DebUnc: Improving Large Language Model Agent Communication With Uncertainty Metrics

Anonymous ACL submission

Abstract

Multi-agent debates have been introduced to improve the accuracy of Large Language Models (LLMs) by having multiple agents discuss solutions to a problem over several rounds of debate. However, models often generate incorrect yet confident-sounding responses, which can mislead others. This issue arises partly because agents do not consider how confident their peers are. To address this, we propose DebUnc, a debate framework that uses uncertainty metrics to assess agent confidence. Confidence is then conveyed through a modified attention mechanism that adjusts token weights, or through textual prompts. Evaluations across benchmarks show that attention-based methods are particularly effective and that performance continues to improve as uncertainty estimation becomes more reliable.

1 Introduction

Large language models (LLMs) have demonstrated impressive performance across various domains, including law, academia, and coding (OpenAI, 2024). However, they are prone to hallucinations, where they confidently generate false or misleading information (Rawte et al., 2023). This poses significant risks in real-world applications. For example, an LLM tutor providing incorrect explanations could mislead students, while a customer service agent giving faulty advice could frustrate users. In high-stakes fields like healthcare, journalism, or finance, hallucinations can have severe consequences, including financial loss or health risks. To mitigate these issues, researchers have explored multi-agent debate, where multiple LLMs propose diverse solutions and critique each other's reasoning over several rounds of debate(Liang et al., 2023).This process has been shown to enhance the reasoning and accuracy of LLMs, outperforming simpler approaches such as majority voting or chain-of-thought prompting (Du et al., 2023).

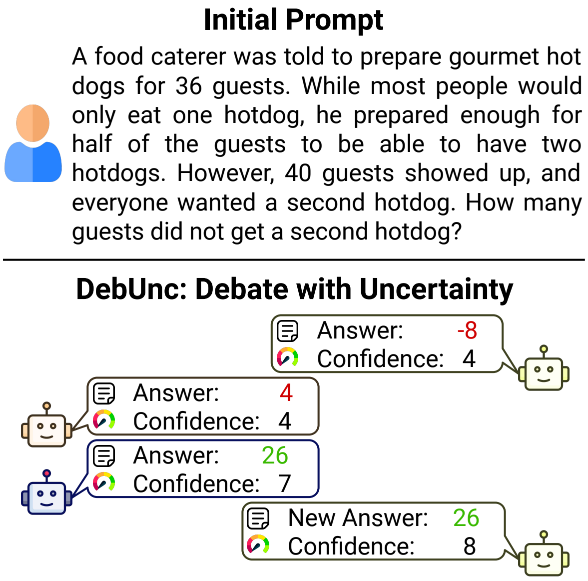

Figure 1: Example three-agent debate. The first agent initially provides an incorrect response but corrects itself after considering the answers and confidence levels of others. Each agent uses a LLM to generate its response and an uncertainty metric to assesses its confidence. Correct answers are shown in green, while incorrect ones are shown in red.

Ideally, debate should help agents recognize and correct errors when some provide incorrect answers. In practice, while agents often agree on the same final answer, it is sometimes incorrect. This issue stems from flawed communication, as LLMs typically respond with high confidence regardless of accuracy (Du et al., 2023). A confidently incorrect response can mislead other agents, causing the system to converge on an incorrect conclusion. Because it is difficult to gauge a LLM's certainty based on its response alone, uncertainty metrics have been developed to provide a more reliable confidence measure.

By contrast, people often use qualifiers such as "I am sure that..." or "I am not sure, but..." to express confidence during discussions. These cues help others assess the reliability of information.

Building on these insights, we present DebUnc, a novel multi-agent debate framework that combines multi-agent **Deb**ates and model **Unc**ertainty metrics. After each round of debate, we measure each agent's confidence with an uncertainty metric. In the following round, both the agents' responses and confidences are shared with the other agents. We explore two methods for communicating agent uncertainty: incorporating the uncertainty directly into the textual prompt, as shown in Figure 2, and shifting the model's attention towards more confident agents, as depicted in Figure 3. We evaluate DebUnc across multiple LLMs, benchmarks, and uncertainty metrics.

Our key contributions are outlined as follows:

1. We introduce DebUnc, a framework designed to quantify and communicate the uncertainty of LLM agents within multi-agent debates.

2. We propose an attention-scaling mechanism that guides the model's focus towards more confident agents, serving as an alternative to conveying uncertainty through textual prompts.

3. We evaluate DebUnc across multiple LLMs, benchmarks, and uncertainty metrics, and find that debates using attention-scaling to communicate confidence outperform those using textual prompts to communicate confidence, as well as debates without any uncertainty communication.

4. We provide insights into future improvements, exploring how more robust uncertainty metrics could further enhance debate performance.

2 Related Work

LLMs are prone to overconfidence and often generate responses regardless of their certainty. This leads to hallucinations, where the information provided by the model is incorrect or unsupported by its training data (Liang et al., 2024; Yadkori et al., 2024; Duan et al., 2024; Yao et al., 2023; Aichberger et al., 2024). Factual accuracy is crucial for building trust in LLM-based systems and enabling real-world applications. Recent work has focused on understanding the causes of hallucinations and developing strategies to mitigate them (Ji et al., 2023; McDonald et al., 2024; Liu et al., 2023).

2.1 Uncertainty in LLMs

One current effort to mitigate hallucinations focuses on measuring a model's uncertainty and enhancing its self-awareness (Kadavath et al., 2022; Amayuelas et al., 2023; Yin et al., 2023).

Accurate confidence estimates help users judge when to trust model outputs (Lin et al., 2022a; Xu et al., 2024) and guide agents in deciding when to use external tools (Han et al., 2024). However, uncertainty estimation remains an open challenge. Ongoing research explores new uncertainty metrics, which typically fall into one of three categories:

Token Probability-Based Uncertainty Metrics compute uncertainty from the token-level probabilities generated by the model. High token probabilities reflect confidence, while low probabilities indicate uncertainty. Notable methods in this category include Mean Token Entropy, Perplexity (Fomicheva et al., 2020), SAR (Duan et al., 2023), RDE (Vazhentsev et al., 2023), and Claim-Conditioned Probability (Fadeeva et al., 2024).

LLM-Generated Uncertainty Metrics rely on the model expressing its own uncertainty. For example, Lin et al. (2022a) fine-tuned GPT-3 (Brown et al., 2020) to respond with both an answer and a confidence level. Another approach is to prompt the model to express its uncertainty without explicit training, which (Tian et al., 2023) found to outperform token probability-based methods for models fine-tuned with reinforcement learning from human feedback (Christiano et al., 2017). However, Xiong et al. (2023) report lower performance compared to token probability-based methods on GPT-3.

Sampling-Based Uncertainty Metrics evaluate uncertainty by sampling multiple responses and assesing variations in meaning. Consistency in meaning across samples indicates confidence, while variations suggest uncertainty. These methods generally outperform token probability-based metrics, but are more resource-intensive due to the need for multiple generations. Examples include Semantic Entropy (Kuhn et al., 2023), LUQ (Zhang et al., 2024a), and other metrics that evaluate meaning diversity (Lin et al., 2023).

2.2 Multi-Agent Debate

With the increased accessibility of LLMs and improvements in their inference times, LLMs are being integrated into more complex systems as autonomous agents (Wu et al., 2023; Li et al., 2023; Hong et al., 2023). A critical component of agent-based systems is collaboration, where models debate each other to refine outputs. (Zhang et al., 2023) Such mechanisms promote divergent thinking (Liang et al., 2023), improve reasoning and factual accuracy (Du et al., 2023), and lead to more reliable evaluations (Chan et al., 2023).

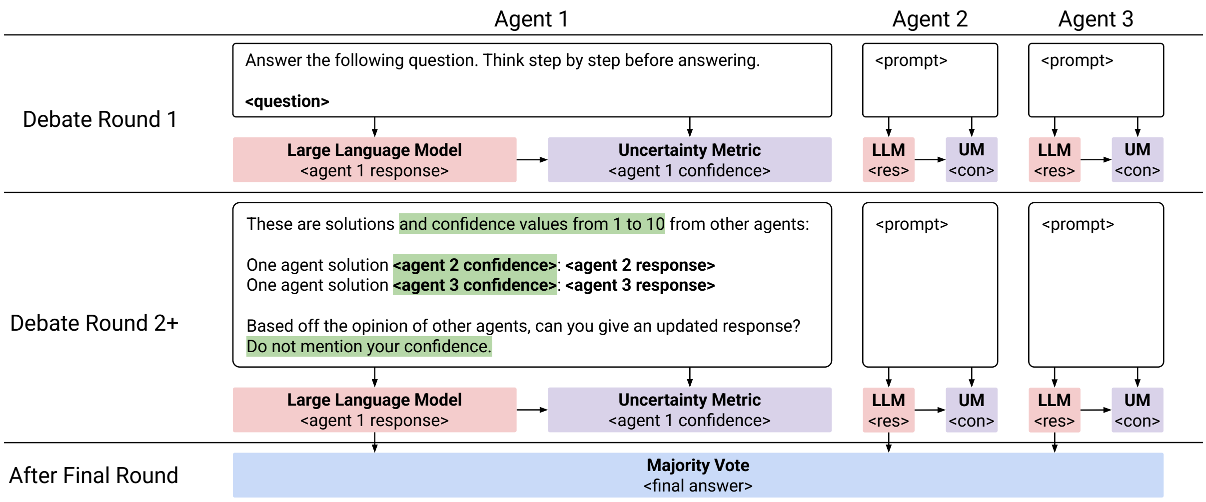

Figure 2: Illustration of the modified multi-agent debate involving three agents. In the first round, each agent independently generates a response to the question, which is evaluated for confidence using an uncertainty metric. The prompt for following rounds includes the responses from other agents 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.

Through discussions, the LLMs can refine their outputs, ultimately achieving higher levels of agreement and producing more factually accurate text (Sun et al., 2024; Feng et al., 2024). (Pham et al., 2023) recognized that text is not be the most effective communication mechanism for LLM agents, since information is lost during the token sampling process, and demonstrated how LLMs can communicate through embeddings. ReConcile (Chen et al., 2023) explored the integration of agent confidence in multi-agent debates, relying on LLM agents to self-report their confidence, which was communicated to other agents through prompts.

2.3 Attention Shifting

Another approach to mitigating hallucinations involves informing the model about the relative importance and reliability of information in the prompt. For instance, if certain parts of the input originate from a highly trusted source, the model should weigh them more heavily than less reliable content. Zhang et al. (2024b) introduced PASTA, a technique that shifts attention away from less relevant tokens toward more important ones, such as instructions. This approach substantially improved the model's ability to follow instructions and incorporate new knowledge.

Building on these ideas, we use uncertainty metrics in multi-agent debates to estimate each agent's confidence, which is then communicated in the next round through prompting or attention shifting.

3 Method

In multi-agent LLM debates, agents often produce confident-sounding yet inaccurate responses, potentially misleading other agents and leading to consensus on incorrect answers (Du et al., 2023). Our goal is to guide agents on which opinions to prioritize based on uncertainty levels. Our modified debate pipeline, illustrated in Figure 2, operates as follows: in each round of debate, every agent generates a response and its uncertainty is assessed. In the subsequent round, each agent shares its response and uncertainty with all other agents. We evaluate on three uncertainty metrics and three methods for communicating agent uncertainty.

3.1 Uncertainty Metrics

Uncertainty metrics assess an LLM's confidence in its responses: high uncertainty indicates low confidence and potential unreliability, while low uncertainty suggests greater confidence and reliability. These metrics generally fall into one of three categories: token probability-based, sampling-based, or LLM-generated methods. For more details, see Uncertainty in LLMs. In our experiments, we focus on token probability-based metrics due to their efficiency and simplicity, as they require only a single generation and do not rely on the model's ability to express uncertainty, unlike sampling-based and LLM-generated metrics, respectively. However, our methods could be used with any uncertainty metric.

Specifically, we selected Mean Token Entropy (Fomicheva et al., 2020) for its simplicity and TokenSAR (Duan et al., 2023), a more advanced approach that recognizes that certain tokens contribute more significantly to a sequence's meaning than others. We utilized implementations from LM-Polygraph, a framework that provides implementations for various uncertainty metrics (Fadeeva et al., 2023). Lastly, to evaluate the potential of future uncertainty metrics, we include a third "Ground Truth" uncertainty metric in our analysis.

Mean Token Entropy. One of the simplest and most computationally efficient uncertainty metrics is Mean Token Entropy, adding negligible runtime overhead to debates (Fomicheva et al., 2020). It is calculated as the average entropy across all tokens generated, with the entropy H 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 introduced TokenSAR, an uncertainty metric that accounts for this. TokenSAR is defined as the weighted average of the negative log probabilities of each generated token, with weights determined by 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. Computing relevance requires running RoBERTa-large (Liu et al., 2019) once per token, for a total of N times. This makes TokenSAR more expensive than Mean Token Entropy, increasing runtime by around 50% in our Llama-3-8B-Instruct experiments on a NVIDIA Titan RTX. Still, it remains much cheaper than multi-generation methods.

Ground Truth Uncertainty. Although metrics such as Mean Token Entropy and TokenSAR provide useful estimates of uncertainty, they are imperfect, and future advancements will likely lead to more accurate metrics. To assess the potential effectiveness of our uncertainty communication methods with improved uncertainty metrics, we introduce a diagnostic metric we refer to as Ground Truth Uncertainty. This metric simulates an ideal uncertainty signal by assigning low uncertainty to correct responses and high uncertainty to incorrect ones:

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

We detail exactly how this metric is applied in the following subsection. Since this metric relies on access to the correct answer, it cannot be used in real-world applications. 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 need to communicate it to the other agents. We explore several 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. However, it may be more natural to express confidence, where higher is better, rather than uncertainty, where lower is better. Humans, for example, often describe their confidence on a scale from 1 to 10. Since LLMs are trained on human data, they may exhibit a similar preference.

Mean Token Entropy and TokenSAR both yield non-negative uncertainties, but their scales differ: Mean Token Entropy has an upper bound that depends on the model's vocabulary size, while TokenSAR has an unbounded maximum. Because of this, absolute uncertainty values are less informative than the relative differences in uncertainty between agents.

In order to convert uncertainty values into confidence values, we perform the following steps. Given a list of uncertainties u for n agents, where u_i is the uncertainty of agent i, we first invert them to obtain unscaled confidence values $c_i = \frac{1}{u_i}$. We then scale these values such that the average confidence s_i of all agents is 5:

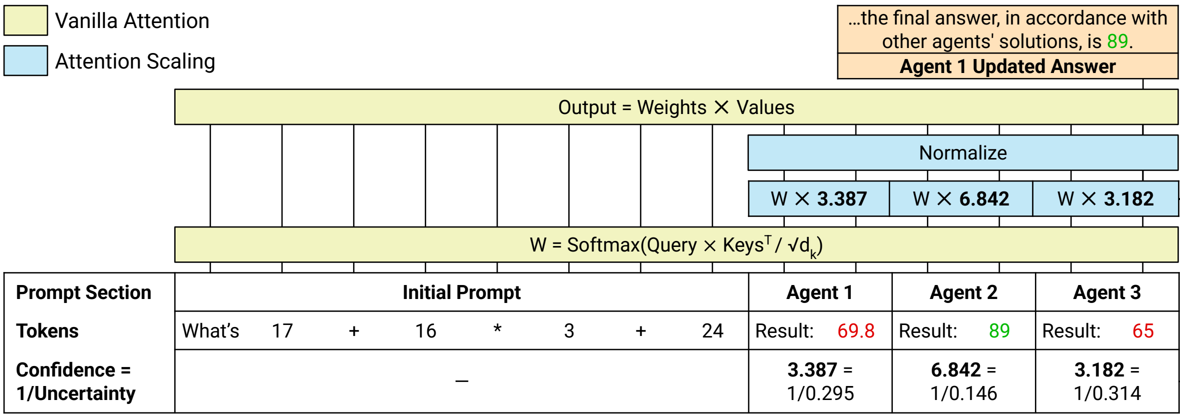

Figure 3: Illustration of the Attention-All method from the perspective of Agent 1. As the second debate round begins, the model's context includes the initial prompt and each agent's responses. Agent 2 provided a correct response with lower uncertainty than Agents 1 and 3, who responded incorrectly. 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.

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

Finally, we clamp the confidence levels to the range of 1 to 10 and round to the nearest integer. When using the Ground Truth uncertainty metric, we set the confidence to 1 if the agent was incorrect and to 10 if the agent was correct.

Attention Scaling. Another approach to communicate confidence levels is to 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 by softmax to ensure they sum to 1. The output embedding is the weighted sum of the value vectors of each token (Vaswani et al., 2017). As a result, the weight of each token determines its influence on the next token generated. By modifying these weights, we can adjust the model's focus to each token in the input. 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. We compute the uncertainty of each agent using an uncertainty metric. In the next round, each agent's prompt will include these responses. As the LLM generates its next response, it computes the normalized attention weights for each preceding token. We divide the weight w_i of every token from agent j by m_j, which is defined as the inverse of agent j's uncertainty when using Mean Token Entropy or TokenSAR.

When using the Ground Truth uncertainty metric, to avoid divide-by-zero errors, we set m_j to 10^{-5} if agent j was incorrect and 1 if agent j was correct.

Formally, the scaled weight a_i for token 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. After this, the attention weights may no longer sum to 1, so another normalization step is needed. We only normalize the weights at indices that were scaled, leaving the weights of other tokens unchanged. The final weight f_i for token i is:

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

We only apply attention scaling to responses from the previous round, as confidence from earlier rounds may become outdated as the agent refines its answer. For instance, in a three-round debate, attention would be rescaled for first-round responses during the second round, and for the second-round responses during the third round. However, first-round responses would not be rescaled during the third round.

Metric	Method	MMLU-0	MMLU-5	GSM8k	Truthful	Arithmetic	Average
—	Standard	0.52 ± 0.02	0.54 ± 0.02	0.51 ± 0.05	0.60 ± 0.03	0.48 ± 0.03	0.53 ± 0.01
Entropy	Prompt	0.52 ± 0.02	0.55 ± 0.03	**0.54** ± 0.04	0.60 ± 0.04	0.48 ± 0.05	0.54 ± 0.01
	Attn-Others	**0.54** ± 0.02	0.57 ± 0.04	0.49 ± 0.03	**0.61** ± 0.03	**0.52** ± 0.09	0.54 ± 0.02
	Attn-All	0.53 ± 0.05	**0.57** ± 0.04	**0.54** ± 0.03	0.60 ± 0.03	**0.52** ± 0.10	**0.55** ± 0.02
SAR	Prompt	**0.54** ± 0.03	0.55 ± 0.05	0.50 ± 0.03	0.62 ± 0.01	0.46 ± 0.04	0.53 ± 0.01
	Attn-Others	0.53 ± 0.04	**0.56** ± 0.04	0.50 ± 0.03	**0.63** ± 0.03	0.50 ± 0.06	0.54 ± 0.01
	Attn-All	0.53 ± 0.02	0.55 ± 0.04	**0.53** ± 0.06	0.61 ± 0.03	**0.50** ± 0.10	**0.54** ± 0.02
Ground-Truth*	Prompt	0.56 ± 0.03	0.56 ± 0.03	0.55 ± 0.04	0.63 ± 0.03	0.54 ± 0.06	0.57 ± 0.01
	Attn-Others	0.61 ± 0.04	0.67 ± 0.03	0.64 ± 0.04	0.63 ± 0.03	0.65 ± 0.04	0.64 ± 0.01
	Attn-All	**0.62** ± 0.03	**0.68** ± 0.06	**0.66** ± 0.05	**0.65** ± 0.02	**0.73** ± 0.05	**0.67** ± 0.01

Table 1: Accuracy comparison ± 95% CI across various benchmarks using different uncertainty metrics (*requires access to the ground truth answer) and communication methods with Mistral-7B. 'MMLU-0' and 'MMLU-5' represent 0 and 5-shot performance on MMLU respectively. The other benchmarks used zero-shot prompting. The 'Average' column shows the average performance over all benchmarks.

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 answer tokens.

We explore two variants of attention scaling:

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

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. In addition, the uncertainty metrics used are based on token probabilities that 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). Tokens were sampled with a temperature of 1 to ensure responses varied. 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 communication method on these samples five times, and report the mean accuracy across the five runs ± the 95% confidence interval.

Parameter	Details
LLMs	Mistral-7B, Llama-3-8B
Temperature	1
Unc. Metrics	Entropy, TokenSAR, Ground Truth
Benchmarks	**MMLU** Broad knowledge **GSM8k** Math problems **TruthfulQA** Misconceptions **Arithmetic** Random equations
Prompting	MMLU: 0/5-shot for Mistral Others: 0-shot
Questions	100 sampled per benchmark
Repetitions	5 runs per method
Debate Setup	3 agents, 3 rounds, same LLM

Table 2: Experiment Design Summary

We evaluated on the following 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 $0 \leq a, b, c, d < 30$.

Metric	Method	MMLU-0	GSM8k	TruthfulQA	Arithmetic	Average
—	Standard	0.65 ± 0.03	0.81 ± 0.04	0.52 ± 0.02	0.52 ± 0.05	0.63 ± 0.01
Entropy	Prompt	0.61 ± 0.02	**0.84** ± 0.05	0.54 ± 0.05	0.53 ± 0.05	0.63 ± 0.02
	Attn-Others	0.64 ± 0.06	0.81 ± 0.03	**0.56** ± 0.05	0.53 ± 0.08	0.63 ± 0.02
	Attn-All	**0.66** ± 0.03	0.81 ± 0.03	0.56 ± 0.04	**0.53** ± 0.05	**0.64** ± 0.01
Ground Truth*	Prompt	0.67 ± 0.03	0.87 ± 0.02	0.58 ± 0.04	0.55 ± 0.05	0.67 ± 0.01
	Attn-Others	**0.78** ± 0.05	0.90 ± 0.03	0.67 ± 0.02	0.56 ± 0.06	**0.73** ± 0.02
	Attn-All	0.75 ± 0.03	**0.90** ± 0.02	**0.68** ± 0.04	**0.56** ± 0.05	0.72 ± 0.01

Table 3: Accuracy comparison ± 95% CI across various benchmarks using different uncertainty metrics (*requires access to the ground truth answer) and communication methods with Llama-3-8B. Zero-shot prompting was used for all benchmarks. The 'Average' column shows the average performance over all benchmarks.

For MMLU on Mistral-7B, we tested both zero-shot and 5-shot prompting with examples from the original MMLU repository's[1] dev set. For other benchmarks, we used only zero-shot prompting. The prompts used are shown in Appendix C.

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.

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 Ground Truth) 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 3 presents the results using Llama-3-8B. The findings show that with the Ground Truth 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.

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.

5.2 Uncertainty Metrics

The best-performing uncertainty metric was, by definition, the Ground Truth metric. Mean Token Entropy ranked next, with debates using it typically 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.

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 from Mistral-7B across all benchmarks and uncertainty incorporation methods.

[1] https://github.com/hendrycks/test

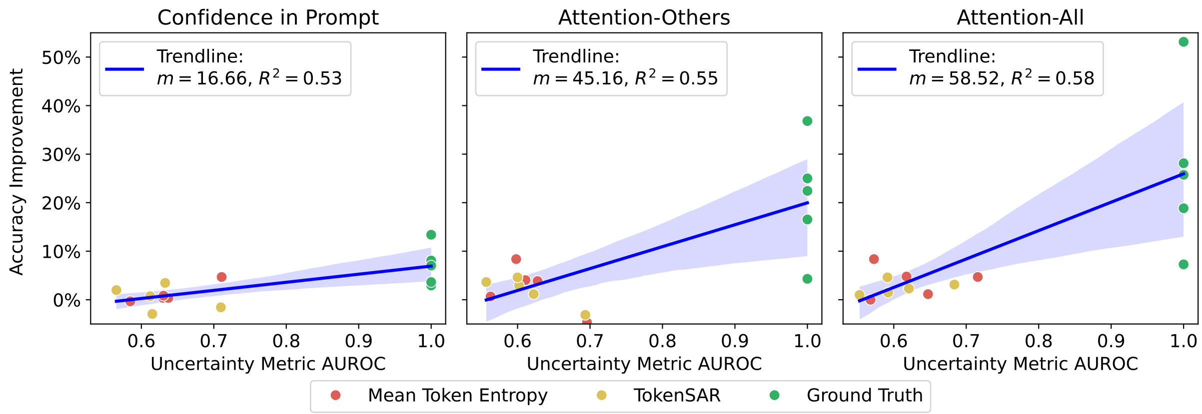

Figure 4: Plots showing the percent increase in accuracy over standard debate versus uncertainty metric AUROC for a given combination of benchmark, uncertainty metric, and trial using Mistral-7B. A higher AUROC indicates better metric performance. The plots are titled by uncertainty incorporation method and color-coded by the uncertainty metric used. 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.

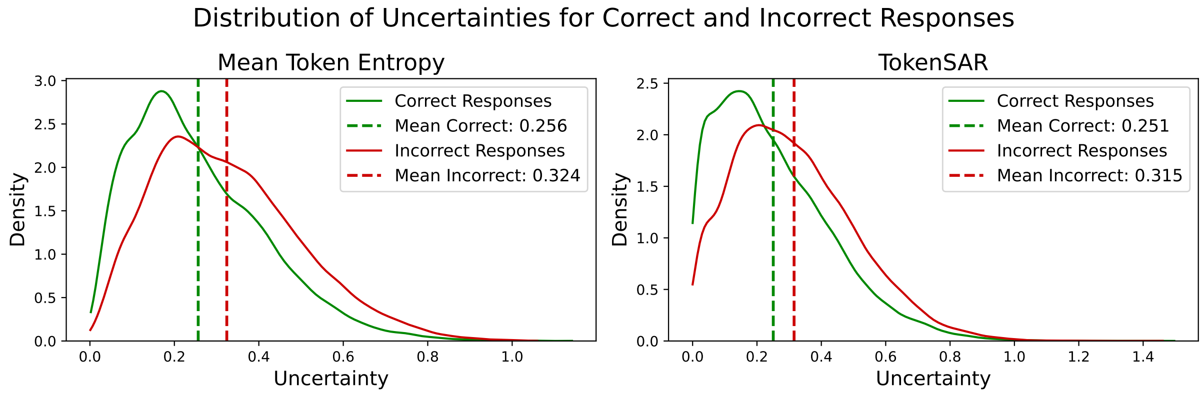

Figure 5: Distribution of uncertainties for correct and incorrect answers across all Mistral-7B experiments, as measured by the uncertainty metrics Mean Token Entropy and TokenSAR. Generally, correct answers exhibit lower uncertainties than incorrect ones, indicating that although not perfect, uncertainty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

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 falls within the margin of error.

6 Conclusion

We introduce DebUnc, a framework that integrates model uncertainty into multi-agent debates. By incorporating confidence levels, agents can assess the reliability of their peers and make more informed decisions when responses differ. Confidence is estimated using an uncertainty metric and conveyed through either attention scaling, which adjusts token weights based on confidence levels, or direct inclusion in prompts. Our results show that attention scaling outperforms textual prompts, highlighting that LLMs can process information beyond traditional text inputs. This technique is broadly applicable to scenarios where certain parts of the input should be prioritized, and we encourage further exploration of its potential.

7 Limitations

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.

Ethics Statement

Deploying large language models (LLMs) to the general public requires them to be trustworthy and reliable, particularly in high-risk scenarios. This concern gains more importance when LLMs interact within agent frameworks. We foresee a future where LLMs interact with other LLMs to achieve complex tasks. In this work, we take a step towards achieving reliable communication among agents by incorporating uncertainty estimations.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. Semantically diverse language generation for uncertainty estimation in language models. *ArXiv preprint*, abs/2406.04306.

AI@Meta. 2024. Llama 3 model card.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *ArXiv preprint*, abs/2305.13712.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv preprint*, abs/2308.07201.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *ArXiv preprint*, abs/2402.09733.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *Preprint*, arXiv:2307.01379.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *ArXiv preprint*, abs/2403.04696.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. Lm-polygraph: Uncertainty estimation for language models. *Preprint*, arXiv:2311.07383.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *ArXiv preprint*, abs/2402.00367.

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

Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. 2024. Towards uncertainty-aware language agent. *Preprint*, arXiv:2401.14016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023.

Metagpt: Meta programming for multi-agent collaborative framework. *ArXiv preprint*, abs/2308.00352.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv preprint*, abs/2302.09664.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *ArXiv preprint*, abs/2305.19118.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *ArXiv preprint*, abs/2401.15449.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *ArXiv preprint*, abs/2305.19187.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Daniel McDonald, Rachael Papadopoulos, and Leslie Benningfield. 2024. Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer, Zhaoran Wang, and Hongxia Yang. 2023. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. 2024. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. *arXiv e-prints*, pages arXiv–2406.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *ArXiv preprint*, abs/2405.20974.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *ArXiv preprint*, abs/2406.02543.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *ArXiv preprint*, abs/2310.01469.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *ArXiv preprint*, abs/2305.18153.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. Luq: Long-text uncertainty quantification for llms. *ArXiv preprint*, abs/2403.20279.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023. Exploring collaboration mechanisms for llm agents: A social psychology view. *ArXiv preprint*, abs/2310.02124.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024b. Tell your model where to attend: Post-hoc attention steering for llms. *Preprint*, arXiv:2311.02262.

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 Ground Truth 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.

<table>
<tr><th>Agent 1</th><th>Agent 2</th><th>Agent 3</th></tr>
<tr><td colspan="3">User, Round 1</td></tr>
<tr><td colspan="3">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</td></tr>
<tr><td colspan="3">Assistant, Round 1</td></tr>
<tr><td>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.

The features most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation are:

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.

Therefore, the correct answer is:

A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur

Answer: A</td><td>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 murmor; long mid-diastolic murmur

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.

Answer: C</td><td>The combination of a displaced apex beat, a pan-systolic murmor, 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

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.

Answer: C.</td></tr>
</table>

Agent 1 (cont.)	**Agent 2 (cont.)**	**Agent 3 (cont.)**
	User, Round 2	
These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 1): “‘<agent 2 response>“‘ One agent solution (confidence level is 1): “‘<agent 3 response>“‘ 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.	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 10): “‘<agent 1 response>“‘ One agent solution (confidence level is 1): “‘<agent 3 response>“‘ 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.	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 10): “‘<agent 1 response>“‘ One agent solution (confidence level is 1): “‘<agent 2 response>“‘ 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.
	Assistant, Round 2	
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: A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur. 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. Answer: A.	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. Answer: C.	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: A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur 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. Answer: A.

Agent 1 (cont.)	**Agent 2 (cont.)**	**Agent 3 (cont.)**
	User, Round 3	
These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 1): "'<agent 2 response>"' One agent solution (confidence level is 10): "'<agent 3 response>"' 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.	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 10): "'<agent 1 response>"' One agent solution (confidence level is 10): "'<agent 3 response>"' 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.	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is 10): "'<agent 1 response>"' One agent solution (confidence level is 1): "'<agent 2 response>"' 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.
	Assistant, Round 3	
Based on the discussions among different agents, the combination of clinical features most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation is: A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur. Explanation: Mixed mitral valve disease is defined as the presence of both mitral stenosis and mitral regurgitation. The clinical presentation of mixed mitral valve disease can vary depending on the severity and predominance of each component. A displaced apex beat, soft first heart sound, pan-systolic murmur, and long mid-diastolic murmur are more indicative of mitral regurgitation than mitral stenosis. The displaced apex beat can result from the increased volume load in the left ventricle due to mitral regurgitation. A soft or absent first heart sound may be heard in mitral regurgitation due to the decreased force of mitral valve closure, while the pan-systolic murmur and long mid-diastolic murmur are characteristic features of mitral regurgitation. 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. Answer: A.	Based on the consensus of other agents, the combination most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation is: A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur. Answer: A.	Based on the information provided by various agents and the consensus reached, the combination of clinical features most suggestive of mixed mitral valve disease with a predominance of mitral regurgitation is: A. Displaced apex beat; soft first heart sound; pan-systolic murmur; long mid-diastolic murmur Here's the explanation: The presence of a displaced apex beat and a pan-systolic murmur suggest mitral valve disease. However, the duration and timing of the diastolic murmur can help distinguish between mitral stenosis and mitral regurgitation. Mitral regurgitation is more commonly associated with a long, holodiastolic or mid-diastolic murmur, while mitral stenosis typically has a short, mid-diastolic murmur. In this scenario, the long mid-diastolic murmur, along with the displaced apex beat and soft first heart sound, makes mixed mitral valve disease with a predominance of mitral regurgitation the most likely diagnosis. Therefore, the answer is: Answer: A.

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
Ground Truth	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 4: 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 Ground Truth 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	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. <question>
	Initial, 5-shot	Answer the following multiple choice question. Examples: <5 example questions and answers> — YOUR TASK 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. <question>
	Debate	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is c): <agent response> One agent solution (confidence level is c): <agent response> 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.
GSM8k	Initial	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. <question>
	Debate	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is c): <agent response> One agent solution (confidence level is c): <agent response> Based off the opinion of other agents, can you provide an updated response? The original problem is: <question> 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.

Benchmark	Prompt Type	Prompt
Arithmetic	Initial	What is the result of <a+b*c+d>? State the final answer at the end of your response.
	Debate	These are solutions and confidence values from 1 to 10 (higher means more confident) to the problem from other agents: One agent solution (confidence level is c): <agent response> One agent solution (confidence level is c): <agent response> 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.
TruthfulQA	Initial	Answer the following multiple choice question: <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 <options>.
	Debate	These are the selections and confidence values from 1 to 10 (higher means more confident) from other agents: One agent solution (confidence level is c): <agent response> One agent solution (confidence level is c): <agent response> 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 <options>.