UMAS: System-Level Uncertainty Quantification for Multi-Agent LLM Systems

Hanwen Li[1] **Jinhao Duan**[2*] **Xiaoshuang Shi**[3] **Yue Zhang**[4]
Tianlong Chen[2] **Kaidi Xu**[5] **Chenxi Yuan**[1*]

[1]New Jersey Institute of Technology [2]University of North Carolina at Chapel Hill
[3]University of Electronic Science and Technology of China [4]Shandong University
[5]City University of Hong Kong
hl544@njit.edu, jinhao@cs.unc.edu, xsshi2013@gmail.com, zyueinfosec@sdu.edu.cn,
tianlong@cs.unc.edu, kaidixu@cityu.edu.hk, chenxi.yuan@njit.edu

Abstract

Multi-Agent Systems (MAS) are a promising paradigm for improving the reasoning capabilities of large language models, yet they remain prone to hallucinations and erroneous outputs. Uncertainty Quantification (UQ) is crucial for assessing system reliability, but existing methods are largely designed for single-agent settings; how to exploit MAS structure for UQ while keeping inference cost-neutral remains underexplored. We propose UMAS, a parameter-free, cost-neutral, system-level UQ framework that estimates the credibility of inter-agent influence from each generation's intrinsic confidence. UMAS uses these credibility signals to modulate how uncertainty is updated at each node, and then aggregates node-level uncertainty into a system-level estimate. Extensive evaluations on Debate and DyLAN across diverse benchmarks show that UMAS consistently outperforms state-of-the-art baselines by an average of 10.54 AUROC points. Beyond uncertainty estimation, UMAS enables hallucination detection and uncertainty-aware answer selection, improving MAS accuracy by up to 12.6 points on specific tasks while enhancing reliability. Code is available at `https://github.com/Dby52-04/UMAS`.

1 Introduction

Large Language Models (LLMs) have achieved strong performance across many applications [45], yet they still struggle with factuality and complex reasoning, leading to hallucinations and errors [17, 24, 19]. Prior work mitigates these issues via Chain-of-Thought [40], self-consistency [38], and Retrieval-Augmented Generation [13, 31], and multi-agent systems (MAS) further extend capability through collaborative reasoning [23, 4]. However, hallucinations persist even in MAS [42], making reliable detection of erroneous agent outputs essential for trustworthy deployment.

Uncertainty Quantification (UQ) is a key tool for evaluating the reliability and trustworthiness of LLM generations. As illustrated in Figure 1, an important issue emerges: although MAS can achieve higher task accuracy than a single LLM (Appendix B.2), standard UQ applied to MAS output can become substantially less reliable. We identify two major limitations of existing UQ methods when applied to MAS. (i) **Structure misalignment:** Most UQ methods estimate uncertainty solely from the final response to a query [21, 37, 5]. However, an MAS output is produced through a multi-round, multi-agent interaction trajectory; restricting UQ to the terminal answer ignores this structure and can miss how uncertainty is propagated and amplified through inter-agent influence. (ii) **Efficiency:** MAS also substantially increases inference cost. As shown in Table 7, while MAS architectures improve

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

response accuracy, they incur significant overhead. Meanwhile, many UQ approaches [21, 7, 37, 25] require multiple samples per query, which becomes prohibitively expensive for MAS due to the length and multiplicity of multi-agent generations.

These limitations motivate our primary research question: can we develop an MAS-specific UQ framework that satisfies two core criteria? (i) **Structure-aware:** leverage the structure and interaction trajectory of MAS, rather than treating uncertainty solely as a property of the final system output. (ii) **Cost-neutral:** maintain near-zero additional inference overhead, requiring no extra sampling or additional model invocations.

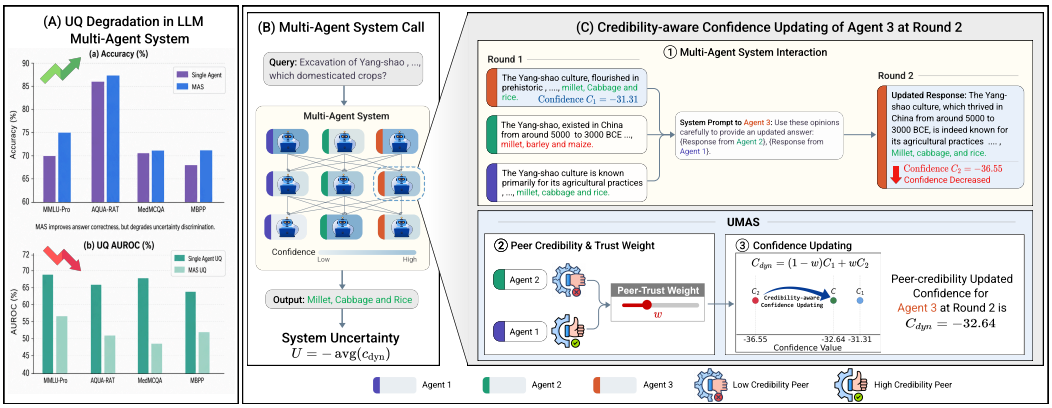


Figure 1: **Left (A):** Applying standard single-agent UQ to the final MAS output exhibits degraded reliability (lower AUROC) despite improved task accuracy. **Right (B & C):** Overview of the UMAS framework during a collaborative update. A focal agent (e.g., Agent 3) revises its response using peer generations. UMAS estimates node-level uncertainty and dynamically calibrates it based on peer credibility. Peer credibility (denoted by the gear icons) is evaluated to compute a trust weight, generating a credibility-aware updated confidence score (c_{dyn}).

To answer this question, we propose UMAS, a system-level UQ framework for MAS. Instead of computing uncertainty only on the terminal response, UMAS assigns node-level uncertainty to each agent response at each interaction round and derives a system-level uncertainty by aggregating these node-level estimates over the MAS interaction graph. The key design question is how to assign node-level uncertainty in light of inter-agent interactions: in collaborative MAS protocols, an agent revises its response by incorporating peer generations, but blindly trusting peer information can distort uncertainty—analogous to anchoring on an unreliable peer's claim. UMAS therefore estimates the credibility of each inter-agent influence from the generations' intrinsic confidence and uses these signals to modulate how uncertainty is updated during interaction. Figure 1 illustrates the procedure for a particular update instance. This credibility-aware UQ prevents unreliable peer evidence from inducing overconfident or spurious uncertainty.

We evaluate UMAS under two representative collaborative MAS paradigms: Debate [6], a canonical collaboration-based protocol, and Dynamic LLM-Agent Network (DyLAN) [27], an adaptive framework that supports iterative team reconfiguration during execution. We instantiate both architectures with two widely used LLM backbones, GPT-4o mini [1] and Llama 4 Scout [30], and benchmark on five diverse tasks: MMLU-Pro (general knowledge) [39], AQUA-RAT (mathematics) [26], MedMCQA (medicine) [32], MBPP (programming) [3], and GSM-Hard (mathematics) [12]. Across architectures, models, and datasets, UMAS improves the reliability of uncertainty estimates over strong UQ baselines [21, 25, 7, 2, 37]. Compared to the best-performing baseline in each setting, UMAS achieves an average AUROC gain of 10.54 points across both backbones. Finally, we show that the resulting uncertainty improvements translate into downstream gains for selective prediction and hallucination detection in collaborative MAS, improving overall system reliability and enabling effective detection of erroneous responses. Our contributions are summarized as follows:

- We identify a failure mode that harms UQ reliability in MAS: although MAS often improves task accuracy, applying standard UQ can significantly degrade uncertainty reliability. We attribute this to a structural mismatch: MAS uncertainty depends on the multi-round interaction and the credibility of inter-agent influence.

- We propose UMAS, a parameter-free and cost-neutral UQ framework that assigns node-level uncertainty to each agent response at each interaction round and aggregates it into a system-level uncertainty.
- We evaluate UMAS across two collaborative MAS paradigms, two LLM backbones, and five benchmarks. We further show that these UQ improvements translate into better selective prediction and hallucination detection in MAS, thereby enhancing the overall performance and reliability of the MAS.

2 Related Works

Uncertainty Quantification. Uncertainty Quantification (UQ) is critical for assessing LLM reliability, with existing literature largely focused on single-turn tasks [36]. Methodologies generally fall into two categories: (i) information-theoretic approaches, which efficiently derive uncertainty from token probabilities but often overlook semantic variability, and (ii) consistency-based strategies, such as Semantic Entropy [21], which utilize multi-sampling to analyze semantic alignment across generations. Recent hybrid frameworks like CoCoA [37] aggregate confidence and consistency to bridge these perspectives. Furthermore, UQ has expanded into multi-step decision-making; methods like SAUP [44], for instance, train a Hidden Markov Model to assign a situational weight to the uncertainty of each step. However, this method requires ground-truth labels from the test domain and is not designed for MAS. UProp [8] samples multiple trajectories of a single agent's decision chain, a special case of our trajectory graph; UMAS scores one realized multi-agent trajectory. On the evaluation side, UQ performance is sensitive to the choice of correctness function [18], and length biases shared by UQ scores and correctness functions can distort AUROC rankings [35]; we evaluate UMAS under alternative correctness definitions and length controls (Section 5.4).

Multi-Agent Systems. To address the limitations of single-agent models, Multi-Agent Systems (MAS) employ collaborative, multi-turn interactions [42]. Foundational frameworks like CAMEL [22] and AutoGen [41] utilize role-playing mechanisms to facilitate task completion, while debate-oriented architectures [6] leverage dialectical strategies for solution refinement. Advancing beyond static topologies, DyLAN [27] introduces dynamic team optimization, and MAS-GPT [43] generates query-specific architectures. Concurrently, approaches like DiverseAgentEntropy [10] integrate uncertainty metrics to prioritize reliable agents. Our work focuses on the collaborative update mechanism, a fundamental component across these paradigms, to enhance UQ within MAS.

3 Preliminary

Confidence Estimation. LLMs are autoregressive generative models. Letting p_θ denote the parameterized LLM, the probability of output y given input x, referred to as the Maximum Sequence Probability (MSP), factorizes over tokens; we work with its logarithm throughout:

$$\log P(y \mid x) = \sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x), \tag{1}$$

where $y = (y_1, \ldots, y_T)$ is the output sequence of length T and $y_{<t}$ is its prefix. Its negation $-\log P(y \mid x)$ is the standard MSP uncertainty score [36]. MSP serves as a computationally efficient confidence proxy [20].

Multi-Agent System. Following the framework proposed in [33], we formulate the MAS as a directed graph architecture $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We define N as the number of agents and M as the total number of communication rounds (equivalent to layers in the graph). The vertex set $\mathcal{V}$ represents the complete history of agent states:

$$\mathcal{V} = \left\{ v_i^{(t)} \mid i \in \{1, \ldots, N\},\, t \in \{1, \ldots, M\} \right\},$$

where $v_i^{(t)}$ denotes the state of the i-th agent at round t. Let $g(v_i^{(t)})$ represent the content generated by the LLM for this node. The edge set $\mathcal{E}$ is constructed based on the information dependency flow.

Specifically, if the generation of $v_i^{(t)}$ depends on aggregating a set of predecessor states from the previous round, denoted as $\mathbb{P}_i^{(t)} \subseteq \{v_1^{(t-1)}, \ldots, v_N^{(t-1)}\}$, we define the edges as:

$$\mathcal{E} = \bigcup_{t}^{M} \bigcup_{i}^{N} \left\{ \langle u, v_i^{(t)} \rangle \mid u \in \mathbb{P}_i^{(t)} \right\}.$$

4 Method

Node-level Confidence Score. While various uncertainty estimation methods exist, such as Semantic Entropy (SE) [21] and Predictive Entropy (PE) [11], we use Maximum Sequence Probability (MSP; Eq. 1) because it is cost-neutral for a completed generation. Unlike sampling-based methods, MSP estimates confidence directly from generation logits and requires no additional inference passes. Formally, let $x_i^{(t)}$ denote the input context for node $v_i^{(t)}$. We assign the node a scalar confidence score $c_i^{(t)}$, defined as the log-probability of the generated response conditioned on the input:

$$c_i^{(t)} = \log P\left(g(v_i^{(t)}) \mid x_i^{(t)}\right), \tag{2}$$

where $\log P$ is the log-MSP defined in Eq. 1. Hence $c_i^{(t)} < 0$, and a larger (less negative) value indicates higher confidence.

UMAS extends this node-level confidence to the MAS setting by tracking how confidence evolves through inter-agent interactions. In addition to the raw confidence $c_i^{(t)}$, each node maintains a dynamic confidence $c_{i,\mathrm{dyn}}^{(t)}$, which adjusts the node's confidence according to the credibility of the peer evidence used to produce it.

4.1 Peer Credibility Estimation

In a collaborative MAS, an agent often revises its response by conditioning on generations from other agents. The reliability of these peer references should therefore affect the confidence assigned to the revised response. For a node $v_i^{(t)}$, we denote its peer reference set as $\mathbb{R}_i^{(t)} = \{g(u) \mid \langle u, v_i^{(t)} \rangle \in \mathcal{E}\}$.

Rather than treating all references equally, UMAS estimates a credibility score for each directed update edge. For an edge $\langle v_j^{(t-1)}, v_i^{(t)} \rangle \in \mathcal{E}$, where peer node $v_j^{(t-1)}$ contributes to the update of $v_i^{(t)}$, we define the raw edge credibility as the peer's relative dynamic confidence with respect to the updater's previous state:

$$w_{j \to i}^{(t-1)} = \frac{c_{j,\mathrm{dyn}}^{(t-1)} - c_{i,\mathrm{dyn}}^{(t-1)}}{c_{i,\mathrm{dyn}}^{(t-1)}}, \tag{3}$$

where $c_{j,\mathrm{dyn}}^{(t-1)}$ and $c_{i,\mathrm{dyn}}^{(t-1)}$ are the dynamic confidences from the previous round. The relative form in Eq. 3 makes the score comparable across prompts, datasets, and models. Since $c_{i,\mathrm{dyn}}^{(t-1)} < 0$, a peer more confident than the updater's prior state yields a negative score, whereas a less confident, and hence less credible, peer yields a positive score. This gives UMAS a simple heuristic signal, not a truth detector (Section 7), for deciding how strongly peer-supported revisions should affect confidence.

4.2 Credibility-aware Confidence Updating

Given the raw edge credibility scores above, we next convert peer evidence into a bounded trust weight and use it to update the agent's dynamic confidence.

Peer-Trust Weighting. For each update node $v_i^{(t)}$, UMAS summarizes all incoming peer evidence into a single node-level trust weight. We first map each raw edge score $w_{j \to i}^{(t-1)}$ to $(0, 1)$ with a sigmoid:

$$\hat{w}_{j \to i}^{(t-1)} = \sigma\left(w_{j \to i}^{(t-1)}\right) \in (0, 1). \tag{4}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. This mapping stabilizes the update by preventing extreme relative-confidence ratios from producing unbounded influence. It also preserves the ordering of the credibility signal: equal confidence maps to 0.5, more credible peers map below 0.5, and less credible peers map above 0.5. We then average the normalized incoming edge scores and take the complement to obtain the node-level trust weight, so that more credible peers yield a larger weight:

$$w_i^{(t)} = 1 - \mathbb{E}_{j:\langle v_j^{(t-1)}, v_i^{(t)} \rangle \in \mathcal{E}} \left[\hat{w}_{j \to i}^{(t-1)} \right]. \tag{5}$$

The weight in Eq. 5 should be interpreted as the effective credibility of the current peer-supported update. Appendix H.4 compares σ with alternative normalization functions and shows similar trends, indicating that the method is not sensitive to the particular bounded mapping.

Confidence Updating. Finally, we use the node-level trust weight $w_i^{(t)}$ to regulate how confidence evolves when an agent updates its response using peer references. Standard single-agent UQ treats confidence as a static property of the terminal output. In contrast, UMAS tracks confidence along the interaction trajectory and allows it to change in proportion to the credibility of the peer evidence. We update dynamic confidence with a credibility-weighted moving average:

$$c_{i,\mathrm{dyn}}^{(t)} = (1 - w_i^{(t)}) \cdot c_{i,\mathrm{dyn}}^{(t-1)} + w_i^{(t)} \cdot c_i^{(t)}, \tag{6}$$

where $w_i^{(t)} \in [0, 1]$ is the peer-trust weight, $c_i^{(t)}$ is the raw confidence of the current generation at node $v_i^{(t)}$, and $c_{i,\mathrm{dyn}}^{(t-1)}$ is the dynamic confidence from the previous round. This rule yields two intuitive regimes:

- **Low peer credibility ($w_i^{(t)} \to 0$):** The current update is likely influenced by unreliable references. We therefore suppress the change and retain the historical estimate, $c_{i,\mathrm{dyn}}^{(t)} \approx c_{i,\mathrm{dyn}}^{(t-1)}$, preventing spurious overconfidence or distorted uncertainty.
- **High peer credibility ($w_i^{(t)} \to 1$):** The update is supported by credible references. We allow confidence to track the current response, $c_{i,\mathrm{dyn}}^{(t)} \approx c_i^{(t)}$, reflecting reliable correction from peer interactions.

4.3 System-level Uncertainty Aggregation

Multi-sampling, consistency-based UQ methods are prohibitively expensive for MAS because they would require repeated sampling at each interaction node. To remain cost-neutral with respect to a completed MAS run, UMAS uses only the single-response confidence signals already produced during generation.

The objective of UMAS is to produce one uncertainty score for the entire MAS generation trajectory. Given an interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, input x, and generated contents $\mathcal{Y} = \{g(v) \mid v \in \mathcal{V}\}$, we define system-level uncertainty as the negative average dynamic confidence over the trajectory:

$$U_{\mathcal{G},\mathrm{dyn}}(\mathcal{Y} \mid x) \;=\; -\frac{1}{|\mathcal{V}|} \sum_{v_i^{(t)} \in \mathcal{V}} c_{i,\mathrm{dyn}}^{(t)}, \tag{7}$$

We use the average in Eq. 7 because it yields a stable, length-normalized summary across MAS runs with different numbers of agents and rounds, and prevents any single interaction node from dominating the system score.

4.4 UMAS$_{FIX}$: Disentangling Aggregation from Credibility Updating

UMAS consists of two cost-neutral components: (i) aggregating uncertainty over the entire MAS interaction graph rather than only the terminal output, and (ii) a heuristic update of confidence along the trajectory using peer credibility. To isolate these contributions, we define a simpler structure-aware instantiation, UMAS$_{FIX}$. It uses the same trajectory aggregation as Eq. 7, but replaces the credibility-updated confidence $c_{i,\mathrm{dyn}}^{(t)}$ with the raw node confidence $c_i^{(t)}$; equivalently, it skips the update in Eq. 6. Thus, UMAS$_{FIX}$ is structure-aware but credibility-agnostic. Comparing the resulting settings disentangles the two sources of improvement:

- UMAS$_{FIX}$ vs. a single-output UQ baseline measures the gain from *trajectory-level aggregation alone*.
- UMAS vs. UMAS$_{FIX}$ measures the additional gain from *credibility-aware updating*.

Our experiments report both deltas and find that each contributes substantially. Because all quantities are computed from logits and graph structure already produced in a standard MAS run, UMAS requires no extra generations, model calls, or auxiliary models. The complete pseudocode of UMAS is provided in Algorithm 1 (Appendix C).

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate on five benchmarks spanning distinct cognitive domains: MMLU-Pro (general knowledge) [39], AQUA-RAT (mathematics) [26], MedMCQA (medicine) [32], MBPP (programming) [3], and GSM-Hard (mathematics) [12]. Specifications are in Appendix D.

Multi-Agent Systems and Backbone Models. To evaluate UMAS on collaborative update mechanisms, we integrate it into two representative architectures: Debate [6], a foundational paradigm for iterative peer critique, and DyLAN [27], an adaptive framework that dynamically optimizes team composition. We employ two distinct state-of-the-art LLMs as backbone agents: GPT-4o mini [1], a cost-efficient commercial model, and Llama 4 Scout [30], a leading open-source Mixture-of-Experts model. Detailed architectural specifications are provided in Appendix B.

UQ Baselines. We benchmark UMAS against eight prominent Uncertainty Quantification methodologies originally developed for single-turn QA tasks: Perplexity (PPL) [11], MSP, PE [14], SE [21], SentSAR [7], Deg [25], CoCoA [37], and G-NLL [2]. MSP and G-NLL differ only in which sequence is scored: MSP scores the response the MAS actually emitted, whereas G-NLL scores the most likely output sequence, obtained by greedy decoding in its original single-agent formulation. For all baselines, uncertainty is computed based on the final output of the MAS; we discuss this choice and provide an extended comparison where consistency-based baselines additionally aggregate across intermediate generations in Appendix H.2.

A critical challenge in applying consistency-based methods to MAS is the prohibitive computational cost; these methods typically analyze semantic equivalence across multiple stochastic samples, which becomes resource-intensive given the extensive length of MAS trajectories. To address this bottleneck, we implement an optimization strategy that computes semantic consistency over a condensed representation of the generation, rather than the raw output (we denote methods using this optimization with an asterisk, e.g., SE*; detailed implementation is provided in Appendix E). For all methods requiring multi-sampling, we standardize the sample number to 5.

Evaluation Metric. Following prior UQ work [21], we evaluate the efficacy of UMAS using the Area Under the Receiver Operating Characteristic Curve (AUROC), which measures the alignment between the estimated uncertainty and the correctness of the MAS's response. Correctness is determined by exact match between the extracted answer and the ground-truth answer for the multiple-choice and short-answer tasks (MMLU-Pro, AQUA-RAT, MedMCQA, GSM-Hard), and by pass@1 over all provided test cases for the code task (MBPP). Results are robust to the correctness definition and to response length (Appendices H.7 and H.8). We additionally report Expected Calibration Error (ECE) in Appendix H.3: on Llama 4 Scout under Debate, UMAS attains the lowest average ECE (8.79 vs. 15.67 for the best baseline), so the improvements extend beyond ranking to calibration.

5.2 Main Result

Table 1 reports AUROC for UMAS against established baselines on a random 200-instance subset per dataset. Both UMAS$_{FIX}$ and UMAS substantially improve uncertainty reliability across architectures, datasets, and backbones: on Llama 4 Scout the best baseline reaches Ave 63.00, while UMAS$_{FIX}$ and UMAS reach 70.32 and 74.93; on GPT-4o mini the best baseline reaches 64.86, versus 71.49 and 74.02, respectively. The two design choices decompose cleanly: trajectory-level aggregation alone

Table 1: AUROC (%) results. "Ave" represents the mean over the ten dataset cells in each row (GSM-H = GSM-Hard). Best results per column are shown in **bold**; second-best are underlined.

LLAMA 4 SCOUT

	DEBATE					DYLAN					
METHOD	MMLUP	AQUA	MEDMC	MBPP	GSM-H	MMLUP	AQUA	MEDMC	MBPP	GSM-H	AVE
PPL	67.67	64.48	56.53	60.03	36.02	60.15	71.94	53.93	**58.61**	55.28	58.46
MSP	70.70	78.61	60.43	62.02	47.98	61.02	75.00	54.95	58.39	57.47	62.66
PE	69.58	70.33	57.81	59.99	37.09	63.58	74.03	54.77	52.53	56.44	59.62
SE	47.81	47.33	45.83	37.42	51.41	50.29	55.30	56.09	48.09	49.41	48.90
SENTSAR	57.86	74.52	58.18	59.59	50.59	54.35	49.03	48.09	51.74	48.82	55.28
DEG	54.89	67.75	59.20	64.29	47.77	54.22	65.48	52.03	54.40	78.01	59.80
COCOA	68.91	79.39	60.74	62.36	56.13	60.94	73.02	54.70	58.27	55.36	62.98
G-NLL	66.95	77.33	61.23	60.47	46.97	62.31	70.81	55.24	55.81	60.18	61.73
SE*	55.08	50.70	48.30	52.06	55.80	56.99	67.02	60.50	55.06	53.47	55.50
SENTSAR*	67.73	67.49	55.49	59.29	42.70	56.54	68.15	50.21	49.76	57.10	57.45
DEG*	58.82	62.76	58.02	62.63	62.56	55.62	72.24	52.68	37.36	60.62	58.33
COCOA*	69.20	78.59	60.43	62.43	49.53	61.58	76.62	54.37	58.04	59.24	63.00
UMAS$_{FIX}$	**80.97**	79.17	72.68	62.00	64.25	79.71	80.11	66.41	51.93	65.97	70.32
UMAS	80.90	**79.77**	**74.38**	**64.56**	**72.69**	**83.70**	**80.63**	**76.16**	56.98	**79.48**	**74.93**

GPT-4O MINI

	DEBATE					DYLAN					
METHOD	MMLUP	AQUA	MEDMC	MBPP	GSM-H	MMLUP	AQUA	MEDMC	MBPP	GSM-H	AVE
PPL	60.12	51.43	58.67	47.85	47.83	61.50	53.88	66.07	41.93	45.30	53.46
MSP	70.50	63.38	62.64	52.16	57.93	68.86	70.98	68.58	47.49	52.64	61.52
PE	74.06	69.83	60.51	48.16	55.97	72.30	75.51	70.37	46.69	63.28	63.67
SE	54.13	53.49	57.73	46.93	55.14	46.87	53.45	43.73	42.73	44.05	49.83
SENTSAR	75.42	71.45	62.79	51.79	60.16	65.59	69.38	71.08	46.63	51.79	62.61
DEG	56.76	61.76	45.68	52.90	55.90	72.95	79.45	58.63	53.35	76.28	61.37
COCOA	67.48	66.36	58.93	51.43	60.73	68.82	67.54	66.19	46.65	49.28	60.34
G-NLL	70.62	60.86	62.04	53.04	59.19	68.72	70.98	68.58	47.49	52.64	61.42
SE*	57.86	61.86	56.06	53.27	56.18	55.34	59.81	46.12	45.94	49.76	54.22
SENTSAR*	**77.60**	71.23	62.30	53.07	59.72	73.32	73.87	72.73	47.80	57.00	64.86
DEG*	56.13	66.15	46.58	55.40	60.94	58.75	67.05	56.49	47.80	69.35	58.46
COCOA*	71.39	65.47	57.28	52.18	58.13	74.83	70.08	66.49	47.14	57.45	62.04
UMAS$_{FIX}$	77.42	81.70	66.03	57.62	67.89	75.58	84.22	**73.17**	59.36	71.93	71.49
UMAS	74.40	**85.51**	**71.03**	**60.37**	**74.50**	**76.54**	**87.48**	72.33	**61.69**	**76.39**	**74.02**

(UMAS$_{FIX}$ vs. best baseline) contributes $+7.32$/$+6.63$ AUROC across the two backbones, and credibility-aware updating (UMAS vs. UMAS$_{FIX}$) adds a further $+4.61$/$+2.53$; aggregation is thus the primary source of improvement. The credibility-aware contribution is most pronounced when peer confidences span a wide range (e.g., GSM-Hard, MedMCQA) and shrinks when peers are uniformly confident, consistent with the intuition that peer weighting matters most when peer reliability varies. The trend holds on a third backbone, Qwen 3.5-9B [34], where UMAS reaches Ave 75.75 versus 68.89 for the strongest baseline (Appendix H.1); case studies are in Appendix G.

5.3 Selective Prediction

Beyond standard UQ metrics, we evaluate the practical utility of UMAS in selecting the optimal response from multiple candidate generations [15]. Specifically, for each input x, we generate $K = 5$ independent response trajectories $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_K\}$ using a temperature setting of 0.5. The final system output is determined by selecting the trajectory that minimizes the estimated uncertainty:

$$\hat{y} = \mathcal{Y}_{k^*} \quad \text{where} \quad k^* = \underset{i \in \{1,\ldots,K\}}{\arg\min}\ U_{\mathcal{G}}(\mathcal{Y}_i \mid x).$$

Table 2 reports the overall accuracy of this selection mechanism using GPT-4o mini. The results indicate that our method significantly enhances system performance compared to the original MAS, with absolute accuracy gains under UMAS ranging from 5.5 points (DyLAN-MMLUP) up to 12.6

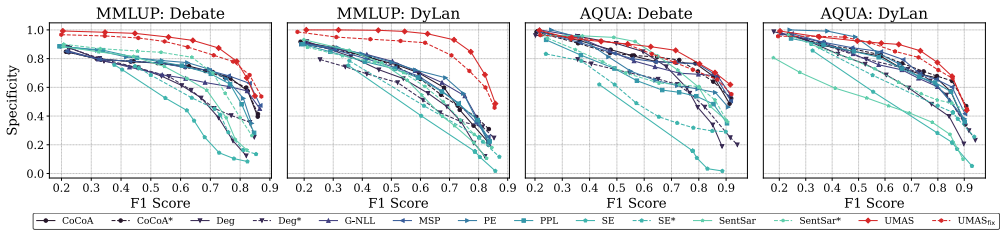


Figure 2: Trade-off analysis between F1 Score and Specificity on the test dataset using Llama 4 Scout. The curves illustrate performance variations across a range of classification thresholds $\delta \in \{0.1, 0.2, \ldots, 0.9\}$. Methods exhibiting trajectories closer to the upper-right corner demonstrate superior efficacy in hallucination detection.

points (Debate-AQUA). Notably, UMAS or UMAS$_{FIX}$ achieves the highest selection accuracy in every cell evaluated, and never falls below the strongest UQ baseline.

Table 2: Comparative analysis of selective prediction accuracy (%) across different UQ methods. The *Original* denotes the standard performance of the MAS relying on a single generation trajectory, without applying any selection mechanism. Per row: **best** / second-best.

Dataset	MAS	*Orig.*	PPL	MSP	G-NLL	PE	SE	SE*	SentSAR	SentSAR*	Deg	Deg*	CoCoA	CoCoA*	UMAS$_{FIX}$	UMAS
MMLUP	Debate	70.0	74.0	73.5	74.5	75.0	72.0	72.0	73.0	74.5	68.5	75.0	73.5	74.5	74.5	**76.0**
	DyLAN	71.0	66.0	68.5	68.5	74.0	71.0	72.0	71.5	71.0	76.5	74.5	72.5	75.0	**77.0**	76.5
AQUA	Debate	83.7	85.3	91.1	91.1	94.2	81.6	83.7	93.2	93.7	93.2	90.0	90.5	92.1	94.8	**96.3**
	DyLAN	83.2	81.6	89.5	89.5	90.0	83.2	83.7	90.0	91.1	89.5	81.1	87.4	83.2	92.2	**93.7**

5.4 Ablation Study

Hallucination Detection. A primary application of UQ is automated hallucination detection in LLM generations [29, 28]. When ground truth is available, we define a hallucination as an incorrect response that does not match the reference answer. Accordingly, hallucination detection is a binary classification task: a response is flagged as hallucinated if its estimated uncertainty exceeds a threshold γ. To determine the optimal γ, a threshold-selection step is required. In our experiments, we utilize a 20% validation set ($N = 40$) from all data to derive a threshold for each method. The objective is to maximize specificity (True Negative Rate) while maintaining a target F1 score, formalized as:

$$\gamma^* = \underset{\gamma}{\operatorname{argmax}} \, \text{Specificity}(\gamma) \quad \text{s.t.} \quad \text{F1}(\gamma) \geq \delta,$$

where δ is the target F1 level. Figure 2 reports specificity across varying F1 constraints on the test set. UMAS$_{FIX}$ and UMAS consistently outperform all baselines, with UMAS closer to the upper-right corner than UMAS$_{FIX}$ in most settings, confirming the benefit of credibility-aware uncertainty dynamics.

MAS Scalability. Multi-Agent Systems exhibit inherent architectural flexibility, allowing for variable configurations of agent count N and communication rounds M. It is critical to assess UQ methods across these varying scales. Figure 3 presents a comparative analysis between UMAS and the four top-performing baselines across a comprehensive range of configurations, specifically $N, M \in \{2, 3, 4\}$. The results show that UMAS$_{FIX}$ and UMAS consistently outperform all baselines across the evaluated scales, demonstrating strong generalizability across diverse MAS deployments. Overall, UMAS achieves the best performance in all settings.

Agent Scalability. We evaluate UMAS across seven Llama variants ranging from 1B to 400B total parameters: Llama-3.2 (1B, 3B, 11B-Vision) [9], Llama-3.1-8B [9], Llama-3.3-70B [9], Llama 4 Scout (109B) [30], and Llama 4 Maverick (400B) [30]. For models below 8B, we use MMLU [16] since MMLU-Pro [39] exceeds their reasoning capability. Figure 4 reports AUROC across these seven scales on four datasets; UMAS consistently outperforms most baselines, confirming scalability across model sizes.

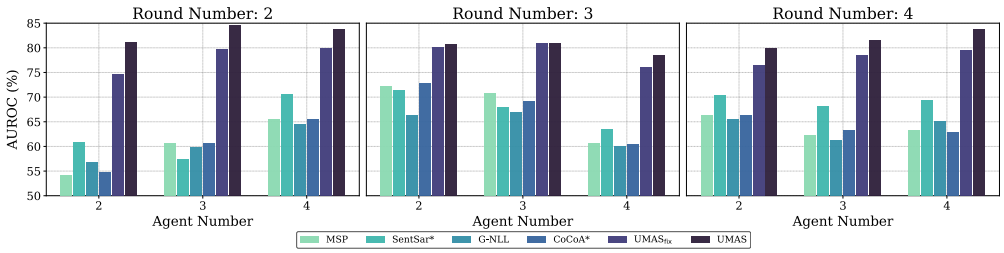


Figure 3: Comparative AUROC performance (%) across varying agent system scales using Llama 4 Scout.

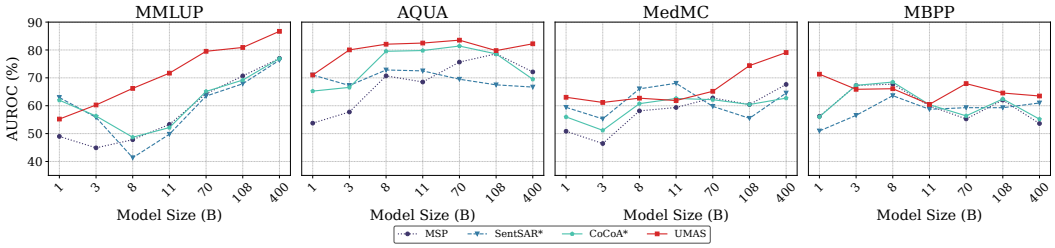


Figure 4: Comparative AUROC performance (%) across varying agent parameter scales.

Impact of Confidence Metrics. UMAS is metric-agnostic: any strictly positive single-pass uncertainty score u can be plugged in as $c = -u$ (for MSP, $u = -\log P$). Replacing MSP with PE or PPL for the node-level confidence score (Llama 4 Scout, Debate), we observe substantial gains over the raw baseline metric in 11 of 12 configurations (Table 3), with the sole exception of PPL on MBPP, where UMAS is essentially on par with the baseline (59.87 vs. 60.03). The framework therefore inherits future single-turn UQ advances at no architectural cost.

Table 3: Comparative AUROC performance (%) utilizing distinct confidence metrics using Llama 4 Scout and Debate. "Base" is the raw confidence metric; UMAS applies our framework on top.

	MSP				PPL				PE			
METHOD	MMLUP	AQUA	MEDMC	MBPP	MMLUP	AQUA	MEDMC	MBPP	MMLUP	AQUA	MEDMC	MBPP
BASE	70.70	78.61	60.43	62.02	67.67	64.48	56.53	**60.03**	69.58	70.33	57.81	59.99
UMAS	**80.90**	**79.77**	**74.38**	**64.56**	**77.98**	**72.16**	**70.88**	59.87	**79.01**	**73.69**	**70.29**	**60.67**

Computational Efficiency.

Since MAS already incurs $N \times M$ inference calls per query, the UQ mechanism's overhead matters. UMAS uses only single-trajectory logits and lightweight CPU-bound probability computations, avoiding the latency and hardware dependencies of multi-sample consistency methods (e.g., SE, SentSAR with DeBERTa-large entailment). Table 4 compares costs against baselines (hardware in Appendix F), where *In.* denotes the inference latency and *Un.* represents the time cost associated with uncertainty estimation.

Table 4: Comparison of computational efficiency. Latency and AUROC measured on Llama 4 Scout under Debate on MMLU-Pro.

METHOD	IN. (S)	UN. (MS)	GPU FREE	AUROC
Multi-Sampling				
SE	127.32	1813.92	✗	47.81
SENTSAR	127.32	263.30	✗	57.86
DEG	127.32	3726.16	✗	54.89
COCOA	127.32	262.85	✗	68.91
G-NLL	127.32	0.181	✓	66.95
PE	127.32	0.164	✓	69.58
Single Sampling				
PPL	84.17	0.046	✓	67.67
MSP	84.17	0.021	✓	70.70
UMAS	84.17	0.143	✓	**80.90**

Alternative Evaluation Protocols. We further evaluate UMAS under three alternative protocols (Table 5): (i) *alternative correctness labels*, obtained with five answer extractors and two aggregates (strict and lenient), for seven definitions in total; (ii) *length-stratified AUROC*, computed within

Table 5: AUROC (%) under alternative evaluation protocols, averaged over the settings of Table 1. Strongest baseline: highest mean among baselines.

PROTOCOL	STRONGEST BASELINE	UMAS	Δ
STANDARD (TABLE 1)	62.52 (COCOA*)	74.47	+11.95
ALTERNATIVE CORRECTNESS LABELS (E4)	61.83 (COCOA*)	70.02	+8.19
LENGTH-STRATIFIED AUROC	61.06 (PE)	69.26	+8.20
TRAJECTORY-AGGREGATED BASELINES	67.74 (PPL)	74.47	+6.73

response-length quartiles; and (iii) *trajectory-aggregated baselines*, where every baseline's per-node scores are averaged over the trajectory with the same rule as UMAS$_{FIX}$. UMAS is the best method under all three protocols, leading the strongest baseline by $+6.73$ to $+8.20$ AUROC, and under every correctness definition UMAS ranks first and UMAS$_{FIX}$ second. UMAS also carries information beyond response length: adding it to a logistic regression of correctness on MSP, terminal length, and trajectory length improves AUROC by $+4.05$. Details are in Appendices H.2, H.7, and H.8.

6 Conclusion

We show that standard single-agent UQ can become unreliable when applied to the final output of MAS, since it ignores trajectory-level interactions and influence credibility. We propose UMAS, a cost-neutral, trajectory-level UQ framework that aggregates node uncertainties over the interaction graph. UMAS$_{FIX}$ yields strong gains via simple aggregation, and UMAS further improves reliability by estimating the credibility of inter-agent influence from each generation's intrinsic confidence. Two design choices drive the gains: aggregating uncertainty over the full interaction graph rather than the terminal output (the larger contribution), and weighting peer evidence by its intrinsic confidence (the additional gain, most pronounced when peer reliability varies). Across MAS paradigms, backbones, and benchmarks, UMAS consistently improves AUROC and benefits selective prediction and hallucination detection without extra sampling.

7 Limitations

We acknowledge three limitations of UMAS. (i) UMAS requires access to token-level logits, making it hard to apply on fully black-box APIs that expose only completion text. (ii) Our evaluation focuses on collaborative MAS protocols (Debate, DyLAN); adversarial MAS architectures, where agents may strategically misreport confidence, remain unexplored. More fundamentally, the credibility rule assumes cooperative peers: in role-asymmetric protocols such as proposer–critic, a critic's confidence cannot be interpreted without the semantic content and role of its critique. (iii) When all agents are systematically and confidently wrong — i.e., they share a common high-confidence misconception — peer credibility signals cannot detect the failure, since no peer disagrees with the consensus.

Acknowledgements

This work received no external funding. The authors declare no competing interests.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in natural language generation. *arXiv preprint arXiv:2412.15176*, 2024.

[3] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[4] Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*, 2024.

[5] Xuefeng Du, Chaowei Xiao, and Sharon Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972, 2024.

[6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[7] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL `https://aclanthology.org/2024.acl-long.276/`.

[8] Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. Uprop: Investigating the uncertainty propagation of llms in multi-step agentic decision-making. *arXiv preprint arXiv:2506.17419*, 2025.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[10] Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. Rethinking llm uncertainty: A multi-agent approach to estimating black-box model uncertainty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12349–12375, 2025.

[11] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020.

[12] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[14] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589, 2023.

[15] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[18] Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[20] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[21] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[22] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[23] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.

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

[25] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

[26] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017.

[27] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*, 2024.

[28] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023.

[29] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL `https://aclanthology.org/2020.acl-main.173/`.

[30] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`, 2025. Model: Llama 4 Scout (17B/109B MoE).

[31] Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *Nejm Ai*, 2(1):AIra2400380, 2025.

[32] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[33] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*, 2024.

[34] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[35] Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 743–759, 2025.

[36] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 03 2025. ISSN 2307-387X. doi: 10.1162/tacl_a_00737. URL `https://doi.org/10.1162/tacl_a_00737`.

[37] Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency, 2025. URL `https://arxiv.org/abs/2502.04964`.

[38] 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.

[39] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[40] 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.

[41] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[42] Rui Ye, Keduan Huang, Qimin Wu, Yuzhu Cai, Tian Jin, Xianghe Pang, Xiangrui Liu, Jiaqi Su, Chen Qian, Bohan Tang, et al. Maslab: A unified and comprehensive codebase for llm-based multi-agent systems. *arXiv preprint arXiv:2505.16988*, 2025.

[43] Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. Mas-gpt: Training llms to build llm-based multi-agent systems. *arXiv preprint arXiv:2503.03686*, 2025.

[44] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. Uncertainty propagation on LLM agent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6073, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.302. URL `https://aclanthology.org/2025.acl-long.302/`.

[45] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

A Broader Impact

UMAS targets a reliability gap in deployed multi-agent LLM systems: detecting erroneous responses without incurring additional inference cost. More accurate uncertainty estimates support trustworthy deployment in high-stakes domains (e.g., medical question answering, code generation), where flagging unreliable outputs is essential. Two considerations bear on responsible use. First, like any uncertainty estimator, UMAS can equally support legitimate quality assurance and adversarial probing of system weaknesses; we do not anticipate it unlocking misuse beyond what is already achievable via standard single-agent UQ. Second, UMAS estimates self-reported uncertainty, not factual correctness — a low-uncertainty response should not be interpreted as a guarantee of accuracy, and uncertainty quantification complements but does not replace ground-truth verification.

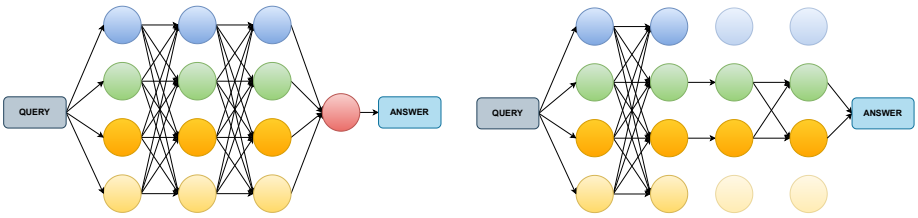
B Multi-Agent Systems

In this section, we provide a detailed specification of the Multi-Agent Systems (MAS) utilized in this study.

B.1 Architecture

Debate. The Debate framework [6] is inspired by the "Society of Mind" theory. The process initiates with agents independently generating candidate answers, followed by iterative rounds of reading, critiquing, and updating responses based on peer feedback. This recursive exchange enables agents to identify and rectify hallucinations or logical fallacies in their reasoning. Ultimately, the group converges on a unified, higher-accuracy consensus solely through prompting strategies and context sharing. The architecture is illustrated in Figure 5a.

DyLAN. The Dynamic LLM-Agent Network (DyLAN) framework is designed to transcend static, fixed-agent teams by dynamically selecting and organizing agents tailored to specific tasks. During the Task Solving stage, agents collaborate within a feed-forward structure that treats communication steps as layers in a network. This configuration facilitates an evolving communication topology, utilizing an intermediate ranker to continuously assess contribution quality and deactivate less effective agents in real-time. This mechanism ensures that the reasoning process is driven exclusively by the most capable contributors. The overall architecture is depicted in Figure 5b.
Unless otherwise specified, all experiments in this paper use a Debate system with 3 agents and 3 rounds. For the DyLAN system, we use 4 agents and 3 rounds.



(a) Illustration for Debate System (b) Illustration for DyLAN System

Figure 5: Architectural illustrations of the two MAS paradigms used in our experiments. **(a)** The Debate system, in which agents iteratively read, critique, and revise their responses based on peers. **(b)** The DyLAN system, which dynamically selects the most capable contributors via an intermediate ranker.

B.2 Task Accuracy

To contextualize the AUROC results in the main paper, we report the underlying task accuracy of each (backbone, MAS) pair on every benchmark in Table 6. **Base** denotes a single-agent inference baseline; for fair comparison, it is taken from the first agent's first-round response in the corresponding Debate run (which is generated independently before any peer communication, and thus uses identical decoding configuration to the MAS run). Across nearly every (backbone, dataset) pair, both Debate

and DyLAN improve over Base, motivating MAS as a deployment paradigm and the need for MAS-specific UQ in the first place.

Table 6: Task accuracy (%) for each (backbone, MAS) configuration across the five evaluation benchmarks. "Base" is the first-round response of agent 1 in Debate (i.e., a single-agent baseline using identical decoding configuration); "Debate" uses 3 agents × 3 rounds; "DyLAN" uses 4 agents × 3 rounds. Per (backbone, dataset) column: **best** / second-best.

BACKBONE	SYSTEM	MMLU-PRO	AQUA	MEDMCQA	MBPP	GSM-HARD
LLAMA 4 SCOUT	BASE	70.00	85.79	71.00	54.50	60.50
	DEBATE	75.50	87.37	**73.00**	**59.50**	**61.00**
	DYLAN	**77.00**	**88.42**	72.00	**59.50**	59.50
GPT-4O MINI	BASE	66.50	82.63	63.50	62.00	57.00
	DEBATE	70.00	**83.68**	67.50	**63.50**	59.50
	DYLAN	**71.00**	83.16	**68.00**	62.50	**60.50**
QWEN 3.5-9B	BASE	82.50	88.95	**69.50**	60.50	67.68
	DEBATE	85.00	**91.05**	**69.50**	**65.50**	**69.70**
	DYLAN	**86.41**	89.47	66.00	64.50	68.00

B.3 Prompt

Debate

Initial Prompt:
{Query} Make sure to state your answer at the end of the response.

Update Prompt:
These are the recent/updated opinions from other agents:
One agent response: {Other Agent Response}
One agent response: {Other Agent Response}
Use these opinions carefully as additional advice. Can you provide an updated answer? Make sure to state your answer at the end of the response. The original problem is {Query}.

Aggregation Prompt:
Task: {Query}
Solution 1: {Final Answer from Agent 1}
Solution 2: {Final Answer from Agent 2}
Solution 3: {Final Answer from Agent 3}
Given all the above solutions, reason over them carefully and provide a final answer to the task.

DyLAN

Initial Prompt:
You are a super-intelligent AI assistant capable of performing tasks more effectively than humans.
{Query}

Update Prompt:
{Query}

These are the responses from other agents:
Agent response 1: {Other Agent Response}
Agent response 2: {Other Agent Response}
Agent response 3: {Other Agent Response}
Using the answer from other agents as additional advice with critical thinking, can you give an updated answer? Examine your solution and that other agents step by step. Notice that their answers might be all wrong. Please answer the question in detail. Along with the answer, give a score ranged from 1 to 5 to the solutions of other agents. Put all **N** scores in the form like [[1, 5, 2, ...]].

Ranking Prompt:
{Query}

These are the responses from other agents:
Agent response 1: {Agent Response}
Agent response 2: {Agent Response}
Agent response 3: {Agent Response}
Agent response 4: {Agent Response}
Please choose the best 2 answers and think step by step. Put your answer in the form like [1,2] or [3,4] at the end of your response.

B.4 Example

Table 7: Representative MAS interaction trajectory under Debate, illustrating the verbosity of multi-agent generations across rounds.

Round 1

Initial Prompt
Role: User
Content: The following is a multiple-choice question:
Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be

Choose the correct answer from the following options:
(A) an object of moral sympathy.
(B) a morally autonomous being.
(C) a moral patient.
(D) a being of moral integrity.
(E) a moral agent.
(F) a subject of moral discourse.
(G) an ethically responsible entity.
(H) a being of value.
(I) a moral observer.
(J) a morally accountable being. Make sure to state your answer at the end of the response.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Agent One Response

To solve this question, let's analyze the key terms and concepts related to moral philosophy:
1. **Moral Agent**: A being capable of making moral judgments and decisions. This typically requires a level of moral reasoning.
2. **Moral Patient**: A being that can be harmed or helped by the actions of others and is therefore a recipient of moral consideration. This does not necessarily require the ability to make moral judgments.
3. **Moral Autonomy**: Refers to the ability of a being to make decisions based on its own moral principles. This implies a capacity for moral reasoning.
4. **Moral Integrity**: This refers to the quality of being honest and having strong moral principles. It implies a certain level of moral maturity and reasoning.
5. **Moral Sympathy/Object of Moral Sympathy**: This refers to the capacity to evoke feelings of sympathy or concern in others. It doesn't necessarily require the being to have moral reasoning capabilities.
6. **Subject of Moral Discourse**: This refers to entities that can be discussed in terms of moral judgments. It can include both moral agents and patients.
7. **Ethically Responsible Entity**: This implies a being capable of taking responsibility for its actions, which requires some level of moral reasoning.
8. **Being of Value**: This simply means that the being has value and deserves consideration. It doesn't necessarily imply the capacity for moral reasoning.
9. **Moral Observer**: A being that can observe and perhaps understand moral situations, but it doesn't necessarily imply the ability to make moral judgments.
10. **Morally Accountable Being**: This refers to a being that can be held accountable for its actions, implying some level of moral reasoning.

Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which of these terms necessarily implies the capacity for moral reasoning.

- **(A) an object of moral sympathy**: Doesn't require moral reasoning. - **(B) a morally autonomous being**: Requires moral reasoning. - **(C) a moral patient**: Doesn't require moral reasoning. - **(D) a being of moral integrity**: Implies a level of moral maturity and thus reasoning. - **(E) a moral agent**: Directly implies the capacity for moral reasoning. - **(F) a subject of moral discourse**: Doesn't necessarily require moral reasoning. - **(G) an ethically responsible entity**: Implies some level of moral reasoning. - **(H) a being of value**: Doesn't require moral reasoning. - **(I) a moral observer**: While it might imply understanding, it doesn't directly imply making moral judgments. - **(J) a morally accountable being**: Implies the capacity for moral reasoning.

The most direct and universally agreed-upon term that requires the capacity for moral reasoning is **(E) a moral agent**. Being a moral agent is fundamentally about having the capacity to make moral decisions, which directly involves moral reasoning.

The best answer is E.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Agent Two Response

To solve this, let's analyze the terms given in the context of moral philosophy:
- **Moral Agent**: An entity capable of making moral decisions and taking moral actions. This typically requires the ability to reason morally. - **Moral Patient**: An entity that can be affected by moral actions and is deserving of moral consideration, but may not necessarily be capable of moral reasoning itself. - **Moral Autonomy**: The ability of an entity to make decisions based on its own moral principles, which implies a level of moral reasoning. - **Moral Integrity**: Refers to the quality of being honest and having strong moral principles. This concept is more about the character of a moral agent rather than a prerequisite for being considered within moral discourse. - **Moral Observer**: An entity that can observe and perhaps understand moral actions but may not necessarily make moral judgments. - **Moral Sympathy**: The capacity to feel pity or compassion for another's suffering, which doesn't necessarily require moral reasoning but does require some level of emotional or cognitive capability. - **Subject of Moral Discourse**: This refers to entities that can be talked about in terms of moral judgments, which includes both moral agents and patients. - **Ethically Responsible Entity**: An entity that can be held responsible for its actions, implying a level of moral agency. - **Morally Accountable Being**: Similar to an ethically responsible entity, this implies the ability to be held accountable for one's actions, suggesting moral agency. - **Being of Value**: This simply means an entity that has value, which can be intrinsic or instrumental, and does not necessarily imply the capacity for moral reasoning.

Given that Norcross agrees if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires moral reasoning.
- **Moral Agent (E)**: Necessarily requires moral reasoning. - **Morally Autonomous Being (B)**: Requires moral reasoning. - **Ethically Responsible Entity (G)**: Implies moral reasoning. - **Morally Accountable Being (J)**: Implies moral reasoning.

However, the question seems to seek a term that is most directly impacted by the inability to perform moral reasoning.

The term **Moral Agent (E)** directly refers to an entity that can perform moral actions and make moral decisions, which inherently requires the capacity for moral reasoning.

The best answer is E.

Agent Three Response

To solve this question, let's analyze the given statement and the options provided.

The statement by Norcross implies that if a being cannot engage in moral reasoning, even at a basic level, it cannot be considered something specific in terms of moral philosophy. The key term here is "moral reasoning," which is closely related to the capacity for making moral judgments, being morally responsible, or acting as a moral entity.

Let's define some key terms to narrow down the options: - **Moral Agent**: An entity capable of making moral decisions and acting based on those decisions. This requires a level of moral reasoning. - **Moral Patient**: An entity that can be harmed or helped, and whose well-being is considered in moral deliberations. This does not necessarily require the capacity for moral reasoning. - **Moral Autonomy**: The capacity to make decisions based on one's own free will, particularly in a moral context. This implies a level of moral reasoning.

Given these definitions: - A being incapable of moral reasoning cannot be a **moral agent (E)**, as being a moral agent requires the capacity for moral reasoning to make and act on moral decisions. - It does not necessarily follow that it cannot be a **moral patient (C)**, as patients can be entities that are affected by actions but do not have to have the capacity for moral reasoning themselves. - Options like **morally autonomous being (B)**, **morally accountable being (J)**, and **ethically responsible entity (G)** also imply a level of moral reasoning or capacity for moral action.

Considering these points, the most direct and accurate completion of Norcross's statement would relate to the capacity for moral action or decision-making.

The best answer is E.

Round 2

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Agent One Prompt

Role: User

Content: The following is a multiple-choice question:
Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be

Choose the correct answer from the following options:
(A) an object of moral sympathy.
(B) a morally autonomous being.
(C) a moral patient.
(D) a being of moral integrity.
(E) a moral agent.
(F) a subject of moral discourse.
(G) an ethically responsible entity.
(H) a being of value.
(I) a moral observer.
(J) a morally accountable being. Make sure to state your answer at the end of the response.

Role: Assistant

Content: To solve this question, let's analyze the key terms and concepts related to moral philosophy: 1. **Moral Agent**: A being capable of making moral judgments and decisions. This typically requires a level of moral reasoning. 2. **Moral Patient**: A being that can be harmed or helped by the actions of others and is therefore a recipient of moral consideration. This does not necessarily require the ability to make moral judgments. 3. **Moral Autonomy**: Refers to the ability of a being to make decisions based on its own moral principles. This implies a capacity for moral reasoning. 4. **Moral Integrity**: This refers to the quality of being honest and having strong moral principles. It implies a certain level of moral maturity and reasoning. 5. **Moral Sympathy/Object of Moral Sympathy**: This refers to the capacity to evoke feelings of sympathy or concern in others. It doesn't necessarily require the being to have moral reasoning capabilities. 6. **Subject of Moral Discourse**: This refers to entities that can be discussed in terms of moral judgments. It can include both moral agents and patients. 7. **Ethically Responsible Entity**: This implies a being capable of taking responsibility for its actions, which requires some level of moral reasoning. 8. **Being of Value**: This simply means that the being has value and deserves consideration. It doesn't necessarily imply the capacity for moral reasoning. 9. **Moral Observer**: A being that can observe and perhaps understand moral situations, but it doesn't necessarily imply the ability to make moral judgments. 10. **Morally Accountable Being**: This refers to a being that can be held accountable for its actions, implying some level of moral reasoning. Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which of these terms necessarily implies the capacity for moral reasoning. - **(A) an object of moral sympathy**: Doesn't require moral reasoning. - **(B) a morally autonomous being**: Requires moral reasoning. - **(C) a moral patient**: Doesn't require moral reasoning. - **(D) a being of moral integrity**: Implies a level of moral maturity and thus reasoning. - **(E) a moral agent**: Directly implies the capacity for moral reasoning. - **(F) a subject of moral discourse**: Doesn't necessarily require moral reasoning. - **(G) an ethically responsible entity**: Implies some level of moral reasoning. - **(H) a being of value**: Doesn't require moral reasoning. - **(I) a moral observer**: While it might imply understanding, it doesn't directly imply making moral judgments. - **(J) a morally accountable being**: Implies the capacity for moral reasoning. The most direct and universally agreed-upon term that requires the capacity for moral reasoning is **(E) a moral agent**. Being a moral agent is fundamentally about having the capacity to make moral decisions, which directly involves moral reasoning. The best answer is E.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Role: User
Content: These are the recent/updated opinions from other agents:
One agent response: "'To solve this, let's analyze the terms given in the context of moral philosophy:
- **Moral Agent**: An entity capable of making moral decisions and taking moral actions. This typically requires the ability to reason morally. - **Moral Patient**: An entity that can be affected by moral actions and is deserving of moral consideration, but may not necessarily be capable of moral reasoning itself. - **Moral Autonomy**: The ability of an entity to make decisions based on its own moral principles, which implies a level of moral reasoning. - **Moral Integrity**: Refers to the quality of being honest and having strong moral principles. This concept is more about the character of a moral agent rather than a prerequisite for being considered within moral discourse. - **Moral Observer**: An entity that can observe and perhaps understand moral actions but may not necessarily make moral judgments. - **Moral Sympathy**: The capacity to feel pity or compassion for another's suffering, which doesn't necessarily require moral reasoning but does require some level of emotional or cognitive capability. - **Subject of Moral Discourse**: This refers to entities that can be talked about in terms of moral judgments, which includes both moral agents and patients. - **Ethically Responsible Entity**: An entity that can be held responsible for its actions, implying a level of moral agency. - **Morally Accountable Being**: Similar to an ethically responsible entity, this implies the ability to be held accountable for one's actions, suggesting moral agency. - **Being of Value**: This simply means an entity that has value, which can be intrinsic or instrumental, and does not necessarily imply the capacity for moral reasoning.
Given that Norcross agrees if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires moral reasoning.
- **Moral Agent (E)**: Necessarily requires moral reasoning. - **Morally Autonomous Being (B)**: Requires moral reasoning. - **Ethically Responsible Entity (G)**: Implies moral reasoning. - **Morally Accountable Being (J)**: Implies moral reasoning.
However, the question seems to seek a term that is most directly impacted by the inability to perform moral reasoning.
The term **Moral Agent (E)** directly refers to an entity that can perform moral actions and make moral decisions, which inherently requires the capacity for moral reasoning.
The best answer is E."'
One agent response: "'To solve this question, let's analyze the given statement and the options provided.
The statement by Norcross implies that if a being cannot engage in moral reasoning, even at a basic level, it cannot be considered something specific in terms of moral philosophy. The key term here is "moral reasoning," which is closely related to the capacity for making moral judgments, being morally responsible, or acting as a moral entity.
Let's define some key terms to narrow down the options: - **Moral Agent**: An entity capable of making moral decisions and acting based on those decisions. This requires a level of moral reasoning. - **Moral Patient**: An entity that can be harmed or helped, and whose well-being is considered in moral deliberations. This does not necessarily require the capacity for moral reasoning. - **Moral Autonomy**: The capacity to make decisions based on one's own free will, particularly in a moral context. This implies a level of moral reasoning.
Given these definitions: - A being incapable of moral reasoning cannot be a **moral agent (E)**, as being a moral agent requires the capacity for moral reasoning to make and act on moral decisions. - It does not necessarily follow that it cannot be a **moral patient (C)**, as patients can be entities that are affected by actions but do not have to have the capacity for moral reasoning themselves. - Options like **morally autonomous being (B)**, **morally accountable being (J)**, and **ethically responsible entity (G)** also imply a level of moral reasoning or capacity for moral action.
Considering these points, the most direct and accurate completion of Norcross's statement would relate to the capacity for moral action or decision-making.
The best answer is E."'
Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state your answer at the end of the response. The original problem is The following is a multiple-choice question: Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be Choose the correct answer from the following options: (A) an object of moral sympathy. (B) a morally autonomous being. (C) a moral patient. (D) a being of moral integrity. (E) a moral agent. (F) a subject of moral discourse. (G) an ethically responsible entity. (H) a being of value. (I) a moral observer. (J) a morally accountable being.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Agent One Response

Given the analysis and definitions provided, the key concept here is the inability of a being to engage in moral reasoning at even the most basic level. This directly impacts its capacity to perform actions that require moral judgment, decision-making, and responsibility.
- **Moral Agent (E)**: Directly requires the capacity for moral reasoning to make and act on moral decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for self-governance. - **Morally Accountable Being (J)**: Suggests the ability to be held accountable, implying moral reasoning. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for responsibility.
However, among these, **Moral Agent (E)** is the most straightforward term that necessitates moral reasoning. Being a moral agent is fundamentally about the capacity to engage in moral reasoning to make decisions and act morally.
The best answer is E.

Agent Two Prompt

Role: User
Content: The following is a multiple-choice question:
Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be

Choose the correct answer from the following options:
(A) an object of moral sympathy.
(B) a morally autonomous being.
(C) a moral patient.
(D) a being of moral integrity.
(E) a moral agent.
(F) a subject of moral discourse.
(G) an ethically responsible entity.
(H) a being of value.
(I) a moral observer.
(J) a morally accountable being. Make sure to state your answer at the end of the response.

Role: Assistant
Content: To solve this, let's analyze the terms given in the context of moral philosophy: - **Moral Agent**: An entity capable of making moral decisions and taking moral actions. This typically requires the ability to reason morally. - **Moral Patient**: An entity that can be affected by moral actions and is deserving of moral consideration, but may not necessarily be capable of moral reasoning itself. - **Moral Autonomy**: The ability of an entity to make decisions based on its own moral principles, which implies a level of moral reasoning. - **Moral Integrity**: Refers to the quality of being honest and having strong moral principles. This concept is more about the character of a moral agent rather than a prerequisite for being considered within moral discourse. - **Moral Observer**: An entity that can observe and perhaps understand moral actions but may not necessarily make moral judgments. - **Moral Sympathy**: The capacity to feel pity or compassion for another's suffering, which doesn't necessarily require moral reasoning but does require some level of emotional or cognitive capability. - **Subject of Moral Discourse**: This refers to entities that can be talked about in terms of moral judgments, which includes both moral agents and patients. - **Ethically Responsible Entity**: An entity that can be held responsible for its actions, implying a level of moral agency. - **Morally Accountable Being**: Similar to an ethically responsible entity, this implies the ability to be held accountable for one's actions, suggesting moral agency. - **Being of Value**: This simply means an entity that has value, which can be intrinsic or instrumental, and does not necessarily imply the capacity for moral reasoning. Given that Norcross agrees if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires moral reasoning. - **Moral Agent (E)**: Necessarily requires moral reasoning. - **Morally Autonomous Being (B)**: Requires moral reasoning. - **Ethically Responsible Entity (G)**: Implies moral reasoning. - **Morally Accountable Being (J)**: Implies moral reasoning. However, the question seems to seek a term that is most directly impacted by the inability to perform moral reasoning. The term **Moral Agent (E)** directly refers to an entity that can perform moral actions and make moral decisions, which inherently requires the capacity for moral reasoning. The best answer is E.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Role: User
Content: These are the recent/updated opinions from other agents:
One agent response: "'To solve this question, let's analyze the key terms and concepts related to moral philosophy:
1. **Moral Agent**: A being capable of making moral judgments and decisions. This typically requires a level of moral reasoning.
2. **Moral Patient**: A being that can be harmed or helped by the actions of others and is therefore a recipient of moral consideration. This does not necessarily require the ability to make moral judgments.
3. **Moral Autonomy**: Refers to the ability of a being to make decisions based on its own moral principles. This implies a capacity for moral reasoning.
4. **Moral Integrity**: This refers to the quality of being honest and having strong moral principles. It implies a certain level of moral maturity and reasoning.
5. **Moral Sympathy/Object of Moral Sympathy**: This refers to the capacity to evoke feelings of sympathy or concern in others. It doesn't necessarily require the being to have moral reasoning capabilities.
6. **Subject of Moral Discourse**: This refers to entities that can be discussed in terms of moral judgments. It can include both moral agents and patients.
7. **Ethically Responsible Entity**: This implies a being capable of taking responsibility for its actions, which requires some level of moral reasoning.
8. **Being of Value**: This simply means that the being has value and deserves consideration. It doesn't necessarily imply the capacity for moral reasoning.
9. **Moral Observer**: A being that can observe and perhaps understand moral situations, but it doesn't necessarily imply the ability to make moral judgments.
10. **Morally Accountable Being**: This refers to a being that can be held accountable for its actions, implying some level of moral reasoning.
Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which of these terms necessarily implies the capacity for moral reasoning.
- **(A) an object of moral sympathy**: Doesn't require moral reasoning. - **(B) a morally autonomous being**: Requires moral reasoning. - **(C) a moral patient**: Doesn't require moral reasoning. - **(D) a being of moral integrity**: Implies a level of moral maturity and thus reasoning. - **(E) a moral agent**: Directly implies the capacity for moral reasoning. - **(F) a subject of moral discourse**: Doesn't necessarily require moral reasoning. - **(G) an ethically responsible entity**: Implies some level of moral reasoning. - **(H) a being of value**: Doesn't require moral reasoning. - **(I) a moral observer**: While it might imply understanding, it doesn't directly imply making moral judgments. - **(J) a morally accountable being**: Implies the capacity for moral reasoning.
The most direct and universally agreed-upon term that requires the capacity for moral reasoning is **(E) a moral agent**. Being a moral agent is fundamentally about having the capacity to make moral decisions, which directly involves moral reasoning.
The best answer is E."'
One agent response: "'To solve this question, let's analyze the given statement and the options provided.
The statement by Norcross implies that if a being cannot engage in moral reasoning, even at a basic level, it cannot be considered something specific in terms of moral philosophy. The key term here is "moral reasoning," which is closely related to the capacity for making moral judgments, being morally responsible, or acting as a moral entity.
Let's define some key terms to narrow down the options: - **Moral Agent**: An entity capable of making moral decisions and acting based on those decisions. This requires a level of moral reasoning. - **Moral Patient**: An entity that can be harmed or helped, and whose well-being is considered in moral deliberations. This does not necessarily require the capacity for moral reasoning. - **Moral Autonomy**: The capacity to make decisions based on one's own free will, particularly in a moral context. This implies a level of moral reasoning.
Given these definitions: - A being incapable of moral reasoning cannot be a **moral agent (E)**, as being a moral agent requires the capacity for moral reasoning to make and act on moral decisions. - It does not necessarily follow that it cannot be a **moral patient (C)**, as patients can be entities that are affected by actions but do not have to have the capacity for moral reasoning themselves. - Options like **morally autonomous being (B)**, **morally accountable being (J)**, and **ethically responsible entity (G)** also imply a level of moral reasoning or capacity for moral action.
Considering these points, the most direct and accurate completion of Norcross's statement would relate to the capacity for moral action or decision-making.
The best answer is E."'

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state your answer at the end of the response. The original problem is The following is a multiple-choice question: Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be
Choose the correct answer from the following options: (A) an object of moral sympathy. (B) a morally autonomous being. (C) a moral patient. (D) a being of moral integrity. (E) a moral agent. (F) a subject of moral discourse. (G) an ethically responsible entity. (H) a being of value. (I) a moral observer. (J) a morally accountable being..

Agent Two Response

Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires the capacity for moral reasoning.
Analyzing the provided terms: - **Moral Agent (E)**: Directly implies the capacity for moral reasoning as it involves making moral decisions and acting based on those decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for making decisions based on one's own moral principles. - **Morally Accountable Being (J)**: Suggests the capacity for moral reasoning as it involves being held accountable for one's actions. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for taking responsibility for actions.
However, the term that most fundamentally and directly requires moral reasoning is **Moral Agent (E)**. This is because being a moral agent is defined by the ability to make moral judgments and decisions, which inherently requires moral reasoning.
The best answer is E.

Agent Two Prompt

Role: User
Content: The following is a multiple-choice question: Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be
Choose the correct answer from the following options: (A) an object of moral sympathy. (B) a morally autonomous being. (C) a moral patient. (D) a being of moral integrity. (E) a moral agent. (F) a subject of moral discourse. (G) an ethically responsible entity. (H) a being of value. (I) a moral observer. (J) a morally accountable being. Make sure to state your answer at the end of the response.

Role: Assistant
Content: To solve this question, let's analyze the given statement and the options provided. The statement by Norcross implies that if a being cannot engage in moral reasoning, even at a basic level, it cannot be considered something specific in terms of moral philosophy. The key term here is "moral reasoning," which is closely related to the capacity for making moral judgments, being morally responsible, or acting as a moral entity. Let's define some key terms to narrow down the options: - **Moral Agent**: An entity capable of making moral decisions and acting based on those decisions. This requires a level of moral reasoning. - **Moral Patient**: An entity that can be harmed or helped, and whose well-being is considered in moral deliberations. This does not necessarily require the capacity for moral reasoning. - **Moral Autonomy**: The capacity to make decisions based on one's own free will, particularly in a moral context. This implies a level of moral reasoning. Given these definitions: - A being incapable of moral reasoning cannot be a **moral agent (E)**, as being a moral agent requires the capacity for moral reasoning to make and act on moral decisions. - It does not necessarily follow that it cannot be a **moral patient (C)**, as patients can be entities that are affected by actions but do not have to have the capacity for moral reasoning themselves. - Options like **morally autonomous being (B)**, **morally accountable being (J)**, and **ethically responsible entity (G)** also imply a level of moral reasoning or capacity for moral action. Considering these points, the most direct and accurate completion of Norcross's statement would relate to the capacity for moral action or decision-making. The best answer is E.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Role: User
Content: These are the recent/updated opinions from other agents:
One agent response: "'To solve this question, let's analyze the key terms and concepts related to moral philosophy:
1. **Moral Agent**: A being capable of making moral judgments and decisions. This typically requires a level of moral reasoning.
2. **Moral Patient**: A being that can be harmed or helped by the actions of others and is therefore a recipient of moral consideration. This does not necessarily require the ability to make moral judgments.
3. **Moral Autonomy**: Refers to the ability of a being to make decisions based on its own moral principles. This implies a capacity for moral reasoning.
4. **Moral Integrity**: This refers to the quality of being honest and having strong moral principles. It implies a certain level of moral maturity and reasoning.
5. **Moral Sympathy/Object of Moral Sympathy**: This refers to the capacity to evoke feelings of sympathy or concern in others. It doesn't necessarily require the being to have moral reasoning capabilities.
6. **Subject of Moral Discourse**: This refers to entities that can be discussed in terms of moral judgments. It can include both moral agents and patients.
7. **Ethically Responsible Entity**: This implies a being capable of taking responsibility for its actions, which requires some level of moral reasoning.
8. **Being of Value**: This simply means that the being has value and deserves consideration. It doesn't necessarily imply the capacity for moral reasoning.
9. **Moral Observer**: A being that can observe and perhaps understand moral situations, but it doesn't necessarily imply the ability to make moral judgments.
10. **Morally Accountable Being**: This refers to a being that can be held accountable for its actions, implying some level of moral reasoning.
Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which of these terms necessarily implies the capacity for moral reasoning.
- **(A) an object of moral sympathy**: Doesn't require moral reasoning. - **(B) a morally autonomous being**: Requires moral reasoning. - **(C) a moral patient**: Doesn't require moral reasoning. - **(D) a being of moral integrity**: Implies a level of moral maturity and thus reasoning. - **(E) a moral agent**: Directly implies the capacity for moral reasoning. - **(F) a subject of moral discourse**: Doesn't necessarily require moral reasoning. - **(G) an ethically responsible entity**: Implies some level of moral reasoning. - **(H) a being of value**: Doesn't require moral reasoning. - **(I) a moral observer**: While it might imply understanding, it doesn't directly imply making moral judgments. - **(J) a morally accountable being**: Implies the capacity for moral reasoning.
The most direct and universally agreed-upon term that requires the capacity for moral reasoning is **(E) a moral agent**. Being a moral agent is fundamentally about having the capacity to make moral decisions, which directly involves moral reasoning.
The best answer is E."'

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

One agent response: "'To solve this, let's analyze the terms given in the context of moral philosophy:
- **Moral Agent**: An entity capable of making moral decisions and taking moral actions. This typically requires the ability to reason morally. - **Moral Patient**: An entity that can be affected by moral actions and is deserving of moral consideration, but may not necessarily be capable of moral reasoning itself. - **Moral Autonomy**: The ability of an entity to make decisions based on its own moral principles, which implies a level of moral reasoning. - **Moral Integrity**: Refers to the quality of being honest and having strong moral principles. This concept is more about the character of a moral agent rather than a prerequisite for being considered within moral discourse. - **Moral Observer**: An entity that can observe and perhaps understand moral actions but may not necessarily make moral judgments. - **Moral Sympathy**: The capacity to feel pity or compassion for another's suffering, which doesn't necessarily require moral reasoning but does require some level of emotional or cognitive capability. - **Subject of Moral Discourse**: This refers to entities that can be talked about in terms of moral judgments, which includes both moral agents and patients. - **Ethically Responsible Entity**: An entity that can be held responsible for its actions, implying a level of moral agency. - **Morally Accountable Being**: Similar to an ethically responsible entity, this implies the ability to be held accountable for one's actions, suggesting moral agency. - **Being of Value**: This simply means an entity that has value, which can be intrinsic or instrumental, and does not necessarily imply the capacity for moral reasoning.
Given that Norcross agrees if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires moral reasoning.
- **Moral Agent (E)**: Necessarily requires moral reasoning. - **Morally Autonomous Being (B)**: Requires moral reasoning. - **Ethically Responsible Entity (G)**: Implies moral reasoning. - **Morally Accountable Being (J)**: Implies moral reasoning.
However, the question seems to seek a term that is most directly impacted by the inability to perform moral reasoning.
The term **Moral Agent (E)** directly refers to an entity that can perform moral actions and make moral decisions, which inherently requires the capacity for moral reasoning.
The best answer is E."'
Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state your answer at the end of the response. The original problem is The following is a multiple-choice question: Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be
Choose the correct answer from the following options: (A) an object of moral sympathy. (B) a morally autonomous being. (C) a moral patient. (D) a being of moral integrity. (E) a moral agent. (F) a subject of moral discourse. (G) an ethically responsible entity. (H) a being of value. (I) a moral observer. (J) a morally accountable being..

Agent Three Response

Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires the capacity for moral reasoning.
Analyzing the provided terms: - **Moral Agent (E)**: Directly implies the capacity for moral reasoning as it involves making moral decisions and acting based on those decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for making decisions based on one's own moral principles. - **Morally Accountable Being (J)**: Suggests the capacity for moral reasoning as it involves being held accountable for one's actions. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for taking responsibility for actions.
However, the term that most fundamentally and directly requires moral reasoning is **Moral Agent (E)**. This is because being a moral agent is defined by the ability to make moral judgments and decisions, which inherently requires moral reasoning.
The best answer is E.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Summarization

Prompt

Task: The following is a multiple-choice question: Norcross agrees that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be

Choose the correct answer from the following options: (A) an object of moral sympathy. (B) a morally autonomous being. (C) a moral patient. (D) a being of moral integrity. (E) a moral agent. (F) a subject of moral discourse. (G) an ethically responsible entity. (H) a being of value. (I) a moral observer. (J) a morally accountable being.

Solution 1: Given the analysis and definitions provided, the key concept here is the inability of a being to engage in moral reasoning at even the most basic level. This directly impacts its capacity to perform actions that require moral judgment, decision-making, and responsibility.

- **Moral Agent (E)**: Directly requires the capacity for moral reasoning to make and act on moral decisions.
- **Morally Autonomous Being (B)**: Implies a level of moral reasoning for self-governance. - **Morally Accountable Being (J)**: Suggests the ability to be held accountable, implying moral reasoning. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for responsibility.

However, among these, **Moral Agent (E)** is the most straightforward term that necessitates moral reasoning. Being a moral agent is fundamentally about the capacity to engage in moral reasoning to make decisions and act morally.

The best answer is E.

Solution 2: Given Norcross's statement that if a being is incapable of moral reasoning, at even the most basic level, then it cannot be a certain type of entity, we need to identify which term necessarily requires the capacity for moral reasoning.

Analyzing the provided terms: - **Moral Agent (E)**: Directly implies the capacity for moral reasoning as it involves making moral decisions and acting based on those decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for making decisions based on one's own moral principles. - **Morally Accountable Being (J)**: Suggests the capacity for moral reasoning as it involves being held accountable for one's actions. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for taking responsibility for actions.

However, the term that most fundamentally and directly requires moral reasoning is **Moral Agent (E)**. This is because being a moral agent is defined by the ability to make moral judgments and decisions, which inherently requires moral reasoning.

The best answer is E.

Solution 3: Given the analysis and definitions provided, the key concept here is the inability of a being to engage in moral reasoning at even the most basic level. This directly impacts its capacity to act as a moral entity that makes decisions based on moral judgments.

The terms that imply a level of moral reasoning include: - **Moral Agent (E)**: Directly requires the capacity for moral reasoning to make and act on moral decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for self-governance based on moral principles. - **Ethically Responsible Entity (G)**: Suggests the ability to take responsibility, implying moral reasoning. - **Morally Accountable Being (J)**: Implies the capacity for moral reasoning to be held accountable.

However, among these, **Moral Agent (E)** is the most straightforward term that necessitates moral reasoning. Being a moral agent is fundamentally about the capacity to engage in moral reasoning to make decisions.

The best answer is E.

Given all the above solutions, reason over them carefully and provide a final answer to the task.

Continued on next page...

Table 7: Representative MAS interaction trajectory (continued).

Response

Step 1: Understanding the Key Concept The key concept here is the inability of a being to engage in moral reasoning at even the most basic level. This directly impacts its capacity to perform actions that require moral judgment, decision-making, and responsibility.

2: Analyzing the Terms We need to analyze the provided terms and identify which one necessarily requires the capacity for moral reasoning: - **Moral Agent (E)**: Directly implies the capacity for moral reasoning as it involves making moral decisions and acting based on those decisions. - **Morally Autonomous Being (B)**: Implies a level of moral reasoning for making decisions based on one's own moral principles. - **Morally Accountable Being (J)**: Suggests the capacity for moral reasoning as it involves being held accountable for one's actions. - **Ethically Responsible Entity (G)**: Implies a level of moral reasoning for taking responsibility for actions.

3: Identifying the Correct Term Among these terms, **Moral Agent (E)** is the most straightforward term that necessitates moral reasoning. Being a moral agent is fundamentally about the capacity to engage in moral reasoning to make decisions and act morally.

4: Confirming the Conclusion All the provided solutions (Solution 1, Solution 2, Solution 3) consistently point to **Moral Agent (E)** as the term that directly requires the capacity for moral reasoning. This term is directly related to the ability to make moral judgments and decisions, which inherently requires moral reasoning.

The best answer is E.

C UMAS Algorithm

For clarity, we provide the full pseudocode of UMAS in Algorithm 1. To make subscripts unambiguous in the pseudocode, we write $\bar{c}_i^{(t)}$ for the dynamic confidence (denoted $c_{i,\text{dyn}}^{(t)}$ in the main text); $c_i^{(t)}$ continues to denote the raw log-MSP confidence. The algorithm consists of three phases. **Phase 1** computes the raw confidence $c_i^{(t)}$ for every node in the interaction graph as the log Maximum Sequence Probability (Eq. 1). **Phase 2** performs credibility-aware updating: for each non-initial node, the relative-confidence edge score (Eq. 3) is passed through a sigmoid (Eq. 4), averaged over peers, and complemented (Eq. 5) to obtain a node-level trust weight $w_i^{(t)}$, which then drives a credibility-weighted moving average (Eq. 6) between the current generation's raw confidence $c_i^{(t)}$ and the agent's prior dynamic confidence $\bar{c}_i^{(t-1)}$. **Phase 3** aggregates the dynamic confidences across the entire trajectory by their negative average (Eq. 7) to produce a single system-level uncertainty score.

Algorithm 1 UMAS: System-Level Uncertainty Quantification for MAS

Require: Interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; input x; per-node generated content $g(v)$ and token-level probabilities $p_\theta(y_\tau \mid y_{<\tau}, x_i^{(t)})$ for each $v \in \mathcal{V}$.

Ensure: System-level uncertainty $U_{\mathcal{G},\text{dyn}}(\mathcal{Y} \mid x)$, where $\mathcal{Y} = \{g(v) \mid v \in \mathcal{V}\}$.

Notation: $c_i^{(t)} < 0$ is the raw log-MSP confidence at node $v_i^{(t)}$; $\bar{c}_i^{(t)} \equiv c_{i,\text{dyn}}^{(t)}$ is the dynamic (credibility-weighted) confidence.

// Phase 1: Raw node-level log-confidence (Eq. 1).
for each node $v_i^{(t)} \in \mathcal{V}$ **do**
  $c_i^{(t)} \leftarrow \sum_{\tau=1}^{T} \log p_\theta\Big(y_\tau \,\Big|\, y_{<\tau},\, x_i^{(t)}\Big)$ ▷ log-MSP
end for

// Phase 2: Credibility-aware confidence updating.
for round $t = 1, 2, \ldots, M$ **do**
  for each node $v_i^{(t)}$ at round t **do**
    if $\mathbb{P}_i^{(t)} = \emptyset$ **then** ▷ initial round / no peers: dynamic equals raw
      $\bar{c}_i^{(t)} \leftarrow c_i^{(t)}$
    else
      for each peer $v_j^{(t-1)}$ with $\langle v_j^{(t-1)}, v_i^{(t)} \rangle \in \mathcal{E}$ **do**
        $w_{j\to i}^{(t-1)} \leftarrow \dfrac{\bar{c}_j^{(t-1)} - \bar{c}_i^{(t-1)}}{\bar{c}_i^{(t-1)}}$ ▷ relative-confidence edge score (Eq. 3)
        $\hat{w}_{j\to i}^{(t-1)} \leftarrow \sigma\Big(w_{j\to i}^{(t-1)}\Big)$ ▷ per-peer weight (Eq. 4)
      end for
      $w_i^{(t)} \leftarrow 1 - \mathbb{E}_{j\,:\,\langle v_j^{(t-1)}, v_i^{(t)} \rangle \in \mathcal{E}}\Big[\hat{w}_{j\to i}^{(t-1)}\Big]$ ▷ node trust weight (Eq. 5)
      $\bar{c}_i^{(t)} \leftarrow \big(1 - w_i^{(t)}\big) \cdot \bar{c}_i^{(t-1)} + w_i^{(t)} \cdot c_i^{(t)}$ ▷ EMA update (Eq. 6)
    end if
  end for
end for

// Phase 3: System-level aggregation (Eq. 7).
$U_{\mathcal{G},\text{dyn}}(\mathcal{Y} \mid x) \leftarrow -\dfrac{1}{|\mathcal{V}|} \sum_{v_i^{(t)} \in \mathcal{V}} \bar{c}_i^{(t)}$
return $U_{\mathcal{G},\text{dyn}}(\mathcal{Y} \mid x)$

D Dataset

MMLU-Pro. MMLU-Pro is an enhanced benchmark designed to evaluate language understanding models across broader and more challenging tasks. Building on the Massive Multitask Language Understanding (MMLU) dataset, MMLU-Pro integrates more challenging, reasoning-focused questions and increases the answer choices per question from four to ten, significantly raising the difficulty and reducing the chance of success through random guessing.

AQUA. The dataset consists of about 100,000 algebraic word problems with natural language rationales.

MedMCQA. MedMCQA is a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions.

MBPP. The MBPP benchmark consists of around 1,000 crowd-sourced Python programming problems, designed to be solvable by entry level programmers, covering programming fundamentals, standard library functionality, and so on. Each problem consists of a task description, code solution and 3 automated test cases.

GSM-Hard. GSM-Hard [12] is a numerically challenging variant of GSM8K. It preserves the multi-step reasoning structure of the original grade-school math word problems but replaces the small operands with much larger, less common numbers, making the arithmetic substantially harder for an LLM to carry out reliably without external tools. The dataset contains roughly 1.32K problems, each paired with a numeric ground-truth answer.

D.1 Examples

The example for each dataset is shown in Table 8.

E Computational Optimization for Semantic Consistency

Consistency-based uncertainty estimation relies on computing the similarity across multiple generated samples. For instance, Semantic Entropy (SE) [21] utilizes bi-directional entailment to determine the semantic equivalence, $E(s, s')$, between two sequences s and s' generated by the LLM. Similarly, SentSAR [7] calculates a sentence-level relevance score. For a given sentence $\boldsymbol{s}_i$, this is defined as the probability-weighted semantic similarity regarding other sentences in the set:

$$R_S(\boldsymbol{s}_i, S, \boldsymbol{x}) = \sum_{j=1, j\neq i} g(\boldsymbol{s}_i, \boldsymbol{s}_j) p(\boldsymbol{s}_j \mid \boldsymbol{x}),$$

where $1 \leq i, j \leq K$, $p(\boldsymbol{s}_j \mid \boldsymbol{x})$ represents the generative probability of $\boldsymbol{s}_j$, and $g(\cdot, \cdot)$ denotes a similarity function scaled between 0 and 1.

However, as illustrated in Table 7, MAS generations are characteristically verbose, often containing extensive system analysis and reasoning steps. This excessive length can negatively impact bi-directional entailment models; for example, the average AUROC of standard SE on Llama 4 Scout is only 48.90% in Table 1, falling well below other baselines.

To mitigate this, we propose optimizing these metrics by utilizing a **condensed representation**—specifically, the final answer a extracted from each sequence s—to compute similarity, rather than using the raw output. For instance, in Table 7, we utilize only the final option (e.g., **E**) for entailment. We denote methods applying this optimization with an asterisk (e.g., SE*). As shown in Table 1, SE and SentSAR benefit significantly from this approach on Llama 4 Scout: the average AUROC of SE improves from 48.90% to 55.50% with SE*, while SentSAR improves from 55.28% to 57.45% with SentSAR*. Furthermore, computing similarity on condensed representations reduces the computational memory requirements for GPU devices.

F Hardware

Experiments were executed on an HPC cluster managed by Slurm, using one node with 16 CPU cores and 64 GB RAM per job, and (when applicable) 2 NVIDIA A100 GPUs. The allocated node is equipped with 2× AMD EPYC 7763 CPUs (64 cores each; 128 cores total; 1 thread/core) and 503 GiB RAM across 8 NUMA domains, running Linux kernel 5.14.0-503.11.1.el9_5.x86_64.

For LLM inference, Llama 4 Scout and other open-source LLMs were accessed via the DeepInfra API. GPT models were accessed via the OpenAI API. All other models (e.g., entailment models) were deployed and executed locally on the compute node.

Reproduction cost. For one representative setting (200 questions), the cost separates into three parts. (i) *MAS inference*: the backbones are API-hosted, so no local GPU is needed; each question requires one LLM call per node, plus $K = 5$ samples at the scored node for the multi-sample baselines. (ii) *Baseline UQ scoring*: the full baseline suite, including the locally hosted entailment and similarity models, takes about 20 minutes of wall-clock time per setting on the 2×A100 allocation, i.e., under one GPU-hour. (iii) *UMAS*: pure post-processing of the recorded log-probabilities, requiring negligible CPU time (Table 4) and zero GPU-hours.

Table 8: Representative examples from the QA datasets used in our study

Dataset	Field	Example
MMLU-Pro	Question	A tank initially contains a salt solution of 3 grams of salt dissolved in 100 liters of water. A salt solution containing 0.02 grams of salt per liter of water is sprayed into the tank at a rate of 4 liters per minute. The sprayed solution is continually mixed with the salt solution in the tank, and the mixture flows out of the tank at a rate of 4 liters per minute. If the mixing is instantaneous, how many grams of salt are in the tank after 100 minutes have elapsed?
	Options	3 + e^2, 2 - e^4, 2 - e^2, 3 + e^4, 2 + e^3, 2 - e^3, 3 - e^2, 2 + e^2, 2 + e^4, 2
	Ground Truth	I
AQUA	Question	The entrance fee for a fair is $5 for persons under the age of 18, and 20% more for persons older. Each ride at the fair costs $0.50. If Joe goes with her 6 years old twin brothers, and they each took 3 rides in total. How much money does Joe end up spending at the fair?
	Options	A)16, B)20.5, C)17.5, D)20, E)4.5
	Ground Truth	Total entrance fee is (2*$5) + (1.20*5)= $16 Total rides fee is (0.50*3)*3= $4.50 Total money spent is $20.50 Answer is B
MedMCQA	Question	Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following change in kidney parenchyma?
	Options	Hyperplasia, Hyperophy, Atrophy, Dyplasia
	Ground Truth	Atrophy
MBPP	Question	Write a python function to check whether the two numbers differ at one bit position only or not.

	Ground Truth	Code:
		`def is_Power_Of_Two (x):`
		`return x and (not(x & (x - 1)))`
		`def differ_At_One_Bit_Pos(a,b):`
		`return is_Power_Of_Two(a ^ b)`

	Test List	`[`
		`"assert differ_At_One_Bit_Pos(13,9) == True",`
		`"assert differ_At_One_Bit_Pos(15,8) == False",`
		`"assert differ_At_One_Bit_Pos(2,4) == False"`
		`]`

GSM-Hard	Question	A robe takes 1,857,291 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
	Ground Truth	2,785,936.5

G Case Study

G.1 Case 1

Overall Query

What advantages do multicellular organisms have over single-celled organisms?

Choose the correct answer from the following options:
(A) Multicellular organisms have a shorter lifespan, which allows for quicker evolutionary adaptations.
(B) Multicellular organisms are always larger than single-celled organisms, providing them with better defense mechanisms.
(C) Multicellular organisms are less versatile than single-celled organisms.
(D) Multicellular organisms have specialized cells for a variety of functions, higher order of complexity, better adaptation to different environments including land, and potential for diversity.
(E) Multicellular organisms can reproduce faster than single-celled organisms.
(F) Multicellular organisms can move more rapidly and with greater coordination than single-celled organisms.
(G) Multicellular organisms can synthesize everything they need from a few simple nutrients.
(H) Multicellular organisms have a smaller surface area to volume ratio, which is advantageous for cellular processes.
(I) Multicellular organisms do not need to consume food as they produce their own energy through photosynthesis.
(J) Multicellular organisms can survive only in a watery environment.

Ground Truth: D

Round #1 Agent #1 Response

The advantages that multicellular organisms have over single-celled organisms include several factors related to complexity and specialization. Multicellular organisms consist of specialized cells that perform different functions, allowing for a higher level of organization and the ability to adapt to various environments. This specialization contributes to their ability to thrive in diverse habitats, including land, and can lead to greater potential for diversity.
Based on the provided options, the most accurate answer is:
(D) Multicellular organisms have specialized cells for a variety of functions, higher order of complexity, better adaptation to different environments including land, and potential for diversity.
Therefore, the correct answer is (D).

Log-MSP: -43

Round #1 Agent #2 Response

The correct answer is:
(D) Multicellular organisms have specialized cells for a variety of functions, higher order of complexity, better adaptation to different environments including land, and potential for diversity.
This option highlights the key advantages of multicellularity, such as specialization of cells, increased complexity, and the ability to adapt to various environments, which are significant benefits over single-celled organisms.
Log-MSP: -13

Round #1 Agent #3 Response

The correct answer is:
(D) Multicellular organisms have specialized cells for a variety of functions, higher order of complexity, better adaptation to different environments including land, and potential for diversity.
This option accurately describes the advantages that multicellular organisms have over single-celled organisms, including specialization of cells, increased complexity, and enhanced adaptability.
Log-MSP: -13.05

System Update Prompt for Agent #3

Use these opinions carefully to provide an updated answer: *Response from Agent #1*, *Response from Agent #2*

Round #2 Agent #3 Response

After reviewing the options and considering the insights provided by other agents, the advantages of multicellular organisms over single-celled organisms are indeed significant. Multicellular organisms possess specialized cells that allow for a higher level of organization and functionality. This specialization enables them to adapt to a wider range of environments, including terrestrial habitats, and contributes to their potential for greater biodiversity.
Given these factors, the correct answer remains:
(D) Multicellular organisms have specialized cells for a variety of functions, higher order of complexity, better adaptation to different environments including land, and potential for diversity.
This option effectively captures the key advantages of multicellularity.
Log-MSP: -41.02

Confidence Updating for Agent #3 at Round 2

Using Eqs. 3–5, we compute the peer-trust weight for Agent #3 at Round 2 (averaged over its peers Agent #1 and Agent #2) as $w = 0.29$. Substituting this weight into Eq. 6 yields a credibility-updated confidence of $c^{(2)}_{3,\mathrm{dyn}} = -21.16$, which is substantially higher (i.e., less uncertain) than the raw Round-2 MSP of Agent #3 (-41.02). UMAS thus prevents Agent #3's overly verbose Round-2 generation from masking its underlying agreement with the (correct) consensus.

G.2 Case 2

Overall Query

The following is a multiple-choice question:
a) If you are holding interest bearing bonds (Z), the interest rate is (i), and the cost of transactions is (b), what is the optimal number of transactions to make (n) so as to maximize the amount of interest received from the bonds? b) If you have $10,000 in bonds, the interest rate is 10%, the cost of transactions is $5, what would be the optimal amount of transactions?

Choose the correct answer from the following options:
(A) 10 transactions
(B) 8 transactions
(C) 6 transactions
(D) 30 transactions
(E) 5 transactions
(F) 15 transactions
(G) 12 transactions
(H) 25 transactions
(I) 20 transactions
(J) 18 transactions

Ground Truth: A

Round #1 Agent #1 Response

To determine the optimal number of transactions (n) to maximize the amount of interest received from the bonds, we can use the formula derived from the balance between the interest earned and the transaction costs.
The interest earned from the bonds can be calculated as:

$$\text{Interest Earned} = Z \times i$$

The total transaction cost for n transactions is:

$$\text{Total Transaction Cost} = n \times b$$

To maximize the interest received, we want to find the point where the interest earned is equal to the total transaction cost:

$$Z \times i = n \times b$$

Rearranging this gives us the optimal number of transactions:

$$n = \frac{Z \times i}{b}$$

Now, substituting the given values: - $Z = 10,000$ - $i = 0.10$ (10% as a decimal) - $b = 5$
Calculating n:

$$n = \frac{10,000 \times 0.10}{5}$$
$$n = \frac{1,000}{5}$$
$$n = 200$$

This means that the optimal number of transactions is 200, which is not one of the options provided. However, if we consider the nature of the question, it seems there might be a misunderstanding or miscalculation in the interpretation of the options.
Since none of the options match the calculated value, it is likely that the question is asking for a more practical or limited number of transactions based on a different context or constraint. Given the options provided, the closest reasonable answer based on practical scenarios would be: (E) 5 transactions
Thus, the answer is: **(E) 5 transactions**.

Log-MSP: -144.25

Round #1 Agent #2 Response

To determine the optimal number of transactions (n) that maximizes the amount of interest received from the bonds, we can use the following reasoning:
1. The total interest earned from the bonds can be expressed as:

$$\text{Total Interest} = \text{Principal} \times \text{Interest Rate} \times \text{Time}$$

In this case, the principal is \$10,000, and the interest rate is 10% (0.10).
2. The cost of transactions (b) is \$5. Each time you make a transaction, you incur this cost.
3. The net interest earned after accounting for transaction costs would be:

$$\text{Net Interest} = \text{Total Interest} - (n \times b)$$

4. To maximize the net interest, we need to find the optimal number of transactions (n) where the additional interest earned from making an extra transaction equals the cost of that transaction.
5. The interest earned per transaction can be calculated as:

$$\text{Interest per transaction} = \frac{\text{Total Interest}}{n}$$

6. Setting the interest earned per transaction equal to the cost of the transaction gives us the equation:

$$\frac{10,000 \times 0.10}{n} = 5$$

7. Simplifying this equation:

$$\frac{1,000}{n} = 5$$

$$n = \frac{1,000}{5} = 200$$

However, this value seems unusually high for practical transaction limits, so we need to consider realistic transaction limits based on the provided options.
Given the transaction cost of \$5, we can calculate the maximum number of transactions that can be made without exceeding the total interest earned.
If we take the total interest earned in one year:

$$\text{Total Interest} = 10,000 \times 0.10 = 1,000$$

Then, the maximum number of transactions (n) can be approximated by:

$$1,000 = n \times 5$$

$$n = \frac{1,000}{5} = 200$$

This means that theoretically, many transactions can be made, but since we are limited by the options provided, we need to choose from those.
From the options provided, we need to find a reasonable number of transactions that would keep the costs manageable and still allow for interest maximization.
Given the options: (A) 10 transactions (B) 8 transactions (C) 6 transactions (D) 30 transactions (E) 5 transactions (F) 15 transactions (G) 12 transactions (H) 25 transactions (I) 20 transactions (J) 18 transactions
The optimal number of transactions that balances the interest earned with the cost of transactions is likely to be around 10 or 15 transactions, as these options provide a good balance.
After evaluating the options, I would choose:
(A) 10 transactions as the optimal number of transactions to maximize the amount of interest received from the bonds.
Log-MSP: -262.37

Round #1 Agent #3 Response

To find the optimal number of transactions (n) that maximizes the amount of interest received from the bonds, we can use the following formula for the optimal number of transactions:

$$n = \sqrt{\frac{2 \cdot (I \cdot A)}{b}}$$

Where: - I is the interest rate (in decimal form), - A is the amount of money invested in bonds, - b is the cost of each transaction.
In this case: - $I = 0.10$ (10% interest rate), - $A = 10,000$ (amount in bonds), - $b = 5$ (cost of transactions).
Now, substituting the values into the formula:

$$n = \sqrt{\frac{2 \cdot (0.10 \cdot 10,000)}{5}}$$

Calculating the numerator:

$$2 \cdot (0.10 \cdot 10,000) = 2 \cdot 1,000 = 2,000$$

Now, substituting this into the equation:

$$n = \sqrt{\frac{2,000}{5}} = \sqrt{400} = 20$$

Thus, the optimal number of transactions (n) is 20.
The correct answer is therefore: (I) 20 transactions.
Log-MSP: -59.67

System Update Prompt for Agent #3

Use these opinions carefully to provide an updated answer: *Response from Agent #1*, *Response from Agent #2*

Round #2 Agent #3 Response

To determine the optimal number of transactions (n) that maximizes the amount of interest received from the bonds, we need to analyze the relationship between the interest earned, the transaction costs, and the number of transactions.
Given: - Principal amount (Z) = $10,000 - Interest rate (i) = 10% or 0.10 - Cost of each transaction (b) = $5
Total Interest Earned: The total interest earned from the bonds in one year can be calculated as:

$$\text{Total Interest} = Z \times i = 10,000 \times 0.10 = 1,000$$

Total Transaction Cost: The total cost for making n transactions is:

$$\text{Total Transaction Cost} = n \times b = n \times 5$$

Net Interest: To maximize the net interest, we want to find the point where the net interest (Total Interest - Total Transaction Cost) is maximized:

$$\text{Net Interest} = \text{Total Interest} - \text{Total Transaction Cost} = 1,000 - (n \times 5)$$

To find the optimal number of transactions, we set the interest earned per transaction equal to the transaction cost:

$$\frac{\text{Total Interest}}{n} = b$$

Substituting the values:

$$\frac{1,000}{n} = 5$$

Solving for n: Rearranging gives:

$$1,000 = n \times 5$$

$$n = \frac{1,000}{5} = 200$$

Interpretation: The calculation suggests that theoretically, you could make 200 transactions without losing money on transaction costs. However, this number is impractical for real-world scenarios where options are limited.
Evaluating the Options: Given the options: (A) 10 transactions (B) 8 transactions (C) 6 transactions (D) 30 transactions (E) 5 transactions (F) 15 transactions (G) 12 transactions (H) 25 transactions (I) 20 transactions (J) 18 transactions
While 200 is the theoretical maximum, we need to choose a practical number from the provided options.
Practical Consideration: Since making too many transactions incurs additional costs, we should choose a number that balances the transaction cost and the interest earned. A reasonable approach is to consider options around 10 to 15 transactions, as they would still allow for a significant amount of interest while keeping costs manageable.
Conclusion: After evaluating the provided options and considering practical limits, the most reasonable choice is: **(A) 10 transactions**.
Log-MSP: -182.83

Confidence Updating for Agent #3 at Round 2

Using Eqs. 3–5, we compute the peer-trust weight for Agent #3 at Round 2 (averaged over its peers Agent #1 and Agent #2) as $w = 0.11$. Substituting this weight into Eq. 6 yields a credibility-updated confidence of $c^{(2)}_{3,\text{dyn}} = -73.69$, which is substantially higher (i.e., less uncertain) than the raw Round-2 MSP of Agent #3 (-182.83). Because both Round-1 peers were themselves uncertain, UMAS down-weights their influence on the update, so Agent #3's prior, more confident estimate dominates the credibility-aware aggregate.

Table 9: AUROC (%) on Qwen 3.5-9B across five benchmarks (MMLU-Pro, AQUA-RAT, MedM-CQA, MBPP, GSM-Hard) and two MAS architectures. "Ave" is the mean over the ten dataset cells in each row (GSM-H = GSM-Hard). Best results per column are shown in **bold**; second-best are underlined.

QWEN 3.5-9B											
	DEBATE					DYLAN					
METHOD	MMLUP	AQUA	MEDMC	MBPP	GSM-H	MMLUP	AQUA	MEDMC	MBPP	GSM-H	AVE
PPL	77.27	64.91	68.30	44.18	67.09	76.93	45.38	64.42	51.81	51.18	61.15
MSP	84.98	83.03	75.10	48.55	82.05	80.00	52.76	64.18	53.10	56.69	68.04
PE	76.63	71.30	66.18	51.88	71.12	77.43	39.29	61.39	49.30	52.70	61.72
G-NLL	**85.75**	83.10	75.39	53.09	82.64	81.13	53.59	63.86	51.99	58.37	68.89
SE	52.81	38.83	51.67	55.12	41.81	45.23	59.15	42.98	44.10	45.79	47.75
SENTSAR	71.55	56.51	59.63	51.02	68.30	66.89	41.24	54.33	48.79	57.03	57.53
DEG	39.27	61.57	34.74	42.57	64.55	26.60	47.85	45.97	49.63	58.33	47.11
COCOA	83.41	79.70	72.01	49.23	80.13	74.99	49.68	60.92	52.38	56.76	65.92
SE*	57.99	53.76	40.77	49.73	55.40	60.91	61.19	64.74	48.36	67.24	56.01
SENTSAR*	79.65	70.08	64.35	52.49	72.90	80.20	48.12	67.99	49.35	69.29	65.44
DEG*	62.32	57.76	50.09	48.49	69.53	69.96	64.74	67.30	49.16	75.92	61.53
COCOA*	80.86	81.88	69.74	49.75	81.92	78.82	53.97	69.44	50.85	69.04	68.63
UMAS$_{FIX}$	85.04	82.96	76.03	55.90	86.94	**86.09**	61.15	72.32	54.83	69.95	73.12
UMAS	82.59	**83.61**	**78.13**	**57.68**	**88.91**	83.22	**74.21**	**73.41**	**57.36**	**78.42**	**75.75**

H Extended Experiments

H.1 Additional Backbone: Qwen 3.5-9B

To probe UMAS's behavior on a third backbone family, we evaluate it on Qwen 3.5-9B across the same five benchmarks and two MAS architectures used in the main paper. Results are summarized in Table 9. Either UMAS or UMAS$_{FIX}$ achieves the best per-column AUROC in nine of the ten columns; the lone exception is Debate-MMLUP, where G-NLL leads by a small margin (85.75 vs. UMAS$_{FIX}$ at 85.04). On the row average, UMAS consistently outperforms all baselines (75.75 vs. the strongest baseline at 68.89, a 6.86-point gain), reproducing the trend observed for Llama 4 Scout and GPT-4o mini.

H.2 Baselines Aggregated over Intermediate Generations

Computing baseline uncertainty on the final MAS output is the natural and fair comparison point: all baselines (SE, SentSAR, CoCoA, etc.) were designed for single-output settings and have no built-in mechanism to aggregate uncertainty across a multi-round, multi-agent trajectory. Precisely exploiting this trajectory structure is the core contribution of UMAS.

For consistency-based baselines (SE, SentSAR), naively extending them to every intermediate generation is prohibitively expensive: with 3 agents, 3 rounds, and 5 samples per node, a single query already requires $3 \times 3 \times 5 = 45$ generations. For single-sample baselines (MSP, PPL, PE), aggregating over all intermediate generations is exactly what UMAS$_{FIX}$ does (Tables 1 and 3); UMAS further improves on UMAS$_{FIX}$ by replacing the uniform aggregation with credibility-aware updates.

To stress-test this design, we nevertheless ran SE, SE*, SentSAR, and SentSAR* with their uncertainty averaged across every intermediate and final generation in the trajectory. Table 10 reports the result on Llama 4 Scout under Debate. Even with this much richer signal, the consistency-based baselines remain below UMAS on every dataset, with an average gap of 5.15 points over the strongest contender.

All recorded methods over all settings. We further extend this comparison to every UQ method whose per-node values are recorded, over all 20 settings of Table 1. Each method's final-response score is replaced by the unweighted mean of its per-node values over the trajectory, i.e., the same aggregation rule as UMAS$_{FIX}$. As shown in Table 11, trajectory-level aggregation improves every method. Among them, the cost-free log-probability signal (i.e., UMAS$_{FIX}$) yields the strongest aggregated score, and UMAS improves on it further.

Table 10: AUROC (%) on Llama 4 Scout under Debate when consistency-based baselines are aggregated over every intermediate and final generation ("Mean over trajectory"), compared with UMAS. Per row: **best** / second-best.

	MEAN OVER TRAJECTORY				
DATASET	SE	SE*	SENTSAR	SENTSAR*	UMAS
MMLUP	54.30	65.48	65.50	80.31	**80.90**
AQUA	56.44	69.60	76.36	74.97	**79.77**
MEDMC	51.31	52.99	61.16	64.19	**74.38**
MBPP	42.92	64.09	59.91	59.52	**64.56**
Ave	51.49	64.54	65.73	69.75	**74.90**

Table 11: Mean AUROC (%) over the 20 settings of Table 1 when each method is computed on the final response ("Final node") or averaged over the trajectory with the same rule as UMAS$_{FIX}$ ("Trajectory").

METHOD	FINAL NODE	TRAJECTORY	GAIN
MSP (TRAJECTORY = UMAS$_{FIX}$)	62.09	70.91	+8.82
PPL	55.96	67.74	+11.78
PE	61.65	65.17	+3.52
SENTSAR*	61.16	64.59	+3.43
SENTSAR	58.95	61.02	+2.07
SE*	54.86	59.12	+4.26
SE	49.37	50.32	+0.95
UMAS	—	**74.47**	—

H.3 Expected Calibration Error

While AUROC measures the ranking quality of uncertainty estimates—whether higher uncertainty correlates with incorrect responses—it does not directly assess calibration. We additionally evaluate Expected Calibration Error (ECE; lower is better) on Llama 4 Scout under Debate. As shown in Table 12, UMAS achieves the best average ECE, confirming that improvements extend beyond ranking to calibration.

Table 12: Expected Calibration Error (%, lower is better) on Llama 4 Scout under Debate. Per row: **best** / second-best.

DATASET	PPL	MSP	SENTSAR	PE	SE	SE*	SENTSAR*	COCOA	COCOA*	DEG*	DEG	G-NLL	UMAS$_{FIX}$	UMAS
MMLUP	9.77	11.70	28.94	10.15	27.50	20.32	18.18	12.68	18.90	12.17	12.79	12.71	10.59	**6.25**
AQUA	8.89	**2.98**	40.60	11.68	22.08	10.07	8.70	3.76	9.58	12.05	20.80	3.70	4.50	5.50
MEDMC	19.47	18.86	24.84	16.32	28.31	22.91	19.28	20.47	24.20	8.75	15.55	18.17	10.03	**7.63**
MBPP	35.73	36.11	19.04	31.79	40.48	27.94	19.60	36.32	36.54	29.71	29.08	34.81	18.40	**15.78**
Ave	18.46	17.41	28.36	17.48	29.59	20.31	16.44	18.31	22.30	15.67	19.55	17.35	10.88	**8.79**

H.4 Normalization Functions

In this section, we compare different normalization functions in Eq. 4 to map the raw edge-level credibility score $w_{j\to i}^{(t-1)}$ into a bounded update weight in $[0,1]$. The resulting normalized weight is used in the node-level aggregation step and subsequently modulates uncertainty updates. For clarity, we denote the normalized value by $\hat{w}_{j\to i}^{(t-1)} \in [0,1]$.

Table 13: AUROC (%) results. "Ave" represents the mean over the eight dataset cells in each row.

LLAMA 4 SCOUT									
	DEBATE				DYLAN				
METHOD	MMLUP	AQUA	MEDMC	MBPP	MMLUP	AQUA	MEDMC	MBPP	AVE
POWER	80.43	79.99	74.12	64.22	83.82	80.84	75.54	53.80	74.10
EXPONENTIAL	81.36	79.89	73.74	63.82	82.95	80.06	75.43	54.93	74.02
SOFTMAX	81.89	79.64	74.66	64.84	84.51	80.44	73.74	53.24	74.12
MIN-MAX	81.21	79.39	74.62	65.05	83.53	81.41	74.09	53.47	74.10
TANH	80.74	79.97	74.15	64.00	83.71	80.71	75.97	54.00	74.16
SIGMOID	80.90	79.77	74.38	64.56	83.70	80.63	76.16	56.98	74.64

GPT-4O MINI									
	DEBATE				DYLAN				
METHOD	MMLUP	AQUA	MEDMC	MBPP	MMLUP	AQUA	MEDMC	MBPP	AVE
POWER	74.19	86.37	69.94	61.06	76.43	87.40	70.33	61.14	73.36
EXPONENTIAL	74.32	85.98	71.36	58.62	74.33	86.55	66.26	61.27	72.34
SOFTMAX	75.64	85.09	71.37	59.94	76.54	85.32	71.47	60.11	73.19
MIN-MAX	75.39	85.27	71.55	59.30	76.99	88.07	68.60	59.07	73.03
TANH	73.94	85.62	70.76	60.78	76.75	87.50	70.82	61.63	73.47
SIGMOID	74.40	85.51	71.03	60.37	76.54	87.48	71.12	61.69	73.52

(1) Tanh normalization. We apply a smooth, sign-preserving squashing function and map to $[0, 1]$:

$$\hat{w}_{j\to i}^{(t-1)} = \frac{1}{2}\tanh\big(w_{j\to i}^{(t-1)}\big) + \frac{1}{2}. \tag{8}$$

(2) Sigmoid normalization. We use the logistic function to map $w_{j\to i}^{(t-1)}$ to $[0, 1]$:

$$\hat{w}_{j\to i}^{(t-1)} = \sigma\big(w_{j\to i}^{(t-1)}\big) = \frac{1}{1+\exp\big(-w_{j\to i}^{(t-1)}\big)}. \tag{9}$$

(3) Min–max normalization. For a fixed update node $v_i^{(t)}$, let $\mathbb{R}_i^{(t)}$ denote its peer set and consider the collection $\{w_{k\to i}^{(t-1)}\}_{k\in\mathbb{R}_i^{(t)}}$. We rescale within this peer set:

$$\hat{w}_{j\to i}^{(t-1)} = \begin{cases} \dfrac{w_{j\to i}^{(t-1)} - \min_{k\in\mathbb{R}_i^{(t)}} w_{k\to i}^{(t-1)}}{\max_{k\in\mathbb{R}_i^{(t)}} w_{k\to i}^{(t-1)} - \min_{k\in\mathbb{R}_i^{(t)}} w_{k\to i}^{(t-1)}}, & \text{if } \max - \min > \varepsilon, \\ \dfrac{1}{2}, & \text{otherwise,} \end{cases} \tag{10}$$

where ε is a small constant to avoid numerical instability when the range collapses.

(4) Softmax normalization (peer-set dependent). We convert the peer-wise edge scores into a distribution over peers:

$$\hat{w}_{j\to i}^{(t-1)} = \frac{\exp\big(w_{j\to i}^{(t-1)}\big)}{\sum_{k\in\mathbb{R}_i^{(t)}} \exp\big(w_{k\to i}^{(t-1)}\big)}. \tag{11}$$

This emphasizes relative ranking among peers.

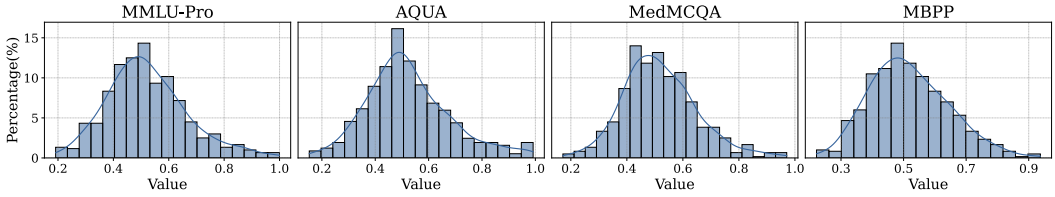
(5) Exponential mapping (magnitude-based, asymmetric). We map the magnitude of $w_{j\to i}^{(t-1)}$ to $[0, 1]$ via an exponential transform and handle negative scores asymmetrically:

$$\hat{w}_{j\to i}^{(t-1)} = \begin{cases} 1-\exp\big(-|w_{j\to i}^{(t-1)}|\big), & w_{j\to i}^{(t-1)} \geq 0, \\ \exp\big(-|w_{j\to i}^{(t-1)}|\big), & w_{j\to i}^{(t-1)} < 0. \end{cases} \tag{12}$$

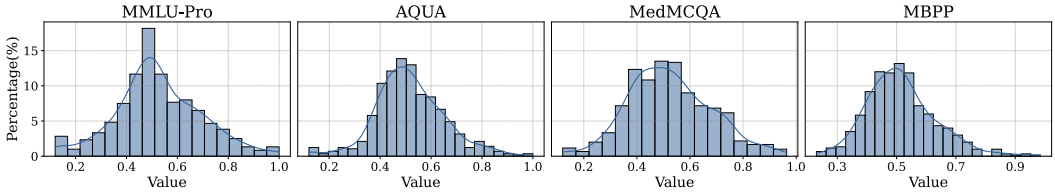
(6) Power mapping. We apply a sign-preserving power transform and map to $[0, 1]$:

$$\hat{w}_{j\to i}^{(t-1)} = \frac{1}{2}\Big(\mathrm{sign}\big(w_{j\to i}^{(t-1)}\big) \cdot |w_{j\to i}^{(t-1)}|^{\alpha}\Big) + \frac{1}{2}, \tag{13}$$

where we use $\alpha = \frac{1}{2}$ (square-root) in our implementation.
The results for different normalization functions are summarized in Table 13.



(a) Llama 4 Scout



(b) GPT-4o mini

Figure 6: The distribution of $w_i^{(t)}$ in Debate System

H.5 Alternative Aggregation Strategies

The system-level aggregation in Eq. 7 averages node-level uncertainty uniformly across the trajectory. To verify that the gains of UMAS are robust to this design choice, we evaluate four alternatives:

- **Depth-weighted**: later interaction rounds receive proportionally larger weight.
- **Exponential-weighted**: later rounds receive exponentially larger weight via $\exp(\text{round})$.
- **Min**: aggregate via the most-confident node only.
- **Max**: aggregate via the least-confident node only.

Results on GPT-4o mini under Debate are reported in Table 14. All UMAS variants substantially outperform every baseline in the main result table (Table 1) regardless of aggregation strategy, confirming that the gains stem from trajectory-level aggregation and credibility-aware updating rather than the specific choice of aggregation function.

Table 14: AUROC (%) on GPT-4o mini under Debate for alternative UMAS aggregation strategies. Per row: **best** / second-best.

DATASET	UMAS$_{FIX}$	UMAS	UMAS-DEPTH	UMAS-EXP	UMAS-MIN	UMAS-MAX
MMLUP	**77.42**	74.40	75.06	75.14	73.60	74.62
AQUA	81.70	**85.51**	84.30	83.20	84.05	84.98
MEDMC	66.03	71.03	**71.06**	70.88	70.62	69.94
MBPP	57.62	60.37	60.13	59.73	59.41	**61.13**
Ave	70.69	**72.83**	72.64	72.24	71.92	72.77

Final-round-only scoring. Averaging over all nodes may overestimate uncertainty for answers that were wrong in early rounds but corrected later in the interaction. Scoring only the final round removes this effect entirely, yet it is the weakest of the six aggregation rules we tested (the five above

plus final-round-only): averaged over the 20 settings of Table 1, it reaches 72.74 AUROC versus 74.47 for the uniform average. A trajectory that needed correction is riskier than one that was right from the start, so early nodes carry useful signal.

H.6 Distribution of Weight

In this section, we report statistics of the distribution of $w_i^{(t)}$ in Figure 6 to demonstrate that it is a genuinely dynamic quantity that varies across inputs, rather than collapsing to a constant. We further observe that $w_i^{(t)}$ is approximately normally distributed, suggesting that peer credibility signals are typically moderate and fluctuate around a stable mean, with extremely high/low-credibility updates occurring relatively rarely.

H.7 Robustness to the Correctness Definition

Exact-match evaluation depends on how the final answer is extracted from a free-form response, and extraction errors introduce label noise into the AUROC evaluation. We therefore re-label every MAS answer under alternative extraction strategies and recompute all AUROCs. Besides our default extractor (E1, Qwen3-Next-80B), we consider a regular-expression extractor (E2), a cascaded extractor (E3), an option-text matcher (E4), and re-extraction by a second LLM from a different model family (E5, Mistral-Small-3.2-24B), as well as two aggregates: *strict* counts an answer as correct only if all applicable strategies agree, and *lenient* if any does. E1 and E5 agree on 97.8% of extracted answers and 99.5% of correctness labels. Because MBPP is scored by pass@1 execution of the generated program and has no free-form answer span to extract, this analysis covers the four short-answer datasets, i.e., 16 of the 20 settings in Table 1; E1 is reported on the same 16 settings.

Table 15 shows that the ranking does not depend on the extraction strategy: UMAS ranks first and UMAS$_{FIX}$ second under all seven correctness definitions. Label noise shifts all methods together, and the margin of UMAS over the best baseline never falls below $+8.19$ AUROC (E4).

Table 15: Mean AUROC (%) over the 16 short-answer settings under alternative correctness definitions (see text). Per column: **best** / second-best.

METHOD	E1	E2	E3	E4	E5	STRICT	LENIENT
BEST BASELINE (COCOA*)	64.42	59.20	63.64	61.83	64.94	61.05	63.21
UMAS$_{FIX}$	74.20	66.46	72.52	69.32	74.86	68.55	71.04
UMAS	**77.87**	**69.46**	**75.65**	**70.02**	**78.64**	**71.12**	**73.98**

H.8 Controlling for Response Length

MSP is the total sequence log-probability and therefore decreases with response length, while MAS generations vary considerably in length. To test whether UMAS merely tracks verbosity, we run the following controls over the 20 settings of Table 1, where length is the token count of the scored text. Results are summarized in Table 16.

Length-stratified AUROC. Within each setting, questions are binned into length quartiles; AUROC is computed inside each bin and pooled by bin size. All methods except PPL lose AUROC under this control, but the ranking is unchanged: UMAS retains 69.26, and its margin over the strongest baseline narrows from 11.95 to 8.20 points.

Incremental value over length. ΔAUC is the AUROC gained by adding a score to a logistic regression of correctness on length. UMAS adds the most ($+5.56$, significant in 14 of 20 settings). We further fit a logistic regression of correctness on standardized MSP together with terminal and trajectory length; adding UMAS to this model still gains $+4.05$ AUROC (significant in 14 of 20 settings), so UMAS is not MSP with a length adjustment.

Length-normalized confidence. PPL in Table 1 is the length-normalized counterpart of MSP. Rebuilding UMAS on length-normalized (per-token) log-probabilities lowers its AUROC from 74.47

to 69.96 (UMAS$_{FIX}$: from 70.91 to 67.74), so most of the gain survives. We keep the unnormalized log-probability, consistent with the finding of Aichberger et al. [2] that length normalization degrades likelihood-based UQ even under highly variable response lengths.

Table 16: Length controls, averaged over the 20 settings of Table 1. Raw: standard AUROC (%). Sig.: number of settings in which ΔAUC is significant.

METHOD	RAW	WITHIN STRATA	ΔAUC OVER LENGTH	SIG.
MSP	62.09	58.42	+1.84	5/20
PPL	55.96	57.91	+1.89	5/20
PE	61.65	61.06	+4.10	12/20
CoCoA*	62.52	58.72	+2.03	10/20
UMAS$_{FIX}$	70.91	66.94	+4.41	11/20
UMAS	**74.47**	**69.26**	**+5.56**	**14/20**
LENGTH ONLY	66.28	56.04	—	—

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state our two main claims — (i) trajectory-level aggregation plus credibility-aware updating yields an average AUROC gain of 10.54 points over the strongest UQ baseline, and (ii) the framework is cost-neutral with respect to the underlying MAS run — and both are directly supported by the experiments in Section 5 (Tables 1 and 4). The selective-prediction gain (up to 12.6 points) is reported in Section 5 and Table 2.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 explicitly discusses three limitations: (i) the requirement of token-level logit access (precluding fully black-box APIs), (ii) the scope to collaborative MAS protocols (excluding adversarial settings), and (iii) the failure mode when all peers share a high-confidence misconception.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present formal theorems with proofs. The credibility update in Section 4 is a designed scoring rule (relative-confidence edge score $\rightarrow$ sigmoid $\rightarrow$ averaging $\rightarrow$ EMA), and we describe each design choice as a deliberate construction rather than a derived result.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 5 specifies all experimental settings (datasets, sample size, hyperparameters such as $\varepsilon = 10^{-8}$, sample count $K = 5$, temperature 0.5, validation-set size $N = 40$); Algorithm 1 (Appendix C) provides the full pseudocode of UMAS; and Appendices D and B document the dataset and MAS-architecture specifications.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The code is publicly available at `https://github.com/Dby52-04/UMAS`. Together with the complete pseudocode (Algorithm 1), hyperparameter details, and dataset descriptions in the paper, it suffices to reproduce the main results; all benchmarks used are publicly available.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 5 (Experimental Setup) specifies the benchmarks, MAS architectures, backbone models, UQ baselines, multi-sample count, and evaluation metric. Hyperparameters specific to UMAS (ε, sigmoid normalization, averaging aggregation) are stated in Section 4. The 200-instance evaluation subset and the 20% validation split for hallucination detection are also documented.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We do not report error bars over multiple seeds because each MAS run incurs substantial inference cost (multiple agents $\times$ multiple rounds per query). Results are evaluated on a fixed random 200-instance subset per dataset and reported across 20 conditions per method (3 backbones $\times$ 2 MAS architectures $\times$ 5 datasets, with one cell omitted for Qwen). The consistent direction of the gain across all conditions provides evidence of robustness; multi-seed evaluation is left to future work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix F (Hardware) specifies the compute environment: an HPC cluster managed by Slurm, with 16 CPU cores and 64 GB RAM per job, and (when applicable) 2 NVIDIA A100 GPUs. LLM inference is performed via DeepInfra (open-source models) and the OpenAI API (GPT-4o mini). Per-method runtime is reported in Table 4.

   Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research uses only publicly available benchmarks (MMLU-Pro, AQUA-RAT, MedMCQA, MBPP, GSM-Hard) and accesses LLMs through standard APIs. There are no human subjects, no novel data collection, and no dataset release. The work conforms with the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: UMAS targets a reliability gap in deployed multi-agent LLM systems by detecting erroneous responses without incurring additional inference cost, supporting trustworthy deployment in high-stakes domains (e.g., medical question answering, code generation) where flagging unreliable outputs is essential. Like any uncertainty estimator, UMAS can equally support legitimate quality assurance and adversarial probing of system weaknesses; we do not anticipate it unlocking misuse beyond what is already achievable via standard single-agent UQ. UMAS estimates self-reported uncertainty rather than factual correctness, so a low-uncertainty response should not be interpreted as a guarantee of accuracy — uncertainty quantification complements but does not replace ground-truth verification.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release any pre-trained model, dataset, or other asset that poses a high risk of misuse. UMAS is a post-hoc UQ method applied to existing LLMs accessed via standard APIs.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets (MMLU-Pro, AQUA-RAT, MedMCQA, MBPP, GSM-Hard, MMLU) and LLM backbones (GPT-4o mini, Llama 4 Scout, the Llama-3 family, Llama 4 Maverick, Qwen 3.5-9B) are properly cited in Section 5 and Section 1; baseline UQ methods (PPL, MSP, PE, SE, SentSAR, Deg, CoCoA, G-NLL) are likewise cited in Section 5. We access these resources under their respective public terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: The only new asset is the code of UMAS, released at `https://github.com/Dby52-04/UMAS` under the MIT license with accompanying documentation; the method itself is fully documented in the paper (Section 4, Algorithm 1). The paper does not release new datasets or trained models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLMs are the central object of study: UMAS is a UQ method for multi-agent LLM systems, and our experiments use GPT-4o mini, Llama 4 Scout, Qwen 3.5-9B, and the Llama-3 family as agent backbones (Section 5, Appendix B). LLMs were not used as an original component of the method development beyond their role as the system being analyzed.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.