—we do not claim that MC-MAS universally improves accuracy, nor that it replaces task-specific reasoning enhancements. Future work should explore adaptive arbitration strategies, lower-cost routing mechanisms, and extensions to more complex multi-agent and long-context settings.

Broader Impact. This work contributes to the development of more reliable and trustworthy agentic AI systems by clarifying the functional role of metacognition in multi-agent language models. By showing that metacognitive arbitration primarily enhances confidence calibration and reduces overconfident errors—rather than generating new reasoning content—our findings support safer deployment of agents in open and uncertain environments where incorrect high-confidence decisions

can be costly. The routing-based perspective facilitates system designs that selectively intervene based on uncertainty, potentially reducing unnecessary computation and mitigating failure modes associated with correlated agent biases. Besides this, our work underscores the importance of understanding when metacognitive mechanisms provide limited benefit, helping practitioners avoid unnecessary computational overhead or false assurances of reliability. Lastly, while our framework is evaluated in controlled benchmark settings, the principles uncovered here are broadly applicable to real-world agentic systems that must balance accuracy, reliability, and efficiency under resource constraints.