

# DEBUNC: IMPROVING LARGE LANGUAGE MODEL AGENT COMMUNICATION VIA UNCERTAINTY METRICS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

To enhance Large Language Model (LLM) capabilities, multi-agent debates have been introduced, where multiple LLMs discuss solutions to a problem over several rounds of debate. However, LLMs often produce incorrect responses that appear confident, which can mislead other agents. This is partly because agents do not express their confidence levels during standard debates. To address this, we introduce DebUnc, a multi-agent debate framework that uses uncertainty metrics to assess agent confidence levels. We adapted the LLM attention mechanism to adjust token weights based on confidence levels and also explored using textual prompts to convey confidence. Our evaluations across various benchmarks show that attention-based methods are particularly effective, and that as uncertainty metrics improve, performance will continue to increase.

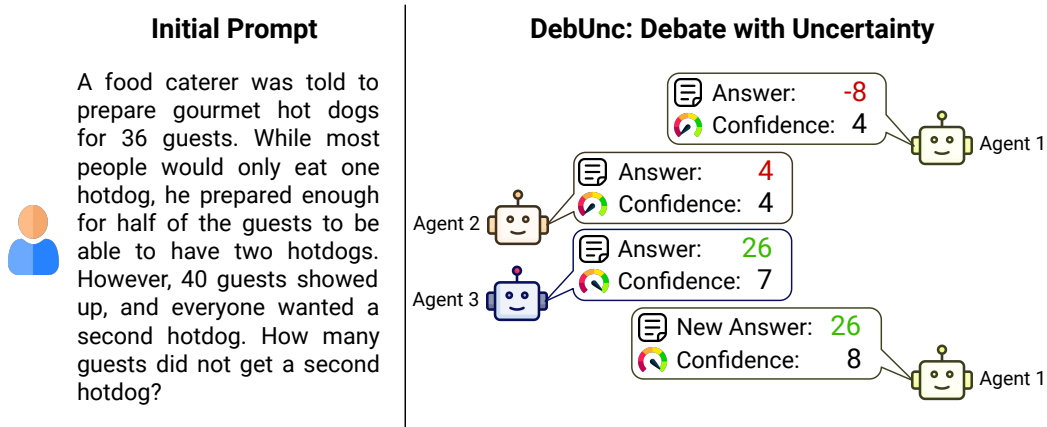


Figure 1: Illustration of a three-agent mathematical debate. Agent 1 initially provides an incorrect response, but corrects itself after seeing the responses and confidence levels from the other agents. Each agent uses a large language model (LLM) to generate text responses and assesses its confidence using an uncertainty metric. The responses and confidence information is shared among the agents, enabling them to decide whom to trust when responses differ. Correct answers are marked in green, while incorrect ones are shown in red.

## 1 INTRODUCTION

Large language models (LLMs) have shown impressive performance in various fields, including law, academia, and coding (OpenAI, 2024). To handle more complex tasks, LLM-powered agents have been developed. These agents observe their environment and take actions, such as communicating with other agents, using tools, or performing reasoning (Wu et al., 2023). The potential of LLM agents increases when multiple agents collaborate. One form of this is multi-agent debate, where agents propose and debate solutions to a problem with other agents. Multi-agent debates have been shown to improve the reasoning and accuracy of LLMs (Du et al., 2023).

054 The factual accuracy of LLMs is crucial for their utility in many real-world applications (Rawte et al.,  
055 2023). For example, an LLM tutor providing incorrect information could harm a student’s learning,  
056 and an LLM customer service agent giving incorrect advice could mislead a customer. In fields such  
057 as healthcare, journalism, or financial services, the consequences of LLM errors can be even more  
058 severe, leading to financial losses or health risks. To address this, multi-agent debate systems have  
059 been proposed. These systems enable multiple agents to generate diverse responses to a problem,  
060 discuss and critique each other’s answers, and ultimately converge on a final solution (Liang et al.,  
061 2023). Ideally, if some agents provide incorrect answers initially, the debate process helps them  
062 recognize and correct their errors.

063 In practice, while agents often confidently agree on the same final answer, that answer is not always  
064 correct. This can be attributed to flawed communication between agents (Du et al., 2023). LLM  
065 agents typically communicate through text and respond with a highly confident tone, regardless of  
066 of the accuracy of their answers. This creates a significant challenge in multi-agent systems, as a  
067 confidently incorrect response from one agent can mislead others, causing all agents to converge  
068 on an incorrect conclusion. By contrast, people often use qualifiers like "I am sure that..." or "I  
069 am not sure, but..." which, though imperfect, provide cues that help others gauge the reliability  
070 of the information. Since the tone of an LLM’s response is not a reliable indicator of its accuracy,  
071 researchers have developed uncertainty metrics to provide a more objective measure of the model’s  
072 confidence.

073 Building on these insights, we present DebUnc, a novel multi-agent debate framework that integrates  
074 multi-agent **Deb**ates with model **Unc**ertainty metrics. After each round of debate, we measure each  
075 agent’s uncertainty with an uncertainty metric. In the following round, both the agents’ responses  
076 and uncertainties are shared with the other agents. We explore two methods for conveying agent  
077 uncertainty: (1) incorporating the uncertainty directly into the textual prompt alongside the agent  
078 responses, as shown in Figure 2, and (2) adjusting the LLM’s attention towards agents’ responses  
079 based on their uncertainty, as depicted in Figure 3. We extensively evaluate DebUnc across multiple  
080 LLMs, benchmarks, and uncertainty metrics, analyzing the results of each uncertainty metric and  
081 method of uncertainty communication.

082 Our key contributions are outlined as follows:

- 083 • We introduce DebUnc, a framework that quantifies and communicates LLM agent uncertainty in  
084 multi-agent debates.
- 085 • We adapt the LLM attention mechanism to adjust token weights based on confidence levels and  
086 also explore the use of textual prompts to communicate confidence.
- 087 • We evaluate DebUnc across multiple LLMs, benchmarks, and uncertainty metrics, and find that the  
088 attention-scaling methods consistently outperforms unmodified debates.
- 089 • We offer insights into how performance will be impacted as uncertainty metrics improve.

## 091 2 RELATED WORK

092 LLMs are known for their overconfidence and their tendency to provide responses to any user query,  
093 regardless of their certainty. This often leads to the generation of factual inaccuracies, known as  
094 hallucinations, where the information provided by the model is incorrect or unsupported by the data  
095 on which it was trained (Liang et al., 2024; Yadkori et al., 2024; Duan et al., 2024; Yao et al., 2023;  
096 Aichberger et al., 2024). Ensuring factual accuracy is crucial for building trust in LLM-based systems  
097 and expanding their use in real-world applications. As a result, there has been a surge in research  
098 focused understanding the mechanisms behind hallucinations and developing strategies to mitigate  
099 them (Ji et al., 2023; McDonald et al., 2024; Liu et al., 2023).

### 102 2.1 UNCERTAINTY IN LLMs

103 Some current research efforts to mitigate hallucinations focus on measuring the model’s uncertainty  
104 and enhancing their self-awareness (Kadavath et al., 2022; Amayuelas et al., 2023; Yin et al., 2023).  
105 If we could accurately measure a model’s confidence, users would have clearer guidance on when to  
106 trust its output (Lin et al., 2022a; Xu et al., 2024), and language agents could better determine when  
107 to access external resources (Han et al., 2024).

108 However, accurately measuring model uncertainty remains an open problem, and ongoing research is  
109 exploring new uncertainty metrics. We identify three primary approaches to this problem:

110 **Token Probability-Based Uncertainty Metrics** evaluate uncertainty based on the probabilities  
111 assigned to each token generated by the model. High token probabilities (close to 1) indicate strong  
112 model confidence, whereas lower token probabilities suggest a spread of probability across several  
113 tokens, signifying uncertainty about the correct choice. Prominent methods in this category include  
114 Mean Token Entropy, Perplexity (Fomicheva et al., 2020), SAR (Duan et al., 2023), RDE (Vazhentsev  
115 et al., 2023), and Claim-Conditioned Probability (Fadeeva et al., 2024).

116 **LLM-Generated Uncertainty Metrics** involve the model explicitly expressing its uncertainty in  
117 its response. Lin et al. (2022a) explored this by fine-tuning GPT-3 (Brown et al., 2020) to provide  
118 both an answer and a corresponding confidence level. Alternatively, the model can be prompted to  
119 express its uncertainty without explicit training. Tian et al. (2023) found that this approach can  
120 outperform token-probability-based methods when applied to LLMs fine-tuned with reinforcement  
121 learning from human feedback (Christiano et al., 2017). However, Xiong et al. (2023) report lower  
122 performance compared to token probability-based methods on GPT-3.

123 **Sampling-Based Uncertainty Metrics** assess uncertainty by generating multiple outputs through  
124 sampling, analyzing the distribution of meanings across the outputs. A consistent meaning across  
125 multiple samples suggests high confidence, while variations indicate lower confidence. This approach  
126 can identify different sequences that convey the same meaning, which token-probability-based metrics  
127 do not account for. However, the need for multiple generations makes these methods more resource-  
128 intensive than the others described. Examples include Semantic Entropy (Kuhn et al., 2023), LUQ  
129 (Zhang et al., 2024), and other metrics that evaluate meaning diversity (Lin et al., 2023).

## 131 2.2 MULTI-AGENT DEBATE

132  
133 With the increased accessibility of LLMs and improvements in their inference times, LLMs are being  
134 integrated into more complex systems as autonomous agents (Wu et al., 2023; Li et al., 2023; Hong  
135 et al., 2023). A critical component of these agent-based systems is the collaboration mechanism,  
136 where models engage in debate with one another. These mechanisms are currently being studied  
137 (Zhang et al., 2023) and have been shown to foster more divergent thinking (Liang et al., 2023),  
138 enhance reasoning and factual accuracy (Du et al., 2023), and lead to more reliable evaluations (Chan  
139 et al., 2023). Through discussions, the LLMs can refine their outputs, ultimately achieving higher  
140 levels of agreement and producing more factually accurate text (Sun et al., 2024; Feng et al., 2024).

141 Pham et al. (2023) recognized that text is not be the most effective communication mechanism for  
142 LLM agents, since information is lost during the token sampling process, and demonstrated how  
143 LLMs can communicate through embeddings. ReConcile Chen et al. (2023) explored the integration  
144 of agent confidence in multi-agent debates, relying on LLM agents to self-report their confidence,  
145 which was communicated to other agents through prompts. Building on these ideas, we employ  
146 uncertainty metrics to estimate agent confidence and explore both prompting and attention scaling to  
147 convey this confidence. We found these enhancements to significantly improve multi-agent debate  
148 performance.

## 149 3 METHOD

150  
151  
152 In human debates, it is often possible to gauge someone’s expertise on a subject by observing the  
153 fluency of their responses, their body language, and other cues. This helps in identifying whose  
154 arguments to consider more seriously when there are conflicting opinions. On the other hand, in  
155 multi-agent LLM debates, agents frequently generate inaccurate responses that sound confident,  
156 which can mislead other agents and result in a consensus on an incorrect response (Du et al., 2023).  
157 Our goal is to advise agents on which other agents’ opinions to prioritize based on their confidence  
158 levels.

159 Our modified debate pipeline, depicted in Figure 2, operates as follows: in each round of debate,  
160 every agent generates a response, and its uncertainty is estimated. In the next round, the responses  
161 and uncertainties from each agent are shared with every other agent. We test three uncertainty metrics  
and three approaches to communicate agent uncertainty.

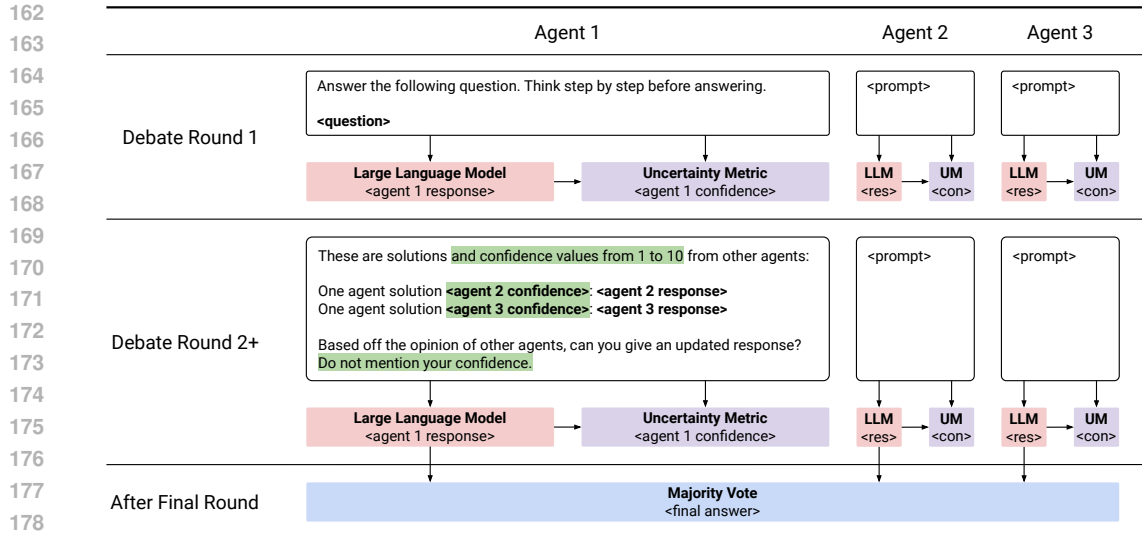


Figure 2: Illustration of the modified multi-agent debate involving three agents. In the first round, each agent independently generates a response to the question, which is evaluated for confidence using an uncertainty metric. The prompt for following rounds includes the responses from other agents in the previous round. Sections of the prompt highlighted in green are used only with the Confidence in Prompt method. Each agent retains access to its complete chat history throughout the debate. After the final round, a majority vote determines the final answer.

### 3.1 UNCERTAINTY METRICS

Uncertainty metrics assess an LLM’s confidence in its responses: high uncertainty implies low confidence and potential unreliability, while low uncertainty suggests greater reliability. These metrics generally fall into three categories: Token Probability-Based, LLM-Generated, and Sampling-Based methods. For more details, see Uncertainty in LLMs

In our experiments, we focus on token probability-based metrics due to their efficiency and flexibility, requiring only a single generation and functioning independently of the model’s ability to express uncertainty. We specifically chose Mean Token Entropy (Fomicheva et al., 2020) for its simplicity and TokenSAR (Duan et al., 2023), which accounts for the fact that some tokens contribute more to a sequence’s meaning than others. We utilize the implementations from LM-Polygraph, a framework with implementations for many uncertainty metrics (Fadeeva et al., 2023). Lastly, to evaluate the potential of future uncertainty metrics, we include a third "oracle" uncertainty metric in our analysis.

**Mean Token Entropy.** One of the simplest and most efficient uncertainty metrics to compute is Mean Token Entropy (Fomicheva et al., 2020). It is the average entropy across all tokens generated, with the entropy of a single token  $X$  defined as:

$$H(X) = - \sum_{x \in V} p(x) \log p(x)$$

Here,  $V$  denotes the model’s vocabulary. Entropy is maximized when  $p(x)$  is uniform over all tokens in the vocabulary, indicating maximum uncertainty. It is minimized when one token has a probability of 1 and all other tokens have a probability of 0, indicating complete certainty in the selected token.

**TokenSAR.** Duan et al. (2023) recognized that some tokens contribute more to a text’s meaning than others and proposed TokenSAR, an uncertainty metric that accounts for this. It is defined as the weighted average of the negative log probabilities for each generated token, where the weights are the relevance scores of the tokens:

$$\text{TokenSAR} = \sum_i^N -\log p(t_i)R(t_i)$$

Here,  $N$  is the number of tokens generated,  $t_i$  is the  $i$ -th token, and  $R(t_i)$  is the relevance of token  $t_i$ . To compute each token’s relevance, RoBERTa-large (Liu et al., 2019) must be run  $N$  times in total. This is more computationally expensive than calculating mean token entropy, but still far less costly than metrics requiring multiple generations.

**Oracle.** While the uncertainty metrics discussed provide valuable insights into agent uncertainty, they are not without limitations, and future advancements will likely lead to more accurate metrics. To assess the potential effectiveness of our methods with improved uncertainty metrics, we include an "Oracle" metric, which simulates an ideal uncertainty metric. This metric yields low uncertainty when the agent is correct and high uncertainty when the agent is incorrect, and is defined as follows:

$$\text{uncertainty} = \begin{cases} 0 & \text{if the response is correct} \\ \infty & \text{if the response is incorrect} \end{cases}$$

In practice, using 0 and  $\infty$  could cause issues with our uncertainty communication methods, so we detail exactly how this metric is applied in the following subsection. It is also important to note that this metric requires knowledge of the ground truth answer, making it impractical for real-world use. Instead, it serves to evaluate the effectiveness of our uncertainty communication methods independently of the performance of the uncertainty metrics themselves, and allows us to anticipate how improvements in uncertainty metrics could affect debate performance.

### 3.2 UNCERTAINTY COMMUNICATION

After computing the uncertainty of each agent, we explore multiple methods to incorporate these uncertainties into the following debate round.

**Confidence in Prompt.** One approach is to include the uncertainties directly in the text prompt for the next debate round, as shown in Figure 2. Mean Token Entropy and TokenSAR yield non-negative uncertainties. For Mean Token Entropy, the range of uncertainties depends on the model’s vocabulary size, while for TokenSAR, the maximum uncertainty is unbounded. Therefore, the exact uncertainty values are less informative than the relative differences in uncertainty between agents.

Rather than expressing their uncertainty as an unbounded non-negative number, humans often express their confidence on 1 to 10 scale, which is more interpretable. Since LLMs are trained on human data, they may exhibit the same preference. As a result, we convert the uncertainties into confidence values. Given a list of uncertainties  $u$  for  $n$  agents, where  $u_i$  is the uncertainty of agent  $i$ , we first invert them to obtain raw confidence values  $r_i = \frac{1}{u_i}$ . We then scale these values such that the average confidence  $s_i$  of all agents is 5:

$$s_i = \frac{r_i}{\sum_{j=1}^n r_j} \cdot (5n - 1) + \frac{1}{n}$$

Finally, we clamp the confidence levels to the range of 1 to 10 and round to the nearest integer.

When using the Oracle uncertainty metric, we set the confidence to 1 if the agent was incorrect and to 10 if the agent was correct.

**Attention Scaling.** As an alternative to including confidence levels in the prompt, we can modify the LLM’s token-generation process to account for each agent’s confidence. Many LLMs use Transformer decoder layers that generate an embedding for the last token and use this embedding to predict the next token (Radford et al., 2018). This embedding is determined by the attention mechanism, which creates "query," "key," and "value" vectors for each token.

The similarity between the "query" vector of the last token and the "key" vector of each token is used to compute a weight for every token. These weights are normalized with a softmax function to ensure they sum to 1, and are used to create the output embedding, which is the weighted sum of the value vectors of each token (Vaswani et al., 2017). The weight of each token determines its influence on the next token generated. By modifying these weights, we can adjust the model’s focus on each token in the input.

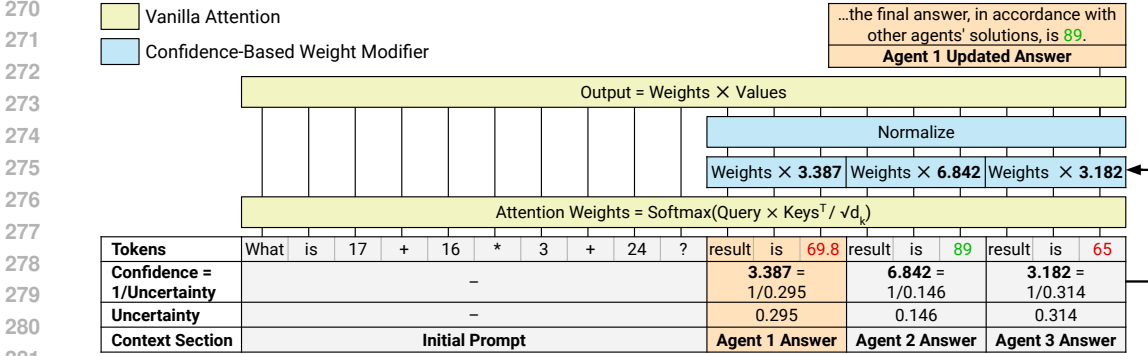


Figure 3: Illustration of the Attention-All method from the perspective of Agent 1. As the second debate round begins, the model’s context includes the initial prompt and each agent’s responses. Agent 2 provided a correct response with lower uncertainty than Agents 1 and 3, who responded incorrectly. Because Agent 2 had a lower uncertainty, the attention weights for tokens constituting Agent 2’s response will be increased, while those for tokens from Agent 1 and Agent 3’s responses will be decreased. This led Agent 1 to switch to the correct answer.

In multi-agent debates, this allows us to shift the model’s focus towards the responses from more confident agents. After each debate round, we have responses from each agent. In the next round, each agent’s prompt will include these responses. We also compute the uncertainty of each agent using an uncertainty metric.

In the next round, as the LLM generates its response, it computes the normalized attention weights for each preceding token. We multiply the weight of every token from agent  $j$  by the inverse of agent  $j$ ’s uncertainty when using Mean Token Entropy or TokenSAR. When using the Oracle metric, to avoid divide-by-zero errors, we set the multiplier to  $10^{-5}$  if the agent was incorrect and 1 if the agent was correct.

Formally, the attention weight for token  $i$   $a_i$  is:

$$a_i = \begin{cases} w_i \cdot m_j & \text{if } i \in t_j, \text{ for any agent } j \\ w_i & \text{otherwise} \end{cases}$$

Here,  $t_j$  is the set of token indices from agent  $j$ . We then normalize the scaled attention weights to ensure that the sum of all token weights equals 1, while leaving the weights of other tokens unchanged. The final weight  $f_i$  for every token  $i$  is calculated as follows:

$$f_i = \begin{cases} a_i \cdot \frac{\sum_{j=1}^n \sum_{k \in t_j} w_k}{\sum_{j=1}^n \sum_{k \in t_j} a_k} & \text{if } i \in t_j, \\ & \text{for any agent } j \\ a_i & \text{otherwise} \end{cases}$$

We only apply attention scaling to the responses from the previous round. For instance, in a three-round debate, attention would be rescaled for the responses from the first round during the second round, and for the responses from the second round during the third round. In the third round of debate, attention would not be rescaled to the first-round responses. Additionally, in order to prevent divide-by-zero errors during normalization, attention is not scaled when computing the embeddings for tokens within the prompt; it is only scaled when generating new tokens.

We explore two variants of attention scaling:

- **Attention-Others**, where agent  $i$  only rescales attention to other agents’ response tokens  $t_j \mid j \neq i$
- **Attention-All**, where agent  $i$  rescales attention to other agents and itself, illustrated in Figure 3

Metric	Method	MMLU-0	MMLU-5	GSM8k	Arith.	Truth.	Average
N/A	Standard	0.520	0.544	0.512	0.478	0.604	0.532 +0.0%
Entropy	Prompt	0.522	0.546	<b>0.536</b>	0.482	0.602	0.538 +1.1%
	Attn-Others	<b>0.540</b>	0.566	0.488	<b>0.518</b>	<b>0.608</b>	0.544 +2.3%
	Attn-All	0.526	<b>0.570</b>	<b>0.536</b>	<b>0.518</b>	0.604	<b>0.551</b> +3.6%
TokenSAR	Prompt	<b>0.538</b>	0.548	0.504	0.464	0.616	0.534 +0.4%
	Attn-Others	0.526	<b>0.560</b>	0.496	<b>0.500</b>	<b>0.626</b>	0.542 +1.9%
	Attn-All	0.532	0.552	<b>0.528</b>	<b>0.500</b>	0.610	<b>0.544</b> +2.3%
Oracle	Prompt	0.562	0.560	0.548	0.542	0.626	0.568 +6.8%
	Attn-Others	0.606	0.666	0.640	0.654	0.630	0.639 +20.1%
	Attn-All	<b>0.618</b>	<b>0.684</b>	<b>0.656</b>	<b>0.732</b>	<b>0.648</b>	<b>0.668</b> +25.6%

Table 1: Accuracy comparison across various benchmarks using different uncertainty estimators and methods with Mistral-7B. ‘MMLU-0’ denotes zero-shot performance on MMLU, while ‘MMLU-5’ represents 5-shot performance. The other benchmarks used zero-shot prompting. The ‘Average’ column shows the average performance for all tests and the % increase over the standard debate.

## 4 EXPERIMENT DESIGN

To evaluate these methods, an open-source LLM is required, as implementing the attention scaling requires modifications to the model source code. Additionally, the uncertainty metrics used rely on token probabilities, which may not be readily available from closed-source models. We ran all of the experiments on Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and re-evaluated the most insightful ones on Llama-3-8B-Instruct (AI@Meta, 2024) to verify that our methods work across different models. Tokens were sampled with a temperature of 1 to ensure variability in the responses. The methods were evaluated on multiple benchmarks:

1. **MMLU** (Hendrycks et al., 2021): A dataset of multiple-choice questions across various subjects.
2. **GSM8k** (Cobbe et al., 2021): A dataset of free-response grade school math problems.
3. **TruthfulQA** (Lin et al., 2022b): A multiple-choice dataset testing the model’s susceptibility to common misconceptions.
4. **Arithmetic**: An randomly generated set of arithmetic problems in the form  $a + b \cdot c + d$  for the Mistral model and  $a + b \cdot c + d - e \cdot f$  for the Llama model, where  $0 \leq a, b, c, d < 30$ .

For MMLU on Mistral-7B, we tested both zero-shot and 5-shot prompting, using examples from the original MMLU repository’s<sup>1</sup> dev set. For the remaining benchmarks, we used only zero-shot prompting. The exact prompts used are shown in Appendix C.

Instead of evaluating on the full datasets, which would be too expensive, we sampled 100 questions from each. We evaluated every combination of uncertainty metric and uncertainty incorporation method on these samples five times, and report the average accuracy across the five runs.

The debates involved three agents and three rounds, with all agents using the same LLM. In the first round, each agent answered the question individually. In the following rounds, each agent was given other agents’ responses, and used this information to refine their answers. A full example debate is shown in Appendix A. The final answer was determined by a majority vote after the third round. This workflow is illustrated in Figure 2.

## 5 RESULTS

In this section, we first analyze the effectiveness of each uncertainty incorporation method, and then analyze the effectiveness of the uncertainty metrics.

<sup>1</sup><https://github.com/hendrycks/test>

Metric	Method	MMLU-0	GSM8k	TruthfulQA	Arithmetic	Average
N/A	Standard	0.654	0.812	0.518	0.520	0.634 +0.0%
Entropy	Prompt	0.614	<b>0.840</b>	0.540	0.528	0.636 +0.4%
	Attention-Others	0.638	0.810	<b>0.562</b>	0.526	<b>0.639</b> +0.9%
	Attention-All	<b>0.658</b>	0.808	0.556	<b>0.530</b>	0.638 +0.4%
Oracle	Prompt	0.668	0.872	0.584	0.546	0.668 +5.5%
	Attention-Others	<b>0.784</b>	0.898	0.674	<b>0.556</b>	<b>0.728</b> +14.9%
	Attention-All	0.754	<b>0.900</b>	<b>0.682</b>	<b>0.556</b>	<b>0.728</b> +15.0%

Table 2: Accuracy comparison across various benchmarks using different uncertainty estimators and methods with Llama-3-8B. Zero-shot prompting was used for all benchmarks. The ‘Average’ column shows the average performance for all tests and the % increase over the standard debate.

### 5.1 UNCERTAINTY INCORPORATION METHODS

Table 1 presents a comparison of the results obtained with Mistral-7B using different combinations of uncertainty metrics (Mean Token Entropy, TokenSAR, and Oracle) and methods (Confidence in Prompt, Attention-Others, and Attention-All). As a baseline, the performance of a standard 3-agent, 3-round debate without any uncertainty metrics is also shown. Overall, Attention-All was the top-performing method, achieving the highest average accuracy across all three uncertainty metrics. It was the only method that consistently matched or exceeded the performance of the standard multi-agent debate on all benchmarks. Table 2 presents the results using Llama-3-8B. The findings show that with the Oracle uncertainty metric, the attention scaling methods significantly outperformed confidence communication via prompting. However, when using mean token entropy, the performance gains were minimal, suggesting that mean token entropy may not be as effective on Llama-3-8B as it was on Mistral-7B.

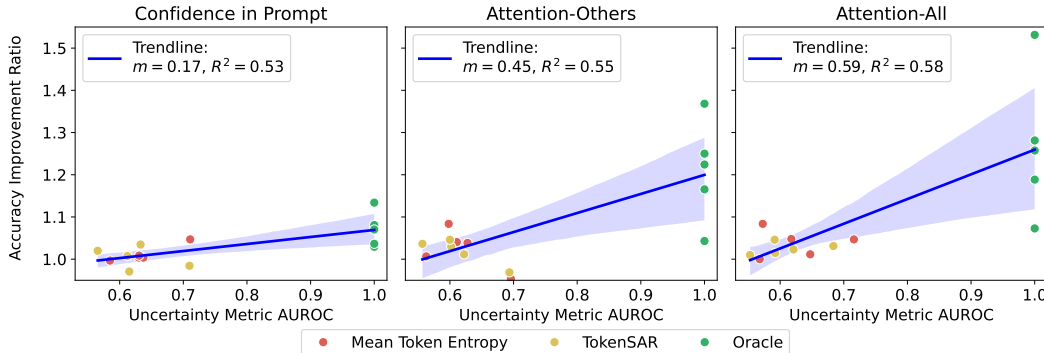


Figure 4: Plots of the ratio of accuracy improvement over a standard debate against the uncertainty metric AUROC for all Mistral-7B experiments. A higher AUROC indicates better metric performance. Each point represents the results on one of the benchmarks (MMLU-0, MMLU-5, GSM8k, Arithmetic, and TruthfulQA). The plots are organized by the method of uncertainty incorporation method (as titled) and the uncertainty metric used (color-coded). The trendlines show that attention-based methods, especially Attention-All, lead to more substantial performance gains as AUROC increases compared to methods that incorporate confidence directly into the prompt.

As shown in Figure 4, Attention-All demonstrates the most significant accuracy improvements as the AUROC of the uncertainty metric increases, with a slope of 0.59 compared to 0.45 for Attention-Others and 0.17 for Confidence in Prompt. The accuracy improvement ratio compares the method’s accuracy to the accuracy observed in a standard debate. AUROC, the area under the receiver operating characteristic curve, represents the probability that a correct answer is assigned a lower uncertainty than an incorrect one. A random uncertainty metric would have an AUROC of 0.5, while a perfect one would have an AUROC of 1.



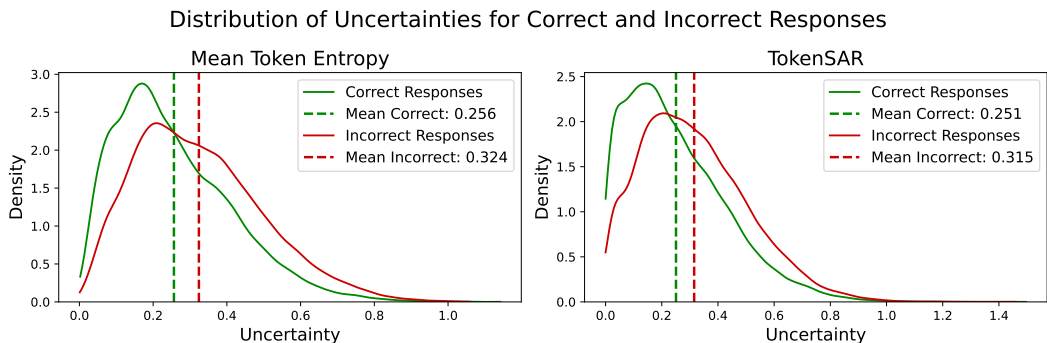
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444445  
446  
447  
448  
449  
450

Figure 5: Distribution of uncertainties for correct and incorrect answers across all Mistral-7B experiments, as measured by the uncertainty metrics Mean Token Entropy and TokenSAR. Generally, correct answers exhibit lower uncertainties than incorrect ones, indicating that although not perfect, uncertainty metrics are useful for distinguishing between accurate responses and those where the agent may be hallucinating.

451  
452

## 5.2 UNCERTAINTY METRICS

453  
454  
455  
456

The best-performing uncertainty metric was the Oracle metric. Mean Token Entropy ranked next, with debates using it consistently achieving higher average accuracies than debates using TokenSAR, as shown in Table 1. Mean Token Entropy achieved an average AUROC across all experiments of 0.627, compared to 0.617 for TokenSAR. Full AUROC results are shown in Appendix B.

457  
458  
459  
460  
461  
462  
463  
464  
465  
466

To further analyze the uncertainty metrics and understand their impact, Figure 5 presents the distributions of uncertainties for responses containing correct answers versus incorrect answers across all benchmarks and uncertainty incorporation methods. The left plot illustrates the uncertainties when using Mean Token Entropy, while the right plot depicts uncertainties when using TokenSAR. As shown in Table 1, Mean Token Entropy achieved slightly higher accuracy than TokenSAR on average. Examining the average uncertainties for correct and incorrect responses, Mean Token Entropy has a ratio of average correct to average incorrect of 1.264, compared to 1.258 for TokenSAR. This suggests that Mean Token Entropy slightly outperforms TokenSAR in differentiating between correct and incorrect responses, while also being less computationally demanding. However, the performance difference is minimal and possibly falls within the margin of error.

467

468

## 6 CONCLUSION

469

470  
471  
472  
473  
474  
475  
476  
477  
478

We introduce DebUnc, a framework that integrates confidence estimations into multi-agent debates. By incorporating confidence levels, agents can assess the reliability of their peers and better determine whom to trust when responses differ. We use uncertainty metrics to estimate the LLM’s confidence based on the probability distribution generated for each token. To convey confidence, we adjusted the LLM’s attention mechanism to scale token weights according to confidence levels and also experimented with using textual prompts. Our results show that attention scaling is more effective, demonstrating that information can be delivered to LLMs through methods beyond traditional text prompts. Attention scaling can be applied in any scenario where certain parts of the input need to be prioritized over others, and we encourage further exploration of this technique.

479

480

## 7 LIMITATIONS

481

482  
483  
484  
485

One limitation of DebUnc is its reliance on open-source LLMs, as attention scaling requires code modifications, and token probabilities may not be accessible from proprietary models. Additionally, attention scaling is sensitive to the order of agent responses within the prompt. For example, if agent 2’s response precedes agent 3’s, unidirectional attention allows agent 2 to influence agent 3’s tokens, potentially leaking information from agent 2 even if agent 3 is more confident.

## REFERENCES

- 486  
487  
488 Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Semantically  
489 diverse language generation for uncertainty estimation in language models. *ArXiv preprint*,  
490 abs/2406.04306, 2024. URL <https://arxiv.org/abs/2406.04306>.
- 491 AI@Meta. Llama 3 model card. 2024. URL [https://github.com/meta-llama/llama3/blob/  
492 main/MODEL\\_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md).
- 493 Alfonso Amayuelas, Liangming Pan, Wenhui Chen, and William Wang. Knowledge of knowledge: Ex-  
494 ploring known-unknowns uncertainty with large language models. *ArXiv preprint*, abs/2305.13712,  
495 2023. URL <https://arxiv.org/abs/2305.13712>.
- 496 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
497 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are  
498 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 499 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and  
500 Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv  
501 preprint*, abs/2308.07201, 2023. URL <https://arxiv.org/abs/2308.07201>.
- 502 Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference  
503 improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023.
- 504 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep  
505 reinforcement learning from human preferences. *Advances in neural information processing  
506 systems*, 30, 2017.
- 507 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
508 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John  
509 Schulman. Training verifiers to solve math word problems, 2021.
- 510 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving  
511 factuality and reasoning in language models through multiagent debate, 2023.
- 512 Hanyu Duan, Yi Yang, and Kar Yan Tam. Do llms know about hallucination? an empirical  
513 investigation of llm’s hidden states. *ArXiv preprint*, abs/2402.09733, 2024. URL <https://arxiv.org/abs/2402.09733>.
- 514 Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura,  
515 and Kaidi Xu. Shifting attention to relevance: Towards the uncertainty estimation of large language  
516 models, 2023.
- 517 Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill  
518 Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy  
519 Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models,  
520 2023. URL <https://arxiv.org/abs/2311.07383>.
- 521 Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy  
522 Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al.  
523 Fact-checking the output of large language models via token-level uncertainty quantification. *ArXiv  
524 preprint*, abs/2403.04696, 2024. URL <https://arxiv.org/abs/2403.04696>.
- 525 Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov.  
526 Don’t hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *ArXiv  
527 preprint*, abs/2402.00367, 2024. URL <https://arxiv.org/abs/2402.00367>.
- 528 Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel,  
529 Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural  
530 machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555,  
531 2020. doi: 10.1162/tacl\_a\_00330. URL <https://aclanthology.org/2020.tacl-1.35>.
- 532 Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Towards uncertainty-aware language agent, 2024.  
533 URL <https://arxiv.org/abs/2401.14016>.

- 540 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob  
541 Steinhardt. Measuring massive multitask language understanding. In *9th International Conference*  
542 *on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net,  
543 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- 544 Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang,  
545 Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent  
546 collaborative framework. *ArXiv preprint*, abs/2308.00352, 2023. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2308.00352)  
547 [2308.00352](https://arxiv.org/abs/2308.00352).
- 549 Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM  
550 hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of*  
551 *the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, Singapore, 2023.  
552 Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL  
553 <https://aclanthology.org/2023.findings-emnlp.123>.
- 554 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,  
555 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,  
556 L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas  
557 Wang, Timoth e Lacroix, and William El Sayed. Mistral 7b, 2023.
- 559 Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas  
560 Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly)  
561 know what they know. *ArXiv preprint*, abs/2207.05221, 2022. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2207.05221)  
562 [2207.05221](https://arxiv.org/abs/2207.05221).
- 563 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for  
564 uncertainty estimation in natural language generation. *ArXiv preprint*, abs/2302.09664, 2023. URL  
565 <https://arxiv.org/abs/2302.09664>.
- 566 Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem.  
567 Camel: Communicative agents for "mind" exploration of large scale language model society. 2023.
- 569 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu,  
570 and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent  
571 debate. *ArXiv preprint*, abs/2305.19118, 2023. URL <https://arxiv.org/abs/2305.19118>.
- 572 Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaying Zhang. Learning to trust your feelings:  
573 Leveraging self-awareness in llms for hallucination mitigation. *ArXiv preprint*, abs/2401.15449,  
574 2024. URL <https://arxiv.org/abs/2401.15449>.
- 575 Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in  
576 words, 2022a. URL <https://arxiv.org/abs/2205.14334>.
- 577 Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human  
578 falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational*  
579 *Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, 2022b. Association for  
580 Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL [https://aclanthology.](https://aclanthology.org/2022.acl-long.229)  
581 [org/2022.acl-long.229](https://aclanthology.org/2022.acl-long.229).
- 582 Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quan-  
583 tification for black-box large language models. *ArXiv preprint*, abs/2305.19187, 2023. URL  
584 <https://arxiv.org/abs/2305.19187>.
- 585 Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating  
586 hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International*  
587 *Conference on Learning Representations*, 2023.
- 588 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike  
589 Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining  
590 approach, 2019.

- 594 Daniel McDonald, Rachael Papadopoulos, and Leslie Benningfield. Reducing llm hallucination using  
595 knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*,  
596 2024.
- 597 OpenAI. Gpt-4 technical report, 2024.
- 599 Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer,  
600 Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through  
601 embeddings. *arXiv preprint arXiv:2310.06272*, 2023.
- 602 Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language under-  
603 standing by generative pre-training. 2018.
- 605 Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models.  
606 *arXiv preprint arXiv:2309.05922*, 2023.
- 607 Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. Towards detecting llms halluci-  
608 nation via markov chain-based multi-agent debate framework. *arXiv e-prints*, pp. arXiv-2406,  
609 2024.
- 611 Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,  
612 and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence  
613 scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*,  
614 2023.
- 615 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,  
616 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike  
617 von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and  
618 Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Confer-  
619 ence on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach,  
620 CA, USA*, pp. 5998–6008, 2017. URL [https://proceedings.neurips.cc/paper/2017/hash/  
621 3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html).
- 622 Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail  
623 Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in  
624 ambiguous tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings  
625 of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long  
626 Papers)*, pp. 11659–11681, Toronto, Canada, 2023. Association for Computational Linguistics.  
627 doi: 10.18653/v1/2023.acl-long.652. URL <https://aclanthology.org/2023.acl-long.652>.
- 628 Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun  
629 Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and  
630 Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.
- 631 Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms  
632 express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint  
633 arXiv:2306.13063*, 2023.
- 634 Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao.  
635 Sayself: Teaching llms to express confidence with self-reflective rationales. *ArXiv preprint*,  
636 abs/2405.20974, 2024. URL <https://arxiv.org/abs/2405.20974>.
- 637 Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to  
638 believe your llm. *ArXiv preprint*, abs/2406.02543, 2024. URL [https://arxiv.org/abs/2406.  
639 02543](https://arxiv.org/abs/2406.02543).
- 642 Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. Llm lies: Hallucinations  
643 are not bugs, but features as adversarial examples. *ArXiv preprint*, abs/2310.01469, 2023. URL  
644 <https://arxiv.org/abs/2310.01469>.
- 645 Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large  
646 language models know what they don’t know? *ArXiv preprint*, abs/2305.18153, 2023. URL  
647 <https://arxiv.org/abs/2305.18153>.

648 Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quan-  
649 tification for llms. *ArXiv preprint*, abs/2403.20279, 2024. URL [https://arxiv.org/abs/2403.](https://arxiv.org/abs/2403.20279)  
650 [20279](https://arxiv.org/abs/2403.20279).

651 Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A  
652 social psychology view. *ArXiv preprint*, abs/2310.02124, 2023. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2310.02124)  
653 [2310.02124](https://arxiv.org/abs/2310.02124).  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

## A EXAMPLE DEBATE

This is a full debate on zero-shot MMLU using Mistral-7B as the LLM. It uses the Confidence in Prompt method for incorporating uncertainty with the Oracle uncertainty metric. For conciseness, the user prompts for rounds 2 and 3 use “<agent  $i$  response>” to reference the response from agent  $i$  in the previous round.

Agent 1	Agent 2	Agent 3
<i>User, Round 1</i>		
Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering.		
Which of these combinations of clinical features is most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation?		
A. Displaced apex beat; soft first heart sound; pan-systolic murmur; short mid-diastolic murmur		
B. Irregularly, irregular pulse; displaced apex beat; ejection systolic murmur; short mid-diastolic murmur		
C. Displaced apex beat; normal first heart sound; pan-systolic murmur; long mid-diastolic murmur		
D. Tapping apex beat; loud first heart sound; pan-systolic murmur; long mid-diastolic murmur		
<i>Assistant, Round 1</i>		
Based on the given clinical features, option C: Displaced apex beat; normal first heart sound; pan-systolic murmur; long mid-diastolic murmur is least suggestive of mixed mitral valve disease with a predominance of mitral regurgitation.	Based on the given clinical features, the combination most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation would be:  C. Displaced apex beat; normal first heart sound; pan-systolic murmur; long mid-diastolic murmur	The combination of a displaced apex beat, a pan-systolic murmur, and a long mid-diastolic murmur is most suggestive of mitral stenosis. Thus, the answer to the question is D. however, we were asked for a combination that is most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation. With this in mind, the correct answer is: C. Displaced apex beat; normal first heart sound; pan-systolic murmur; long mid-diastolic murmur
The features most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation are:	Explanation: The presence of a displaced apex beat and a pan-systolic murmur points towards mitral valve disease. However, the duration and timing of the additional diastolic murmur helps distinguish the underlying etiology. A long mid-diastolic murmur is more suggestive of functional mitral regurgitation, which is a component of mixed mitral valve disease with a predominance of mitral regurgitation. Therefore, option C is the correct answer.	Explanation: Mixed mitral valve disease refers to a condition in which both mitral regurgitation and mitral stenosis are present. The presence of a predominantly regurgitant jet suggests a more advanced state of the disease. A displaced apex beat can be seen in conditions of volume overload, such as mitral regurgitation, and is less common in mitral stenosis. A normal first heart sound and a long mid-diastolic murmur are more likely to be heard in mitral regurgitation than mitral stenosis, as they represent the opening and closing of the mitral valve during diastole. Therefore, option C, with a displaced apex beat, normal first heart sound, and a long mid-diastolic murmur, is the most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation.
1. Displaced apex beat: This can occur due to chronic volume overload or dilatation of the left ventricle from mitral regurgitation.		
2. Soft or absent first heart sound: This can be due to the decreased force of mitral valve closure during diastole.		
3. Pan-systolic murmur: This is typically heard in mitral regurgitation due to the continuous regurgitant jet throughout the cardiac cycle.		
4. Long mid-diastolic murmur: This can occur due to the continued mitral regurgitant jet during diastole.	<b>Answer: C</b>	
Therefore, the correct answer is:		
A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur		
<b>Answer: A</b>		<b>Answer: C.</b>

Agent 1 (cont.)	Agent 2 (cont.)	Agent 3 (cont.)
<i>User, Round 2</i>		
<p>These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:</p> <p>One agent solution (confidence level is 1): ""&lt;agent 2 response&gt;""</p> <p>One agent solution (confidence level is 1): ""&lt;agent 3 response&gt;""</p> <p>Based off the opinion of other agents, can you give an updated response? Do not mention your confidence. Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD.</p>	<p>These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:</p> <p>One agent solution (confidence level is 10): ""&lt;agent 1 response&gt;""</p> <p>One agent solution (confidence level is 1): ""&lt;agent 3 response&gt;""</p> <p>Based off the opinion of other agents, can you give an updated response? Do not mention your confidence. Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD.</p>	<p>These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:</p> <p>One agent solution (confidence level is 10): ""&lt;agent 1 response&gt;""</p> <p>One agent solution (confidence level is 1): ""&lt;agent 2 response&gt;""</p> <p>Based off the opinion of other agents, can you give an updated response? Do not mention your confidence. Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD.</p>
<i>Assistant, Round 2</i>		
<p>Based on the information provided by the other agents, the combination most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation is:</p> <p>A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur.</p> <p>The presence of a displaced apex beat, soft first heart sound, and prolonged pan-systolic murmur indicate mitral valve disease. The duration and timing of the diastolic murmur can be used to distinguish between mitral stenosis and mitral regurgitation. A long, holodiastolic murmur is more commonly heard in mitral regurgitation than mitral stenosis. Therefore, the presence of a long mid-diastolic murmur with a displaced apex beat and soft first heart sound suggests mixed mitral valve disease with a predominance of mitral regurgitation.</p> <p>Answer: A.</p>	<p>Based on the clinical features provided and the consensus of other agents, option C: Displaced apex beat; normal first heart sound; pan-systolic murmur; long mid-diastolic murmur is the most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation.</p> <p>Answer: C.</p>	<p>Based on the discussions among different agents, the most suggestive combination of clinical features for mixed mitral valve disease with a predominance of mitral regurgitation is:</p> <p>A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur</p> <p>Explanation: Mixed mitral valve disease is characterized by the co-existence of mitral stenosis and mitral regurgitation. The presence of a displaced apex beat, soft first heart sound, pan-systolic murmur, and long mid-diastolic murmur can suggest a diagnosis of mixed mitral valve disease with a predominance of mitral regurgitation. The displaced apex beat can occur due to chronic volume overload or dilatation of the left ventricle, as seen in mitral regurgitation. The soft or absent first heart sound can be due to the decreased force of mitral valve closure in mitral regurgitation. The pan-systolic murmur is typically heard in mitral regurgitation due to the continuous regurgitant jet throughout the cardiac cycle. The long mid-diastolic murmur can occur due to the continued mitral regurgitant jet during diastole. Therefore, the combination of these clinical findings, A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur, is most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation.</p> <p>Answer: A.</p>

	Agent 1 (cont.)	Agent 2 (cont.)	Agent 3 (cont.)
	<i>User, Round 3</i>		
813	These are solutions and confidence	These are solutions and confidence	These are solutions and confidence
814	values from 1 to 10 (higher means	values from 1 to 10 (higher means	values from 1 to 10 (higher means
815	more confident) to the problem from	more confident) to the problem from	more confident) to the problem from
816	other agents:	other agents:	other agents:
817	One agent solution (confidence level is	One agent solution (confidence level is	One agent solution (confidence level is
818	1): ""<agent 2 response>""	10): ""<agent 1 response>""	10): ""<agent 1 response>""
819	One agent solution (confidence level is	One agent solution (confidence level is	One agent solution (confidence level is
820	10): ""<agent 3 response>""	10): ""<agent 3 response>""	1): ""<agent 2 response>""
821	Based off the opinion of other agents,	Based off the opinion of other agents,	Based off the opinion of other agents,
822	can you give an updated response? Do	can you give an updated response? Do	can you give an updated response? Do
823	not mention your confidence. Think	not mention your confidence. Think	not mention your confidence. Think
824	step by step before answering. The last	step by step before answering. The last	step by step before answering. The last
825	line of your response should be of the	line of your response should be of the	line of your response should be of the
826	following format: 'Answer:	following format: 'Answer:	following format: 'Answer:
827	\$LETTER' (without quotes) where	\$LETTER' (without quotes) where	\$LETTER' (without quotes) where
	LETTER is one of ABCD.	LETTER is one of ABCD.	LETTER is one of ABCD.
828	<i>Assistant, Round 3</i>		
829	Based on the discussions among	Based on the consensus of other	Based on the information provided by
830	different agents, the combination of	agents, the combination most	various agents and the consensus
831	clinical features most suggestive of	suggestive of mixed mitral valve	reached, the combination of clinical
832	mixed mitral valve disease with a	disease with a predominance of mitral	features most suggestive of mixed
833	predominance of mitral regurgitation	regurgitation is: A. Displaced apex	mitral valve disease with a
834	is:	beat; soft first heart sound;	predominance of mitral regurgitation
835	A. Displaced apex beat; soft first heart	pan-systolic murmur; long	is:
836	sound; pan-systolic murmur; long	mid-diastolic murmur.	A. Displaced apex beat; soft first heart
837	mid-diastolic murmur.	<i>Answer: A.</i>	sound; pan-systolic murmur; long
838	Explanation:		mid-diastolic murmur
839	Mixed mitral valve disease is defined		Here's the explanation:
840	as the presence of both mitral stenosis		The presence of a displaced apex beat
841	and mitral regurgitation. The clinical		and a pan-systolic murmur suggest
842	presentation of mixed mitral valve		mitral valve disease. However, the
843	disease can vary depending on the		duration and timing of the diastolic
844	severity and predominance of each		murmur can help distinguish between
845	component.		mitral stenosis and mitral regurgitation.
846	A displaced apex beat, soft first heart		Mitral regurgitation is more commonly
847	sound, pan-systolic murmur, and long		associated with a long, holodiastolic or
848	mid-diastolic murmur are more		mid-diastolic murmur, while mitral
849	indicative of mitral regurgitation than		stenosis typically has a short,
850	mitral stenosis. The displaced apex		mid-diastolic murmur. In this scenario,
851	beat can result from the increased		the long mid-diastolic murmur, along
852	volume load in the left ventricle due to		with the displaced apex beat and soft
853	mitral regurgitation. A soft or absent		first heart sound, makes mixed mitral
854	first heart sound may be heard in		valve disease with a predominance of
855	mitral regurgitation due to the		mitral regurgitation the most likely
856	decreased force of mitral valve closure,		diagnosis. Therefore, the answer is:
857	while the pan-systolic murmur and		<i>Answer: A.</i>
858	long mid-diastolic murmur are		
859	characteristic features of mitral		
860	regurgitation.		
861	Therefore, the combination of these		
862	clinical findings, A. Displaced apex		
863	beat; soft first heart sound;		
	pan-systolic murmur; long		
	mid-diastolic murmur, is most		
	suggestive of mixed mitral valve		
	disease with a predominance of mitral		
	regurgitation.		
	<i>Answer: A.</i>		



## B FULL UNCERTAINTY METRIC RESULTS

Estimator	Method	MMLU-0	MMLU-5	GSM8k	Arithmetic	TruthfulQA	Avg.
Entropy	Prompt	0.630	0.637	0.711	0.631	0.585	0.639
	Attn-Others	0.628	0.611	0.696	0.598	0.563	0.619
	Attn-All	0.647	0.618	0.716	0.573	0.568	0.624
TokenSAR	Prompt	0.633	0.612	0.710	0.615	0.566	0.627
	Attn-Others	0.622	0.602	0.694	0.600	0.557	0.615
	Attn-All	0.621	0.592	0.684	0.591	0.553	0.608
Oracle	Prompt	1.000	1.000	1.000	1.000	1.000	1.000
	Attn-Others	1.000	1.000	1.000	1.000	1.000	1.000
	Attn-All	1.000	1.000	1.000	1.000	1.000	1.000

Table 3: This table displays the uncertainty metric AUROC values for each experiment run with Mistral-7B. ‘MMLU-0’ denotes zero-shot prompting on MMLU, while ‘MMLU-5’ represents 5-shot prompting. The other benchmarks used zero-shot prompting. The ‘Avg.’ column shows the average AUROC for all tests. By definition, the Oracle metric achieved perfect AUROC scores. Among the remaining metrics, Mean Token Entropy delivered slightly better performance than TokenSAR despite being cheaper to run.

## C PROMPTS

The table below displays the prompts used to facilitate the debates. Text in **green** is only included when using the Confidence in Prompt method, and text in **blue** is specific to each debate.

Benchmark	Prompt Type	Prompt
MMLU	Initial, 0-shot	<p>Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering.</p> <p>&lt;question&gt;</p>
	Initial, 5-shot	<p>Answer the following multiple choice question.</p> <p>Examples:</p> <p>&lt;5 example questions and answers&gt;</p> <p>—</p> <p>YOUR TASK</p> <p>Answer the following question. Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD.</p> <p>&lt;question&gt;</p>
	Debate	<p>These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:</p> <p>One agent solution (confidence level is <math>c</math>): &lt;agent response&gt;</p> <p>One agent solution (confidence level is <math>c</math>): &lt;agent response&gt;</p> <p>Based off the opinion of other agents, can you give an updated response? Do not mention your confidence. Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of ABCD.</p>
GSM8k	Initial	<p>Answer the following math problem. The last line of your response should be of the following format: 'Answer: \$INTEGER' (without quotes) where INTEGER is the integer answer. Think step by step before answering.</p> <p>&lt;question&gt;</p>
	Debate	<p>These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:</p> <p>One agent solution (confidence level is <math>c</math>): &lt;agent response&gt;</p> <p>One agent solution (confidence level is <math>c</math>): &lt;agent response&gt;</p> <p>Based off the opinion of other agents, can you provide an updated response? The original problem is:</p> <p>&lt;question&gt;</p> <p>Do not mention your confidence. The last line of your response should be of the following format: 'Answer: \$INTEGER' (without quotes) where INTEGER is the integer answer.</p>

972	<b>Benchmark</b>	<b>Prompt Type</b>	<b>Prompt</b>
973	Arithmetic	Initial	What is the result of $\langle a+b*c+d \rangle$ ? State the final answer at the end of your response.
974		Debate	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents:
975			One agent solution (confidence level is $c$ ): $\langle \text{agent response} \rangle$
976			One agent solution (confidence level is $c$ ): $\langle \text{agent response} \rangle$
977			Based off the opinion of other agents, can you provide an updated answer? Do not mention your confidence. State the final answer at the end of your response.
978	TruthfulQA	Initial	Answer the following multiple choice question:
979			$\langle \text{question} \rangle$
980			Think step by step before answering. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of $\langle \text{options} \rangle$ .
981		Debate	These are the selections and confidence values from 1 to 10 (higher means more confident) from other agents:
982			One agent solution (confidence level is $c$ ): $\langle \text{agent response} \rangle$
983	One agent solution (confidence level is $c$ ): $\langle \text{agent response} \rangle$		
984			Can you double check that your response is correct? Do not mention your confidence. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of $\langle \text{options} \rangle$ .
985			
986			
987			
988			
989			
990			
991			
992			
993			
994			
995			
996			
997			
998			
999			
1000			
1001			
1002			
1003			
1004			
1005			
1006			
1007			
1008			
1009			
1010			
1011			
1012			
1013			
1014			
1015			
1016			
1017			
1018			
1019			
1020			
1021			
1022			
1023			
1024			
1025			