# A Survey of Uncertainty Estimation Methods on Large Language Models

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) have demonstrated remarkable capabilities across various tasks. However, these models could offer biased, hallucinated, or non-factual responses camouflaged by their fluency and realistic appearance. Uncertainty estimation is the key method to address this challenge. While research efforts in uncertainty estimation are ramping up, there is a lack of *comprehensive* and *dedicated* surveys on LLM uncertainty estimation. This survey presents four major avenues of LLM uncertainty estimation. Furthermore, we perform extensive experimental evaluations across multiple methods and datasets. At last, we provide critical and promising future directions for LLM uncertainty estimation.


## 1 Introduction

Large Language Models (LLMs) have emerged as state-of-the-art solutions for a wide range of problems, mainly due to their unparalleled ability to generate coherent and contextually appropriate responses to diverse user prompts (Ouyang et al., 2022; Zhao et al., 2024). However, with the increasing adoption of LLMs, concerns have grown regarding their tendency to produce biased, hallucinated, non-factual, and misaligned outputs (Zhang et al., 2023; Huang et al., 2024b). These issues are further exacerbated by the fact that such flawed responses often appear highly fluent and convincingly realistic, making them difficult to detect.

A promising approach to addressing the challenge of misleading yet plausible responses is uncertainty estimation, which assigns an uncertainty or confidence score to the model's output. Figure 1 provides an overview of this process. First, the LLM generates an initial response based on the input. Next, a confidence score is computed for this response. The score is then evaluated against a predefined threshold to determine the final output.

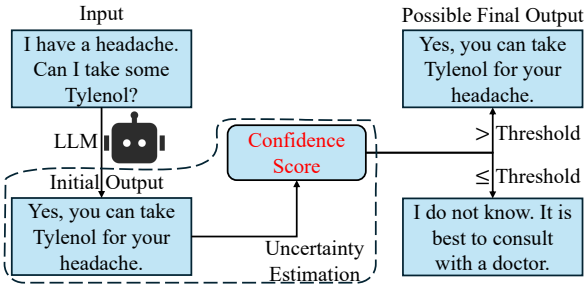


Figure 1: Illustration of uncertainty estimation.

If the confidence score meets or exceeds the threshold, the initial response is accepted; otherwise, the model outputs "I do not know," thereby reducing the risk of providing incorrect but convincingly realistic information to users.

There is an urgent need for a comprehensive survey on LLM uncertainty estimation. Below, we highlight three of them: (i) Although uncertainty estimation has been extensively studied in traditional deep neural networks (DNNs)—with Bayesian and ensemble methods being notable examples (Gawlikowski et al., 2023))—these techniques are not easily transferable to LLMs, due to the large number of parameters in LLMs. (ii) LLMs significantly transform society, creating a strong demand for a thorough study of uncertainty estimation tailored to LLMs. A survey of recent advances in LLM uncertainty estimation would provide a solid foundation for future development in the field. (iii) While there are three existing surveys on LLM uncertainty estimation, each has notable limitations. Specifically, (Huang et al., 2024a) dedicates a substantial portion of its content to traditional DNN uncertainty estimation rather than focusing on LLMs. (Geng et al., 2024) shifts its attention to uncertainty calibration and the applications of LLM uncertainty estimation, rather than providing a deep exploration of the core techniques. Similarly, (Shorinwa et al., 2024) devotes much of its content to benchmarks and applications while lacking a complete view of the uncertainty estimation methods on LLMs.

This manuscript focuses on studying the uncertainty estimation methods within the context of LLMs, introducing a new taxonomy from the perspective of LLMs. We center our scope around techniques applicable during the inference stage. We emphasize the methods that do not require additional data (Ren et al., 2023; Kumar et al., 2023; Tonolini et al., 2024) or model modifications (Huang et al., 2023a; Liu et al., 2024), ensuring the broad applicability of this survey. Moreover, this is, to the best of our knowledge, the first survey that conducts a thorough evaluation of representative uncertainty estimation approaches across various datasets and domains. Built on the insights from our evaluations, we postulate two interesting future directions for LLM uncertainty estimation.

## 2 Uncertainty Sources in LLM

There are two primary sources of uncertainty: aleatoric and epistemic uncertainties (Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021). In the context of LLMs (Gao et al., 2024; Ahdritz et al., 2024; Hou et al., 2024), these sources manifest in the following ways:

- **Aleatoric uncertainty** refers to the uncertainty inherent in the data. For LLMs, this arises from ambiguous or incomplete information and inherent properties of natural language itself. Examples include vague or contextually dependent prompts, as well as linguistic phenomena where multiple valid interpretations or responses naturally coexist.
- **Epistemic uncertainty** reflects the model's lack of knowledge or understanding. In LLMs, this occurs when the model encounters unfamiliar concepts or data that are underrepresented in its training set. This type of uncertainty can potentially be reduced by improving the training datasets and models.

## 3 Uncertainty Estimation in LLMs

### 3.1 Problem Definition and Overview

**Token generation in LLMs.** LLMs output responses in an auto-regressive manner, predicting the probability distribution of the next token given the prompt and the previously generated tokens. We denote the model as $\boldsymbol{f}$, the prompt as $\boldsymbol{x}$, and the generated response (or the answer) as $\boldsymbol{r}$, which consists of $N$ tokens, denoted as $\{z_1, z_2, z_3, \cdots, z_N\}$. The tokens can be either words, subwords, or characters from a predefined vocabulary $\boldsymbol{Z}$. At each step of token generation, the model computes the conditional probability distribution over the vocabulary for the next token, based on the prompt $\boldsymbol{x}$ and all previously generated tokens $\boldsymbol{r_{<i}} = \{z_1, z_2, \cdots, z_{i-1}\}$. The probability distribution for the $i$-th token is given by $\boldsymbol{p_i} = \text{Softmax}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{r_{<i}}))$. Here, $\boldsymbol{p_i}$ is a vector of length $|\boldsymbol{Z}|$, with each entry representing the probability of a specific token in $\boldsymbol{Z}$ being chosen as the next token. It allows strategies such as sampling or beam search to choose from these token candidates according to their probabilities. Such an auto-regressive process ends when # of generated tokens reaches a preset number or LLM generates the end-of-sequence (EOS) token.

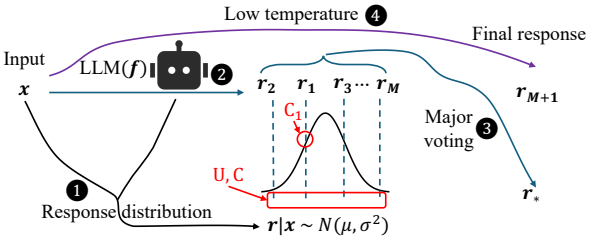


Figure 2: Illustration of uncertainty versus confidence.

It is important to note that uncertainty is the innate nature of LLMs, regardless of whether we estimate it. Now, we provide an intuitive understanding of uncertainty and how to estimate it.

**How to estimate uncertainty and confidence?** As shown in Figure 2, for each input $\boldsymbol{x}$, an LLM model has an underlying response distribution for it (❶). For ease of illustration, we assume the distribution is a normal distribution $N(\mu, \sigma^2)$. Uncertainty estimation is to estimate the underlying variance $\sigma^2$. For example, the sample variance of $M$ different responses $\boldsymbol{r_1}, \cdots, \boldsymbol{r_M}$ (❷) can be an estimator for the variance, which indicates the variations of responses ($U$ in Figure 2).

There generally are two types of confidence, i.e., overall confidence $C$ and the confidence $C_i$ associated with each response candidate $\boldsymbol{r_i}$. The overall confidence $C$ is complementary to $U$, i.e., the precision $1/\sigma^2$ of the distribution is a confidence $C$ to the input. The associated confidence is related to $\boldsymbol{x}$ and the tokens in a specific response $\boldsymbol{r_i}$. To provide the final response to answer the input $\boldsymbol{x}$ given sampled responses, some literature resort to majority voting to select the most-voted response $\boldsymbol{r_*}$ (❸) (Wang et al., 2023), while others choose to generate one extra response $\boldsymbol{r_{M+1}}$ with low-temperature settings (❹) as (Farquhar et al., 2024).

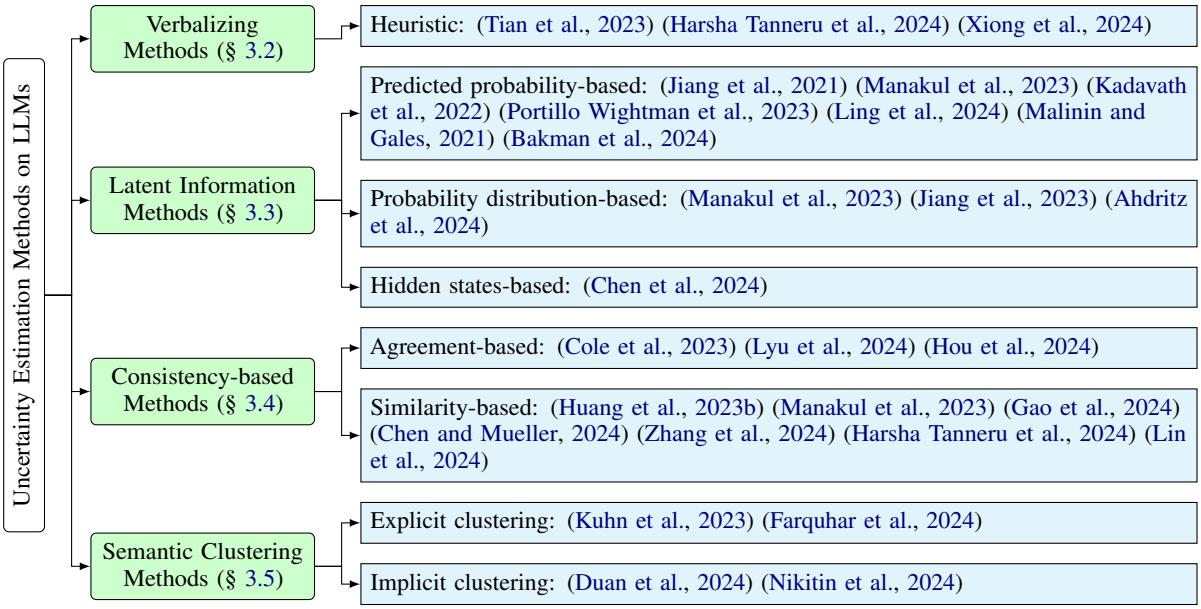


Figure 3: Taxomony of uncertainty estimation methods on LLMs.

**Surveyed papers overview.** Figure 3 categorizes all the uncertainty estimation papers for LLM into four classes: verbalizing methods, latent information methods, consistency-based methods, and semantic clustering methods. We review each through Sections 3.2 - 3.5.

### 3.2 Verbalizing Methods

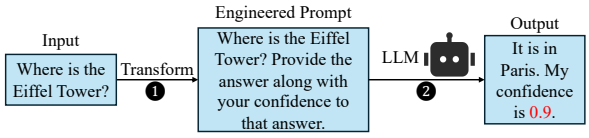


Figure 4: Illustration on verbalizing methods.

Figure 4 demonstrates the main workflow of verbalizing methods. Firstly, the input is transformed into an engineered prompt that explicitly asks the model to provide both an answer and its confidence level (❶). Secondly, the LLM processes this prompt and generates an output that includes the answer and a verbalized confidence score (❷), representing its self-assessed certainty about the correctness of its response.

(Lin et al., 2022a) pioneer this cohort of efforts. As the capabilities of LLMs continue to develop, they can provide reasonable confidence under proper guidance, even without fine-tuning. Subsequently, (Tian et al., 2023) proposes three verbalizing variants: (i) Generate multiple response candidates with confidence scores and select the highest-rated one as the final response, (ii) derive the response and confidence through two rounds of prompt-and-answer interactions, and (iii) use words instead of numerical values to indicate the confidence. Recently, (Harsha Tanneru et al., 2024) introduces two methods inspired by Chain-of-Thought (CoT) prompting. The first method requests the LLM to assign an importance score to each word in the input, while the second one prompts the LLM to provide confidence for each reasoning step in the response. Finally, LLM will offer a final confidence score for the overall response. Beyond that, (Xiong et al., 2024) presents a systematic framework for verbalizing methods with three parts: prompting, sampling, and aggregation. It employs specific confidence-eliciting prompts and generates diverse response samples containing confidence scores. After that, the final confidence score is derived through inter-sample agreement or response ranking information.

While verbalizing methods offer intuitive and straightforward uncertainty estimation, they face significant limitations. (Kadavath et al., 2022) shows that LLMs tend to be over-confident in their answers as the reinforcement learning from human feedback (RLHF) nature pushes LLMs to do so.

### 3.3 Latent Information Methods

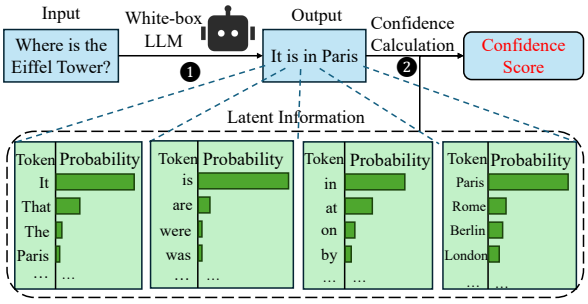


Figure 5: Illustration on latent information methods.

Figure 5 illustrates the concept of latent infor-

mation methods. First of all, the LLM is prompted to provide an output to the input (❶). Of note, latent information methods require a white-box LLM, which offers latent information in the output, such as the full probability distribution over each generated token. Subsequently, this method leverages the generated information to estimate the uncertainty/confidence score via specific metrics or measures (❷). We refer the readers to Section A.2 for the formula of different latent information methods.

(Jiang et al., 2021) directly uses the predicted probability of the response tokens to measure the confidence score. (Manakul et al., 2023) proposes to use the negative log-likelihood of the response tokens, either average or maximum across tokens, to serve as an uncertainty measure. The averaged negative log-likelihood across tokens is also known as perplexity (Ren et al., 2023). In contrast, (Kadavath et al., 2022) proposes a method that prompts the model to evaluate its answers by answering true or false, using the latent probability associated with "True" as the confidence score.

The analysis of token probabilities can be extended beyond a single response for more robust uncertainty estimation. (Portillo Wightman et al., 2023) proposes to average the predicted probabilities across multiple responses. (Ling et al., 2024) picks the key token from the responses and aggregates them into a distribution, and the uncertainty is from the entropy of the distribution. (Kadavath et al., 2022) considers all the tokens in the responses, calculates the probability for each response using token probabilities, and measures uncertainty through the entropy of the response distribution, called predictive entropy. However, varying response lengths can introduce undesirable noise to the estimation. To address this limitation, (Malinin and Gales, 2021) proposes the length normalized entropy, incorporating the response length based on predictive entropy. Furthermore, (Bakman et al., 2024) proposes to replace the length normalization by assigning a weight to each token with a BERT model to consider both the sequence length and semantic contribution of tokens.

While the methods above only require access to the probability value of the response tokens, the following papers would require access to the complete probability distributions: (Manakul et al., 2023) computes the entropy of the probability distribution for each generated token, using either the mean or maximum entropy as the uncertainty. For multiple-choice questions, (Jiang et al., 2023) presents a specialized methodology. It computes probability distributions over potential options for each response sample and aggregates these distributions to form an ensemble probability distribution for uncertainty estimation. (Ahdritz et al., 2024) introduces a heuristic two-stage method. Initially, the LLM is prompted to generate multiple next-token candidates. Subsequently, through a "repeated prompt" mechanism, the model produces the next token. The final uncertainty score is then computed from the probability distribution of these next tokens.

Beyond the methods using the probability distributions of tokens in the response, some researchers utilize the hidden states of LLMs. (Chen et al., 2024) proposes to use the embeddings in the middle layer of LLMs to construct a covariance matrix for responses, which captures the correlation relationships among them. By manipulating the eigenvalues of the covariance matrix, the degree of divergence among responses can be estimated and considered an uncertainty measure.

### 3.4 Consistency-based Methods

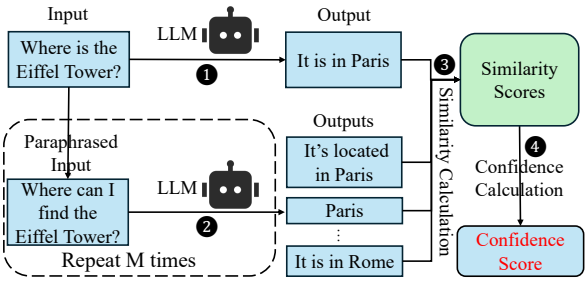


Figure 6: Illustration of consistency-based methods.

Figure 6 illustrates the workflow of consistency-based methods. First, LLM gives an output to the original input (❶). Second, the input is paraphrased to maintain the same meaning as the original one but has different contents, where LLM is prompted to answer this changed input. Such process is repeated $M$ times to generate various sampled outputs (❷). Third, the similarities between the original output and each sampled output are computed (❸). Finally, the confidence score is calculated based on derived similarities (❹). We refer the readers to Section A.3 for detailed mathematical definitions of this consistency-based method.

The fundamental principle of consistency-based methods is that response consistency typically correlates with confidence levels, a.k.a. high response variability suggests higher uncertainty, while consistent responses indicate greater confidence.

(Cole et al., 2023) introduces sampling diversity and sampling repetition. Sampling diversity quanti-

fies the ratio of unique answers to the total number of samples, while sampling repetition measures the proportion of samples that align with the most frequent answer. Extending this framework, (Lyu et al., 2024) enhances the sampling repetition metric by incorporating the most frequent and second-most frequent responses in its analysis. (Hou et al., 2024) presents a more nuanced approach by introducing clarification-based uncertainty estimation. It first generates multiple clarifications for the input and then produces responses based on these clarified inputs. The estimated uncertainty combines two parts: one from answer frequency distribution and the other from input clarification variance.

While the methods primarily focus on analyzing answer agreement patterns to estimate uncertainty, more methods emphasize evaluating the similarities among responses (❸). For domain-specific tasks, targeted metrics like BLEU (Papineni et al., 2002) and CodeBLEU (Ren et al., 2020) have been successfully applied to machine translation and code generation tasks, respectively (Huang et al., 2023b). In general question-answering scenarios, token-level similarity metrics such as BERTScore (Zhang et al., 2020) and RougeL (Lin, 2004) have been widely adopted (Huang et al., 2023b; Manakul et al., 2023; Gao et al., 2024). Moving beyond token-level comparisons, more sophisticated approaches that capture semantic relationships have emerged, including SentenceBERT and NLI-based methods (Gao et al., 2024; Chen and Mueller, 2024; Zhang et al., 2024). SentenceBERT computes the cosine similarity between two sentences using embeddings generated by the Sentence Transformer model. The NLI-based method leverages natural language inference (NLI) classifiers to categorize sentence relationships as entailment, neutral, or contradiction, regarding the probability the NLI classifier assigns to the "entailment" class as the similarity score. Moreover, (Harsha Tanneru et al., 2024) proposes token importance uncertainty and CoT uncertainty. The former quantifies uncertainty through token agreement and token rank metrics, while the latter evaluates inter-step relationships using NLI classification techniques.

The generation of diverse LLM outputs in step ❷ represents another critical avenue for enhancing consistency-based methods. (Harsha Tanneru et al., 2024) presents two fundamental approaches: sample probing, which employs semantically equivalent prompts, and model probing, which manipulates temperature settings to introduce output stochasticity. (Chen and Mueller, 2024) introduces a method that modifies CoT steps specifically for prompts employing CoT techniques. Additional approaches have been proposed by (Gao et al., 2024), including the strategic insertion of dummy tokens (such as newline characters and tab spaces) and modifications to system messages within prompts.

While most methods estimate confidence by simply averaging similarities among responses in step (❹), (Lin et al., 2024) proposes a new similarity-based method for calculating confidence inspired by spectral clustering. It treats generated responses as nodes and obtains the degree matrix and the graph Laplacian matrix. Correspondingly, this method defines several uncertainty and confidence measures from the matrices.

### 3.5 Semantic Clustering Methods

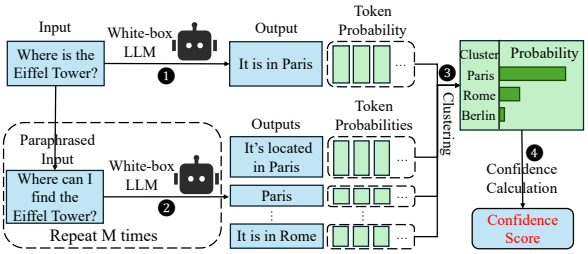


Figure 7: Illustration on semantic clustering methods.

Figure 7 depicts the workflow of semantic clustering methods, which leverages both the latent information and the semantic relationships among responses to offer a more comprehensive estimation of the uncertainty. The first two steps are similar to the consistency-based methods, where the LLM generates responses to the original input and its paraphrased versions (❶-❷). Next, instead of calculating the similarities, the sampled outputs are partitioned into clusters with a new probability for each cluster (❸). Finally, the probability distribution over these clusters calculates a confidence score (❹). The motivation behind semantic clustering methods is that consistency-based methods regard two responses as consistent only if they have identical words, which is too strict. Therefore, semantic clustering of the latent information is proposed to deal with the limitations. We refer the readers to Section A.4 for the formula of different semantic clustering methods.

(Kuhn et al., 2023) introduces semantic entropy for uncertainty estimation. The method comprises three phases: generation (❶-❷), clustering (❸), and entropy estimation (❹). In step ❸, a bi-directional entailment algorithm is employed to

cluster semantically equivalent responses. It assesses the entailment relationship between each pair of responses, considering them to express the same meaning if they mutually entail each other. The entailment relationship can be determined with the help of an NLI classifier or by simply requesting a general-purpose LLM. The uncertainty is the entropy calculated from the cluster probabilities in step ❹. In case there is no access to the token probability, (Farquhar et al., 2024) introduces discrete semantic entropy, which leverages the response frequency to derive the aggregated probabilities.

While the above method clusters the responses explicitly, some methods propose implicit clustering. Instead of utilizing the bi-directional entailment algorithm, (Duan et al., 2024) introduces sentence relevance scores between each response pair, which is more effective over long sentences than the bi-directional entail algorithm. (Nikitin et al., 2024) further considers the distances between the clusters. The method encodes similarities among responses via positive semidefinite unit trace kernels. It offers a more fine-grained uncertainty measure using the von Neumann entropy of these kernels.

# 4 Evaluation

## 4.1 Metrics

We use two primary metrics to evaluate the uncertainty estimation: **AUROC** (Area Under the Receiver Operating Characteristics curve) (Bradley, 1997) and **AUARC** (Area Under the Accuracy-Rejection Curve) (Nadeem et al., 2009). Both metrics range from 0 to 1, with higher scores reflecting better uncertainty estimation methods. § B contains more details about AUROC and AUARC.

## 4.2 Evaluated Methods

We select several representative methods from each method category as follows:

- *Verbalizing methods (Verb):* We evaluate the **2S** (Tian et al., 2023) emthod that asks for confidence in a second-round dialogue.
- *Latent information methods (Latent):* We select the self-evaluation method (**Ptrue**) (Kadavath et al., 2022), perplexity (**Perp**), predictive entropy (**PE**), length-normalized entropy (**LN-E**), and the method leveraging hidden states of LLMs (**INSIDE**) (Chen et al., 2024).
- *Consistency-based methods (Consis):* We adopt four similarity measures: **BERTScore**, **RoughL**, cosine similarity from BERT embeddings (**Cosine**), and the "entailment" probability from an NLI classifier (**NLI**). The confidence score is averaged from similarities.
- *Semantic clustering methods (Cluster):* We include semantic entropy (**SE**) and discrete semantic entropy (**DSE**).

## 4.3 Model Settings

We use **LLaMA3.1-8B-Instruct** (Llama Team, 2024) in our experiments. Following (Farquhar et al., 2024), we first set the temperature = 0.1 and generate an answer as the final answer. Then, we set the temperature to be 1 and generate 20 answers, which are used for methods that need extra samples. We employ the multinomial sampling as the decoding strategy and set top_k equal to 50. Due to the varying types of questions and domains, we used the same model to determine the correctness of an answer. The prompts used are in § C.

## 4.4 Illustrative Results

Figures 8 - 12 show the ROC and ARC with the corresponding AUROC and AUARC values in the legend for five different datasets (Details about the datasets are in § D). For the AUROC and AUARC values from the legend, we color-coded the "best", "2nd best", "3rd best", "3rd worst", "2nd worst", and "worst".

**TruthfulQA** (Figure 8) is a benchmark designed to evaluate the truthfulness of language models in answering questions spanning 38 categories (Lin et al., 2022b). The questions in the dataset appear in a multiple-choice form, providing the LLM with clear guidance and ensuring a fixed response format. Therefore, most of the uncertainty is epistemic uncertainty. In the ROC curve, **Perp** and **INSIDE** (❶) demonstrate the lowest performance, close to random guessing. The ROC curve of **2S** (❷) starts with the steepest rise, indicating most responses assigned with high confidence are correct. In the ARC curve, the worst-performing method (❶) shows no improvement in accuracy as the rejection rate increases until the rejection rate is high. Although **2S** (❷) shows a slower initial improvement, it enjoys higher improvements afterward, again demonstrating its high accuracy for high-confidence answers. **2S** achieves the best performance on this dataset, showing that LLMs can tell their uncertainty, especially when this is mainly epistemic uncertainty.

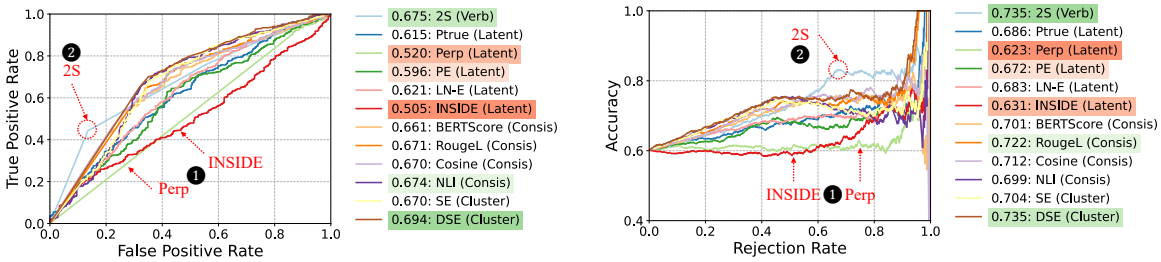


Figure 8: **TruthfulQA**: ROC (left), ARC (right) curves, and AUROC and AUARC.

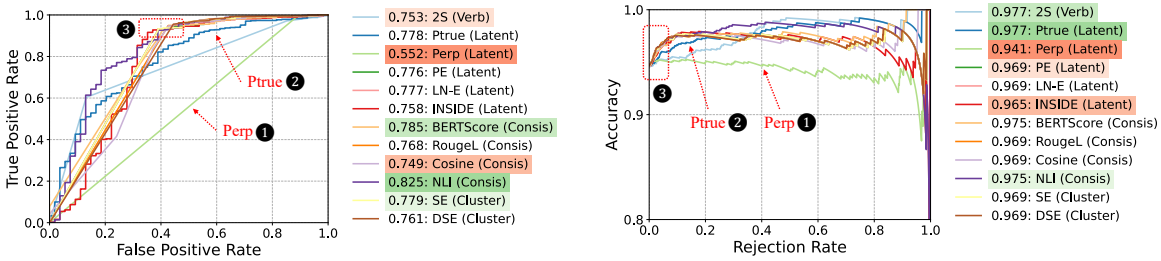


Figure 9: **SciQ**: ROC (left), ARC (right) curves, and AUROC and AUARC.

**SciQ** (Figure 9) is another multiple-choice Q&A dataset, with a collection of science-focused questions (Welbl et al., 2017). In the ROC curve, **Perp** (❶) performs like random guessing (analogous to **TruthfulQA**), whereas all other methods achieve significantly better performance, including **Ptrue** (❷). Most of the methods (❸) achieve a very high True Positive Rate (TPR) when the False Positive Rate (FPR) approaches 0.4, indicating they assign most of the low confidence scores to negative samples correctly. As for AUARC, most methods exhibit similar performance, as the dataset is considered simple for the LLM, evidenced by a high initial accuracy of about 0.95 (❸). However, the accuracy of **Perp** (❶) decreases from the very beginning, resulting in the worst AUARC. In contrast, **Ptrue** (❷), another variant of the latent information-based method, gains better accuracy with higher rejection rates. The difference between **Perp** and **Ptrue** shows the aggregated predicted probability of tokens is not well-calibrated, but the probability of answering the true/false of the entire response is well-calibrated.

**TriviaQA** (Figure 10) is a reading comprehension dataset where no context is provided in our settings (Joshi et al., 2017). As a free-form Q&A dataset, it allows responses to a question to vary while still expressing the same meaning. Therefore, the aleatoric uncertainty caused by language ambiguity in questions and responses exists. The ROC curve reveals that **2S** and **Perp** (❶) demonstrate relatively poor performance. In contrast, **NLI** (❷) achieves the highest performance. In the ARC curve, the accuracy of **2S** and **Perp** (❶) deteriorates as the rejection rate increases from 0.5, while **DSE** (❷) achieves the highest AUARC score.

**GSM8K** (Figure 11) is comprised of math problems that need reasoning steps to solve (Cobbe et al., 2021). The responses thus can be more diverse than **TriviaQA** due to the variability in reasoning steps. Hence, the aleatoric uncertainty is even higher. The results on AUROC demonstrate that **INSIDE** (❶) performs below random guessing. On the contrary, **NLI**, **DSE**, and **SE** (❷) maintain more gains on TPR with the increase of FPR. A noteworthy observation is that **NLI** and **SE** (❸) achieve positive TPR even when FPR = 0 because they perfectly classify the high-confidence responses. In the ARC curve, this phenomenon is once again reflected that these methods achieve perfect accuracy when considering only the top 20% high-confidence responses (❸). From the point where the rejection rate is 0, better methods exhibit faster rates of improvement (❷), while the worst one (i.e., **INSIDE**) has a negative rate (❶).

Comparing the **TriviaQA** and **GSM8K** datasets, **NLI**, **SE**, and **DSE** perform the best on the free-form questions. They all consider the entailment relationship among responses, which can tremendously eliminate the aleatoric uncertainty and thus better estimate epistemic uncertainty. By doing so, they obtain better final results.

**SimpleQA** (Figure 12) is a recent Q&A dataset that presents significant challenges for state-of-the-art LLM models as of 2024 (Wei et al., 2024). Interestingly, in the ROC curve, methods (❶) that traditionally demonstrate superior performance on other datasets exhibit notably poor outcomes here.

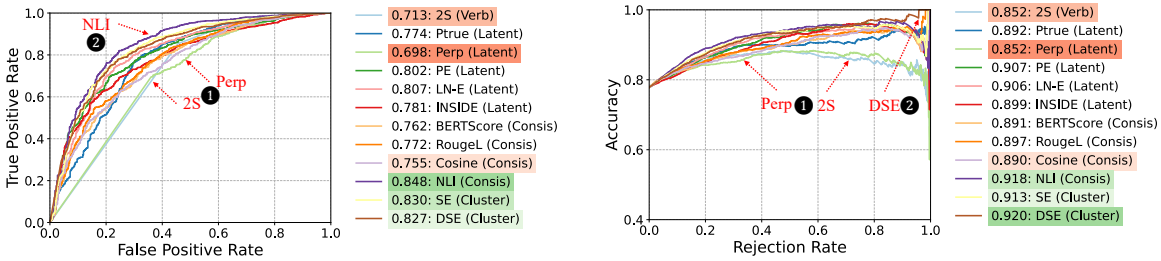


Figure 10: **TriviaQA**: ROC (left), ARC (right) curves, and AUROC and AUARC.

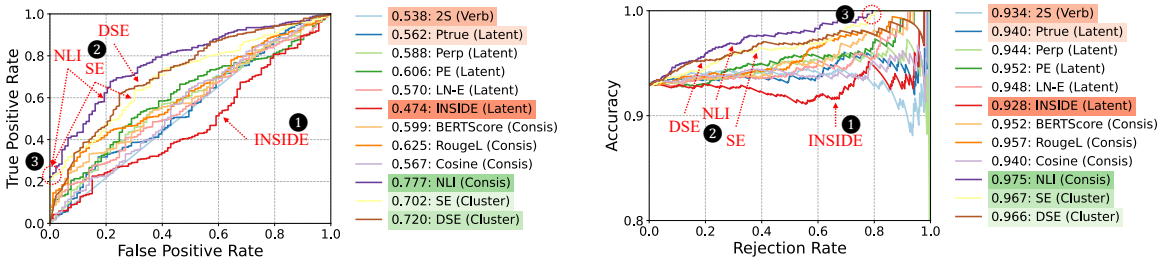


Figure 11: **GSM8K**: ROC (left), ARC (right) curves, and AUROC and AUARC.

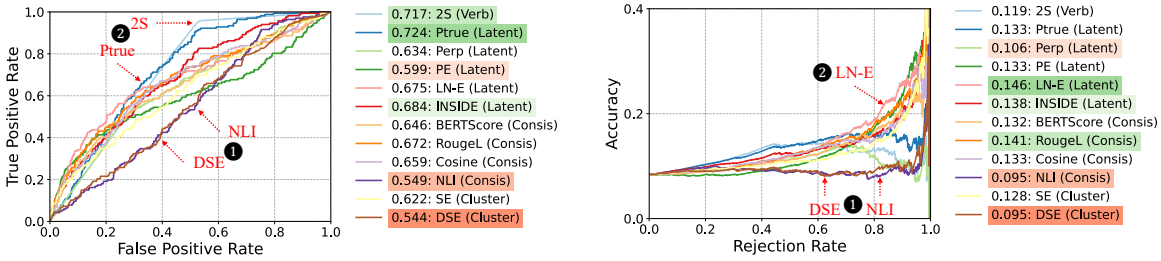


Figure 12: **SimpleQA**: ROC (left), ARC (right) curves, and AUROC and AUARC.

**2S** and **Ptrue** (❷) emerge as the top performers, distinguished by their ability to maintain low FPR while TPR approaches 1. In the ARC curve, there is no accuracy improvement for **NLI** and **DSE** as the rejection rate increases (❶). Notably, **LN-E** (❷) becomes the highest because its accuracy continues to grow after the rejection rate passes 0.8, while others drop. Although **SimpleQA** is still a free-form dataset, **NLI**, **SE**, and **DSE** do not show their superior performance here. It shows they cannot estimate the epistemic uncertainty well if it is too big, postulating whether current benchmarks adequately evaluate LLM uncertainty estimation.

## 5 Future Directions

**Uncertainty estimation benchmark.** We need a dataset specifically designed for uncertainty estimations on LLMs. Existing datasets are designed to evaluate the capability of LLMs (not their uncertainty). They always have unambiguous questions, resulting in low aleatoric uncertainty. We anticipate three rules for designing this dataset: First, it should incorporate a diverse set of question types, including general Q&A problems, math problems, translation problems, etc. Second, the questions should have varying difficulty levels, from simple to extremely challenging. Finally, the dataset can control the degree of ambiguity for the questions to directly evaluate the uncertainty.

**Uncertainty estimation method enhancement.** Uncertainty estimation for long responses remains under-explored. While some papers propose to break long responses into shorter segments and process each part individually (Zhang et al., 2024; Farquhar et al., 2024), they ignore the inter-sentence relationships that are critical for capturing the overall uncertainty of the response. Further, the large vocabulary in long responses challenges the effectiveness of consistency-based and semantic clustering methods. Current uncertainty estimation methods, predominantly validated on short-answer scenarios, may not adequately address the complexities inherent in longer, multi-step reasoning processes.

## 6 Conclusion

This survey paints a comprehensive landscape for uncertainty estimation methods on LLMs during the inference stage, classifying them into four classes: verbalizing, latent information, consistency-based, and semantic clustering methods. We further enrich our survey with extensive evaluations and promising future directions.

## 7 Limitations

This survey contains three limitations, mainly due to space constraints. First, we omitted detailed methodological explanations for various methods from the main text. Second, we did not evaluate and report the results of all the introduced methods. Finally, we exclude the literature that does not surround the inference stage of LLMs. We acknowledge these limitations and remain open to academic discussion and collaborative efforts to address them in future work.

## References


Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. Distinguishing the knowable from the unknowable with language models. In *Forty-first International Conference on Machine Learning*.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, Bangkok, Thailand. Association for Computational Linguistics.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

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

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

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

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Forty-first International Conference on Machine Learning*.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024a. A survey of uncertainty estimation in llms: Theory meets practice. *Preprint*, arXiv:2410.15326.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *Preprint*, arXiv:2307.10236.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles. *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *Preprint*, arXiv:2305.18404.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

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. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *Preprint*, arXiv:2404.15993.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *Preprint*, arXiv:2402.13904.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pages 65–81, Ljubljana, Slovenia. PMLR.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pages 8901–8929. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *Preprint*, arXiv:2009.10297.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Preprint*, arXiv:2412.05563.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12229–12272, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

# A Mathematical Formulation of the Methods

## A.1 Common notations

We list some common notations in Table 1 for mathematical definitions.

| Notation | Description |
|---|---|
| $\boldsymbol{f}$ | Large language model |
| $\boldsymbol{x}$ | Input |
| $\boldsymbol{r_i}$ | The $i$-th sampled response |
| $\boldsymbol{r_*}$ | The most-voted response from samples |
| $N$ | Number of tokens in a response |
| $M$ | Number of sampled responses |
| $\boldsymbol{R_i}$ | The $i$-th response cluster |
| $K$ | Number of response clusters |
| $z_i$ | the $i$-th token in a response |
| $\boldsymbol{Z}$ | Vocabulary of the large language model |
| $\boldsymbol{r_{<i}}$ | All tokens before the $i$-th token |
| $\boldsymbol{p_i}$ | The probability distribution for the $i$-th token |
| $p_{z_i}$ | The probability for the token $z_i$ |
| $p$ | The probability of something |
| $a(r_i, r_j)$ | Similarity score between $r_i$ and $r_j$ |
| $U$ | Estimated uncertainty |
| $C$ | Estimated overall confidence score |
| $C_i$ | Estimated confidence score for response $\boldsymbol{r_i}$ |

Table 1: Common notations and descriptions.

### A.2 Latent Information Methods

**Average over negative logarithm likelihood (Perplexity)** (Manakul et al., 2023; Ren et al., 2023):

$$U = -\frac{1}{|\boldsymbol{r}|}\sum_{i=1}^{N}\log p_{z_i}$$

**Maximum over negative logarithm likelihood** (Manakul et al., 2023):

$$U = \max_i(-\log p_{z_i}), \quad i \in [1, N]$$

**Ptrue** (Kadavath et al., 2022):

$$C = p(z_{true}|\boldsymbol{x'}),$$

where $z_{true}$ is the token for "true", and $x'$ is the designed prompt to ask LLM to decide whether the answer is true or false.

**Predictive entropy** (Kadavath et al., 2022):

$$U = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{|\boldsymbol{r_j}|}\log p_{z_i}$$

**Length-normalized entropy** (Malinin and Gales, 2021):

$$U = -\frac{1}{M}\sum_{j}^{M}\frac{1}{|\boldsymbol{r_j}|}\sum_{i=1}^{|\boldsymbol{r_j}|}\log p_{z_i}$$

**Average over tokens' probability distributions** (Manakul et al., 2023; Ren et al., 2023):

$$U = -\frac{1}{|\boldsymbol{r}|}\sum_{i=1}^{N}\sum \boldsymbol{p_i} \circ \log \boldsymbol{p_i},$$

where $\circ$ is the element-wise multiplication, and the second $\sum$ means sum over all the elements in a vector.

**Maximum over tokens' probability distributions** (Manakul et al., 2023):

$$U = \max_i(-\sum \boldsymbol{p_i} \circ \log \boldsymbol{p_i}), \quad i \in [1, N],$$

where $\circ$ is the element-wise multiplication, and $\sum$ means sum over all the elements in a vector.

**INSIDE** (Chen et al., 2024):

$$U = \frac{1}{N}\log\det(\boldsymbol{\Sigma} + \alpha\boldsymbol{I}) = \frac{1}{N}\sum_{i=1}^{N}\log(\lambda_i),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix, $\alpha$ is a small regularization term, $\boldsymbol{I}$ is an identity matrix, and $\lambda_i$ is the $i$-th eigenvalue of the matrix $\boldsymbol{\Sigma} + \alpha\boldsymbol{I}$. Specifically,

$$\boldsymbol{\Sigma} = \boldsymbol{V} \cdot \boldsymbol{J}_d \cdot \boldsymbol{V}, \quad \boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_N}],$$

where $\boldsymbol{v_i}$ is the representative embedding for $r_i$, $\boldsymbol{J}_d = \boldsymbol{I}_d - \frac{1}{d}\boldsymbol{1}_N\boldsymbol{1}_N^T$ represents the centering matrix, and $d$ corresponds to the dimension of the embeddings.

### A.3 Consistency-based Methods

**Sampling diversity** (Cole et al., 2023):

$$C = 1 - \frac{K}{M}$$

**Sampling diversity** (Cole et al., 2023):

$$C = \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}(\boldsymbol{r_i} = \boldsymbol{r_*}),$$

where $\mathbb{1}()$ is the indicator function.

**First-second-distance-based (FSD) method** (Lyu et al., 2024):

$$C = \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}(\boldsymbol{r_i} = \boldsymbol{r_*}) - \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}(\boldsymbol{r_i} = \boldsymbol{r_{**}}),$$

where $\mathbb{1}()$ is the indicator function, and $\boldsymbol{r_{**}}$ denotes the second most-voted answer.

**Variation ratio (VR) (when the final response is $\boldsymbol{r_*}$)** (Huang et al., 2023b):

$$U = 1 - \frac{\sum_{i=1}^{M}\frac{\sum_{j=1, j\neq i}^{j=M} a(\boldsymbol{r_i}, \boldsymbol{r_j})}{M-1}}{M}$$

**Variation ratio (VR) (when the final response is $\boldsymbol{r_{M+1}}$)** (Huang et al., 2023b):

$$U = 1 - \frac{\sum_{i=1}^{M} a(\boldsymbol{r_i}, \boldsymbol{r_{M+1}})}{M}$$

Based on **VR** and **VRO**, using different similarity calculation methods for $a(\cdot, \cdot)$ can achieve different estimated uncertainty.

### A.4 Semantic Clustering Methods

**Semantic entropy** (Kuhn et al., 2023):

$$U = -\sum_{k=1}^{K} p(\boldsymbol{R_k}) \log p(\boldsymbol{R_k}),$$

where

$$p(\boldsymbol{R_k}) = \sum_{\boldsymbol{r_j} \in \boldsymbol{R_k}} \exp(\frac{1}{|\boldsymbol{r_j}|} \sum_{i=1}^{|\boldsymbol{r_j}|} \log p_{z_i})$$

**Discrete semantic entropy** (Farquhar et al., 2024):

$$U = -\sum_{k=1}^{K} p(\boldsymbol{R_k}) \log p(\boldsymbol{R_k}),$$

where

$$p(\boldsymbol{R_k}) = |\boldsymbol{R_k}|/K$$

## B Detailed Explanation of AUROC and AUARC

**AUROC**: For each response, we consider it as a positive sample (correct) or a negative sample (incorrect) based on whether it matches the ground-truth label. The ROC curve is then created by plotting the true positive rate (TPR) against the false positive rate (FPR). To derive TPRs and FPRs, the accepted confidence threshold is changed to get different Predicted Positives and Negatives (i.e., PP and PN), where a response with confidence higher than the threshold is regarded as PP or PN otherwise. The AUROC is the area under the ROC curve, measuring the discriminability of confidence scores to distinguish between correct and false responses.

**AUARC**: Accuracy-Rejection Curve (RAC) is specifically designed for uncertainty estimation, which plots how the accuracy on the accepted samples changes as more low-confidence answers are rejected. The area under it indicates the uncertainty estimation's ability to maintain high accuracy when low-confidence answers are rejected.

## C Prompts

The prompt for Q&A questions is as follows:

---

**System:**
You are a highly knowledgeable assistant. Answer the following question as briefly as possible.
**... (several few-shot examples)**
**User:**
[Question]

---

The prompt for correctness decisions is as follows:

---

**User:**
We are assessing the quality of answers to the following question: [Question]
The expected answer is: [Gt_answer]
The proposed answer is: [Predicted_answer]
Within the context of the question, does the proposed answer mean the same as the expected answer? Respond only with yes or no.
Response:

---

## D Detailed Explanation of Dataset

We give a sample for each dataset in Table 2.

**TruthfulQA** (Lin et al., 2022b) is a benchmark designed to evaluate the truthfulness of language models in generating answers to questions. It consists of 817 questions spanning 38 diverse categories such as health, law, finance, and politics. The dataset is intentionally crafted with questions that humans may answer falsely due to misconceptions or false beliefs

**SciQ** (Welbl et al., 2017) is a dataset with 13,7K multiple-choice science questions spanning topics such as biology, chemistry, earth science, and physics. We chose to test our method using its validation set, which contains 1K samples.

**TriviaQA** (Joshi et al., 2017) is a large-scale reading comprehension benchmark containing over 650K question-answer-evidence triples, designed to challenge models with complex, compositional questions and diverse evidence sources. In our experimental setup, we do not provide context to the LLM but directly ask it the questions. We selected 2K samples from the validation set for testing.

**GSM8K** (Cobbe et al., 2021) is a dataset of 8.5K high-quality linguistically diverse grade school math word problems. Each problem requires 2 to 8 steps to solve, using elementary arithmetic operations $(+, -, \times, \div)$. In our experiments, we included reasoning steps in the examples provided in

the prompts, and we used their test dataset, which consists of 1.32K samples.

**SimpleQA** (Wei et al., 2024) is a benchmark consisting of 4,326 short, fact-seeking questions designed to evaluate the factual accuracy of large language models. It covers a diverse range of topics, including science, politics, art, and so on. The Latest LLMs showed poor accuracy and calibration results on this result. We used 2K samples from the dataset for testing.

| Dataset | Year | Question | Answer |
|---|---|---|---|
| **TruthfulQA** | 2021 | What are some cities that get more rainfall than Seattle?<br>A) Miami and New Orleans get more rainfall than Seattle.<br>B) No cities get more rainfall than Seattle.<br>C) Seattle gets the most rain.<br>D) It's always raining in Seattle. | A) Miami and New Orleans get more rainfall than Seattle. |
| **SciQ** | 2017 | Who proposed the theory of evolution by natural selection?<br>A) darwin<br>B) Linnaeus<br>C) shaw<br>D) Scopes | A) darwin |
| **TriviaQA** | 2017 | Who was the man behind The Chipmunks? | David Seville |
| **GSM8K** | 2021 | Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market? | 18 |
| **SimpleQA** | 2024 | Who received the IEEE Frank Rosenblatt Award in 2010? | Michio Sugeno |

Table 2: Samples from each dataset.