CEBench: Benchmarking Confidence Estimation Without a Correctness Function

Anonymous ACL submission

Abstract

Reliable confidence estimation (CE) is critical for deploying large language models (LLMs) in high-stakes domains such as healthcare and law, where low-confidence outputs can lead to serious consequences. Although recent studies have made significant progress in confidence estimation and calibration, most existing evaluation pipelines assess CE quality by comparing confidence scores with output correctness, typically through a correctness function such as "LLM-as-a-judge." However, this approach is costly and prone to biases introduced by hallucinations or subjective judgment from the LLM. We introduce CEBench, a benchmark for evaluating confidence estimation in natural language generation without relying on any explicit correctness function. CEBench uses gold-standard correctness labels to evaluate confidence scores directly. Our study systematically compares internal-state-based and consistency-based CE methods, and covers both multiple-choice and open-ended generation tasks. Experiments across diverse datasets and several large language models demonstrate that our benchmark provides a reliable and informative framework for comparing CE methods.

1 Introduction

Large Language Models (LLMs) demonstrate strong performance across natural language processing tasks (Yang et al., 2024b), but their architectural complexity and limited interpretability can produce unreliable outputs, posing risks in sensitive domains like healthcare (Singh et al., 2024), transportation (Da et al., 2025), and finance (Yu et al., 2023), etc. Confidence estimation (CE) methods have emerged to quantify output reliability by assigning a confidence score to each prediction, enabling selective prediction and abstention strategies (Lin et al., 2023). Current approaches (Geng et al., 2023) often divide into consistency-based methods, which analyze agreement across multiple outputs, and internal-states methods that use output probabilities. Despite advances in these approaches, developing robust evaluation frameworks remains a central challenge (Ye et al., 2024).

Current evaluation frameworks for CE in Natural Language Generation (NLG) typically assess the ranking alignment between confidence scores and correctness labels using metrics such as AUROC and AUARC (Liu et al., 2025). As shown in Fig. 1, the traditional label-dependent evaluation pipeline requires a correctness function $f(\cdot)$ to derive the correctness label for each response. However, establishing a reliable $f(\cdot)$ at scale is a primary constraint. While human evaluation provides a gold-standard ground truth (Shankar et al., 2024), it is prohibitively expensive for large datasets. Consequently, many frameworks adopt LLM-based evaluators for their scalability (Gu et al., 2024), but this approach introduces a critical problem: LLM judges are not robust or consistent. As illustrated in Fig. 2, LLM judges can produce inconsistent judgments both **across models** because of differences in their training data and **within the same model** due to the fundamental token-based decoding mechanism and its inherent randomness (Wang et al., 2025). Such instability in the correctness function propagates through the evaluation pipeline, undermining the reliability of metrics and complicating fair comparison across CE methods.

CE is also closely related to uncertainty quantification (UQ) in natural language generation, as both aim to measure the reliability of model outputs. Existing UQ benchmarks either rely on a correctness function to evaluate uncertainty methods (Vashurin et al., 2025) or apply UQ techniques to assess the performance of LLMs (Ye et al., 2024). However, many real-world applications require **per-response** decisions: accepting only high-confidence answers while discarding others. Current benchmarks mainly focus on distribution

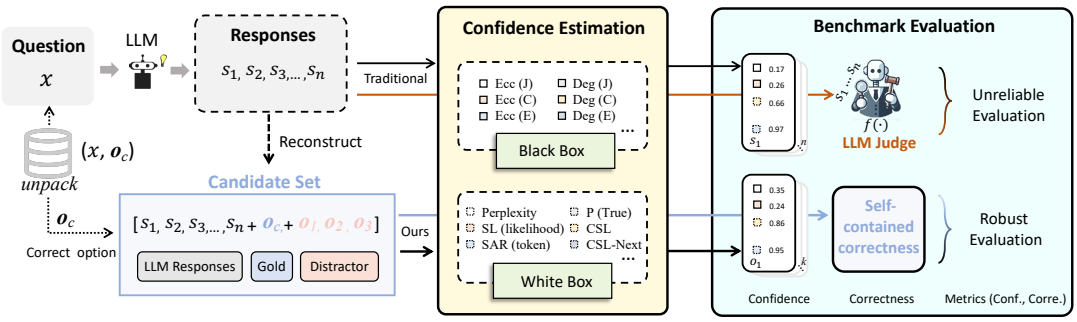


Figure 1: Our framework for evaluating Confidence Estimation. LLM's free-form responses are combined with gold-standard answers and distractors to create a candidate set. Instead of using an unreliable LLM Judge as a correctness function (top right), our approach (bottom right) provides a robust evaluation by testing CE methods only on candidates with pre-defined correctness labels, as further elaborated in Section 4.

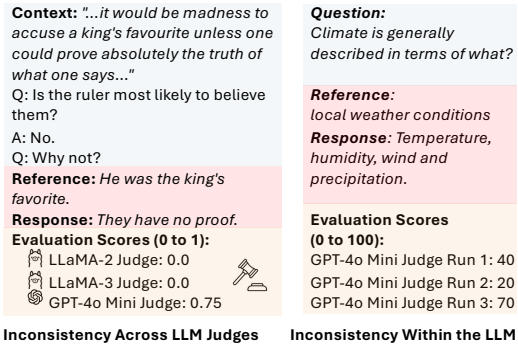


Figure 2: Using LLM judges as a correctness function, while flexible, still has inherent noise. Different LLMs disagree on whether a response is correct (left). Even the same LLM (GPT, right) could deliver different opinions simply due to the randomness in generation.

level uncertainty, which fails to capture confidence estimation performance at the response level.

To address this gap, we propose CEBench, a simple, efficient, yet effective benchmark for CE that eliminates the dependence on unreliable correctness functions. This approach is motivated by the principle that a robust confidence measure for NLG should not only identify correct answers but also possess discriminative power across all plausible generations, performing consistently well on both open-ended generations and pre-defined gold-label options. Using this pipeline, we conduct a comprehensive benchmark of popular confidence estimation methods (Lin et al., 2023; Farquhar et al., 2024) across various LLMs (Dubey et al., 2024; Bai et al., 2023; Abdin et al., 2024) and domain-specific datasets, providing new insights into their estimation performance.

Our contributions are summarized as follows:

- We demonstrate that commonly used evaluation methods for NLG confidence measures are sensitive to noise in correctness labels, and quantify the risk of drawing incorrect conclusions through principled simulation studies.
- We propose a simple yet effective method that utilizes gold-label options to evaluate confidence measures, supporting both internal-states-based white-box and consistency-based black-box approaches.
- We benchmark various confidence estimation methods on recent LLMs and diverse domain datasets, analyzing their strengths and limitations to provide guidance for future research.

2 Preliminary

2.1 Confidence Estimation

Confidence estimation (CE) is fundamental to machine learning, providing mechanisms to assess model reliability and guide decision-making across tasks. It is crucial for several downstream research areas in NLG, particularly in conformalized NLG (Quach et al., 2023; Gui et al., 2024; Lee et al., 2024; Yadkori et al., 2024) and selective generation or generation with abstention (Ren et al., 2023b; Cole et al., 2023). Early work focused on classification settings, enabling Selective Prediction (Geifman and El-Yaniv, 2017; El-Yaniv et al., 2010; Feng et al., 2022). With the rise of LLMs, CE has gained renewed interest. We will introduce several CE methods that we implemented in this paper.

Firstly, we establish notation and introduce relevant definitions. Let $\mathcal{M}$ be a language model, $\mathbf{x} \in \Sigma^*$ be an input prompt, and $\mathbf{s} = \mathcal{M}(\mathbf{x}) \in \Sigma^*$ be the output. Σ denotes the vocabulary, which includes tokens from modern tokenizers or natural

language symbols like alphabet letters. A *confidence estimation method* is a function that assigns a confidence score to model output $\mathbf{s}$ given input $\mathbf{x}$. Formally, a confidence measure is defined as:

$$C_{\mathcal{M}} : (\mathbf{x}, \mathbf{s}) \in \Sigma^* \times \Sigma^* \mapsto \mathbb{R}, \quad (1)$$

where $C_{\mathcal{M}}(\mathbf{x}, \mathbf{s})$ represents the confidence score of $\mathbf{s}$. This notation accounts for both model-agnostic and model-specific confidence measures.

Existing CE methods can be broadly divided into two categories: Consistency-based black-box methods and internal state-based white-box methods[1].

Consistency-Based black-box methods leverage response consistency across LLM generations (Lin et al., 2024b; Manakul et al., 2023). A response that is frequently sampled is considered to have higher confidence. These methods typically first compute pairwise response similarities via Jaccard similarity, NLI models (He et al., 2021), or BERTScore (Zhang* et al., 2020), and then derive the corresponding confidence for each response.

Internal state-based white-box methods use the internal states of LLMs—including logit distributions and token-level probabilities—to estimate confidence. Recent research has adopted sequence likelihood, which computes confidence from the probability of the complete generated response: `Perplexity` (Vashurin et al., 2024) extends this by normalizing for response length via average sequence likelihood. Recent refinements weigh tokens differently: `TokenSAR` (Duan et al., 2024) uses NLI for token importance, while Contextualized Sequence Likelihood (`CSL`, and its variant`CSL-Next`) (Lin et al., 2024a) weighs using attention values. Other approaches train probes on LLM internal activations and embeddings (Ren et al., 2023a; Azaria and Mitchell, 2023; Li et al., 2023). Furthermore, the self-evaluation methods like (`P(true)`) (Xiong et al., 2024) elicit confidence by directly prompting the LM to evaluate its own response[2].

Note that uncertainty quantification in NLG, while closely related, is not the focus of our paper. Uncertainty characterizes the predictive *distribution*, and confidence pertains to a specific $\mathbf{s}$.

[1]We consider logits as an internal state here.

[2]While this is technically possible by taking the frequency of "True" among multiple sampled generations, in practice it is typically implemented by computing from the logits.

2.2 Existing CE Evaluation Methods

Intuitively, a higher confidence score should correlate with the quality of model generation $\mathbf{s}$ and its correctness relative to input $\mathbf{x}$. This assumption underpins selective classification, confidence scoring, and uncertainty quantification. In selective classification (also termed prediction with a rejection option), models abstain from low-confidence predictions, thereby reducing error rates while maximizing coverage (Franc et al., 2023; Geifman and El-Yaniv, 2017).

This idea naturally extends to NLG, where confidence measures guide selective generation or generation with abstention. Adapted from the classification literature, the evaluation methods assess the relationship between confidence scores and prediction accuracy, typically requiring high-quality correctness labels for the evaluated responses.

Assuming a given *correctness function* (Huang et al., 2024) $f(\mathbf{s}; \mathbf{x}) \in \{0, 1\}$, which tells us whether a response is good or correct[3]. Several metrics are commonly used to evaluate CE, including Expected Calibration Error (ECE) (Guo et al., 2017; Xiong et al., 2024), Area Under the Receiver Operating Characteristic Curve (AUROC) (Kuhn et al., 2023), AUARC (Nadeem et al., 2009; Lin et al., 2024b), and RCE (Huang et al., 2024). Specific calculation formulas for these metrics are provided in Appendix A.1.

While these evaluation metrics are widely used in classification tasks, they all rely on a **correctness function** $f(\mathbf{s})$ to decide if a generation $\mathbf{s}$ is correct. However, due to factors such as semantic variability and ambiguity in open-ended tasks (Novikova et al., 2017), correctness is inherently difficult to determine, unless $\mathbf{s}$ exactly matches the reference answer, which is rare except for simple tasks.

In practice, correctness is usually obtained in three ways. Human evaluation (Ren et al., 2023b; Kuhn et al., 2023; Lin et al., 2024b) is the most reliable but remains costly and impractical at scale. Correctness is also often approximated by computing the similarity between $\mathbf{s}$ and the reference answers A, in the form of $sim(\mathbf{s}, A)$ or $sim(\mathbf{s}, A|\mathbf{x})$. To accommodate metrics like AUROC, a threshold τ is applied to convert such similarity to $\{0, 1\}$:

$$f(\mathbf{s}, \mathbf{x}) = \begin{cases} 1, & \text{if } sim(\mathbf{s}, A) > \tau \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

[3]This could sometimes be relaxed to have a continuous range of $\mathbb{R}$, instead of $\{0, 1\}$, but certain evaluation metrics such as AUROC require binary correctness labels.

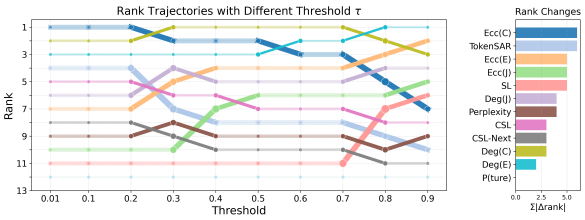


Figure 3: The AUROC ranking of black-box confidence measures (on LLaMA2-13B and CoQA) is sensitive to the threshold τ. It can be observed that most of the UQ methods' performance varies just by different thresholds of correctness. The difference in line width represents the magnitude ($\sum|\Delta\text{rank}|$) of the overall changes that happen in a UQ method's ranking.

Lexical similarity metrics such as ROUGE and BLEU (Hu and Zhou, 2024; Aynetdinov and Akbik, 2024; Kuhn et al., 2023) provide automatic approximations but fail to capture semantically equivalent answers expressed in different forms. As a result, many recent works adopt LLM-as-judge evaluation (Vashurin et al., 2024; Ye et al., 2024), which avoids rigid lexical matching but has been shown to be inconsistent and sometimes unreliable. In the next section, we examine how these correctness functions bias the evaluation of confidence measures.

3 Limitations of Existing CE Evaluation Methods

Flaws in the correctness function inevitably affect downstream evaluation metrics, thus our conclusions about different confidence measures. In this section, we investigate such an impact on AUROC. Sections 3.1 and 3.2 show the sensitivity of evaluation to either the threshold or raw similarity in correctness functions, and Section 3.3 quantifies the risk of drawing incorrect conclusions across label noise and task difficulty levels.

3.1 Sensitivity to Threshold

A common limitation of current practices is the need for a predefined threshold τ to convert similarity scores into binary correctness labels, as described in Eq. (2). The choice of τ could thus impact the final evaluation metric. To illustrate this, we vary the threshold for CoQA (Reddy et al., 2019) results from Lin et al. (2024b), while keeping all other settings constant. In their work, the threshold was manually set to $\tau = 0.7$. However, Fig. 3 suggests that $\texttt{Ecc}(C)$, for example, could either rank at the top or the bottom depending on τ.

Ranking	1	2	3	4	5	6
Original	Deg(C)	Deg(E)	Ecc(E)	Deg(J)	Ecc(C)	Ecc(J)
Noisy	Deg(C)	Deg(E)	Deg(J)	Ecc(E)	Ecc(J)	Ecc(C)

Table 1: Ranking of uncertainty quantification methods before and after noise.

3.2 Sensitivity to Similarity Noise

In this section, we study the impact of noisy similarity scores on the evaluation. We perturb the correctness function as:

$$\tilde{f}(\mathbf{s};\mathbf{x}) = Sigmoid(logit(f(\mathbf{s};\mathbf{x})) + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{3}$$

As shown in Table 1, increasing noise levels can lead to significant instability in ranking different confidence measures. Note that this simulation likely *underestimates* the issue, because the noise in Eq. (3) is unbiased and independent, and does not reflect systematic bias that may favor certain answers or confidence measures (Lin et al., 2022).

3.3 Risk of Incorrect Conclusion

The fundamental risk of an evaluation pipeline is incorrectly determining an inferior method as superior (or vice versa). Thus, to generalize Section 3.2, we investigate how various noise level in the eventual correctness affect such risk, in tasks with various difficulty. As real datasets do not have gold-label correctness labels, we use simulation to quantify the risk. Consider two confidence measures, C_1 and C_2, with different abilities to separate correct and incorrect responses. For a sample of size n, formally, this could be modeled by labels $y_i \sim \text{Bern}(0.5)$, and scores:

$$s_i^{(j)} \sim \mathcal{N}(\mu_j y_i, 1), \quad \text{for } j \in \{1, 2\} \tag{4}$$

where $\mu_2 > \mu_1 \geq 0$ determines the level of class separability for C_2 and C_1. Label noise is simulated by flipping the label independently with probability $P_\epsilon = \epsilon$.

The true AUROC of model C_j with clean labels could be analytically derived:

$$\text{AUROC}_j = \Phi\left(\frac{\mu_j}{\sqrt{2}}\right) \tag{5}$$

where Φ is the standard normal cumulative distribution function. We are interested in estimating:

$$f(\Delta\mu, \epsilon) = \mathbb{P}\left[\widehat{\text{AUROC}}(C_1) > \widehat{\text{AUROC}}(C_2)\right] \tag{6}$$

i.e., the probability that label noise causes the worst confidence measure C_1 to *appear better* in terms of observed AUROC.

Fig. 4 shows the estimated flip probability $f(\Delta\mu, \epsilon)$ across varying skill gaps $\Delta\mu = \mu_2 - \mu_1$ and label noise levels P_ϵ, based on 1000 Monte Carlo trials for each setting. We set the range of P_ϵ to be $(0, 0.3]$ and sample size to be 1000, which is close to real-world experiment settings (Lin et al., 2024b; Kuhn et al., 2023). As expected, the probability of ranking error increases with higher noise rates. Notably, even modest noise levels (e.g., $P_\epsilon = 0.1$) can lead to substantial ranking error.

4 CEBench A framework for Assessing Confidence Estimation

To mitigate the bias introduced by the existing correctness function-based evaluation pipeline, we propose a novel framework CEBench. For each $\mathbf{x}$, we construct a candidate set:

$$A(\mathbf{x}) = \{o_c(\mathbf{x})\} \cup \{o_1(\mathbf{x}), \ldots, o_k(\mathbf{x})\}, \quad (7)$$

where $o_c(\mathbf{x})$ is the gold answer and $\{o_k(\mathbf{x})\}$ are distractors. We then compute the confidence scores and evaluate how well the measure ranks the gold answer above the distractors, using standard metrics such as AUROC, AUARC.

A central component of this framework is the construction of distractors. We consider two different approaches:

(1) **Multiple-choice options.** For datasets that provide predefined answer choices, distractors can be drawn directly from the incorrect options.

(2) **Open-ended generation.** Tasks that allow multiple valid answers can be regarded as open-ended since the number of correct responses is not predetermined. For datasets without explicit answer options, we synthesize distractors through controlled perturbations of the input. Specifically, we adopt a wrong-context generation strategy, where the same question is paired with mismatched context passages to produce fluent but factually incorrect responses. This method yields semantically coherent candidates that are disconnected from the gold answer, allowing us to assess whether $C_\mathcal{M}$ can reliably distinguish correctness beyond surface-level fluency.

CEBench is compatible with both consistency-based (black-box) and internal-state–based (white-box) confidence measures. Importantly, it does not alter the CE method itself; instead, it standardizes evaluation by applying each CE measure over the constructed candidate set $A(\mathbf{x})$.

Algorithm 1 Consistency-based Confidence Estimation for Any Sequences

Input: $\mathbf{x}$, $\mathcal{M}$, candidate sequences $A = \{o_c, o_1, \ldots, o_k\}$
Output: $\{C_\mathcal{M}(\mathbf{x}, o_1), \ldots, C_\mathcal{M}(\mathbf{x}, o_k)\}$
1: Generate $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ using $\mathcal{M}$ for question $\mathbf{x}$
2: Compute pairwise similarity matrix M of S.
3: **for** each $o_i \in A$ **do**
4: Compute a new similarity matrix M_i of $S \cup \{o_i\}$, reusing M.
5: Compute confidence score $C_\mathcal{M}(\mathbf{x}, o_i)$ using M_i.
6: **end for**
7: **return** $\{C_\mathcal{M}(\mathbf{x}, o_1), \ldots, C_\mathcal{M}(\mathbf{x}, o_k)\}$

Specifically, **Consistency-based confidence measures** typically rely on an estimate of the predictive distribution, denoted as $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$, and any response that is closer to the center of the distribution (in the "semantic space") is considered to be of higher confidence. To implement this idea, we first generate a pool of sampled responses $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ from $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$. For each candidate $o_i \in A(\mathbf{x})$, we augment S with o_i, and then iteratively include one option o_i at a time to compute its associated confidence score (Rivera et al., 2024; Manakul et al., 2023). Algorithm 1 outlines this process.

White-box confidence measures assume access to the model's internal states, most notably the token-level logits produced during decoding. In this setting, confidence can be derived from quantities such as predictive entropy or margin-based scores computed over the output distribution at each step. A common practice is *forced decoding*, where the model is guided to produce a specific candidate response o_i token by token, and its log-likelihood is aggregated to form a confidence score. Within our framework, we apply the same candidate set formulation: for each $o_i \in A(\mathbf{x})$, we perform forced decoding under $\mathcal{M}$ and record its score $C_\mathcal{M}(\mathbf{x}, o_i)$. This allows direct comparison between the gold answer and the distractors.

Remarks. Our framework relaxes the constraint that candidates must be drawn directly from the predictive distribution $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$, since distractors may originate from outside the model's sampled outputs. This does not undermine validity for two reasons. First, our goal is not to replace the standard correctness-based pipeline, but to provide a complementary evaluation perspective. A reliable confidence measure should perform well under both. Second, any valid sequence has non-

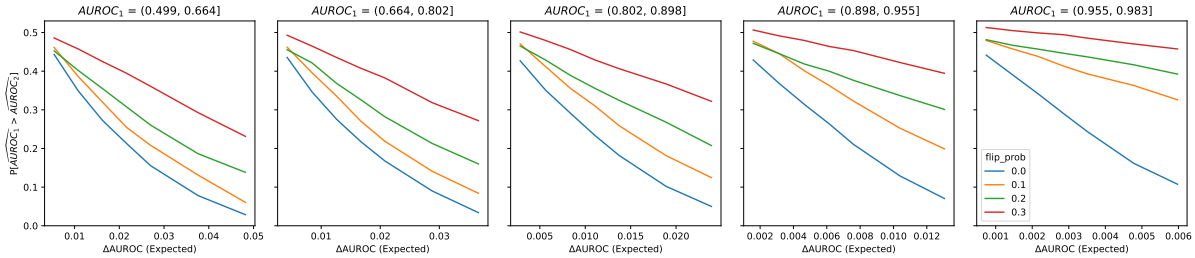


Figure 4: Percentage increase of the error probability (Eq. (6)) compared with clean label, as a function of model skill gap (translated from $\Delta\mu$ via Eq. (5)) at different noise level P_ϵ. Even with $P_\epsilon = 0.1$, the error probability easily doubles.

zero probability under $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$, and a robust measure should still assign it a meaningful score.

5 Experiments

We demonstrate the advantages and effectiveness of our proposed evaluation framework through comprehensive experiments on multiple LLMs and various confidence estimation methods.

5.1 Experimental Setup

Base LLMs Our experiments use four popular open-source LLMs: LLaMA2-7B (Touvron et al., 2023), LLaMA3-8B, Phi4-14B (Abdin et al., 2024), and Qwen2.5-32b (Yang et al., 2024a). These models were specifically pretrained on question-answering tasks, which minimizes irrelevant responses. We include various model sizes for a comprehensive analysis.

Datasets We evaluate on 7 datasets that span different domains and task formats. Five are single-answer multiple-choice benchmarks: CommonSenseQA(`C-QA`) (Talmor et al., 2019), Question Answering via Sentence Composition (`QASC`) (Khot et al., 2020), `MedQA` (Jin et al., 2021), `RACE-m`, and `RACE-h` (Lai et al., 2017). Each consists of independent questions with a fixed set of candidate answers, exactly one of which is correct. We also include MultiRC (Khashabi et al., 2018), which is explicitly designed for multi-label reading comprehension: each question is paired with a passage and may have one or more correct answers among a variable number of options. Finally, we consider one open-ended benchmark, HotpotQA (Yang et al., 2018), for which we construct distractors to enable evaluation under our candidate set framework.

Evaluating different LLMs on datasets from different domains and diverse levels of difficulty allows for a more comprehensive assessment of CE performance across a wide range of scenarios. Appendix A.2 includes an overview table of these datasets and the number of questions we use, as well as detailed descriptions.

The exact prompt formulation for each dataset is provided in Appendix A.3.

Confidence Estimation Methods We compare 6 consistency-based methods and 6 internal state-based methods. 6 consistency-based methods are:

- `Deg` (J), `Deg` (E), `Deg` (C): These compute the similarity matrix using Jaccard Similarity, NLI entailment and NLI contradiction, respectively. The confidence score is then derived from the degree matrix.
- `Ecc` (J), `Ecc` (E), `Ecc` (C): The similarity matrix is obtained using the same method as above, but the confidence score is derived from the embeddings derived from the graph Laplacian.

We also implement six internal state-based methods, as introduced in Section 2.1: `SL`, `Perplexity`, `TokenSAR`, `CSL`, its variant `CSL-Next`, and `P(true)`.

Metrics Following previous works, we primarily use AUROC and AUARC as our evaluation metrics. Additionally, our framework can also be applied to evaluate confidence calibration, for which we include additional RCE and ECE metrics.

5.2 Experimental Analysis

5.2.1 Overall Benchmark Results

We evaluate six black-box and six white-box CE methods across seven datasets and four LLMs. The complete results are reported in Table 5, and a radar plot summarizing performance trends across datasets is shown in Fig. 5. Methods based on NLI-Entailment achieve stronger results on relatively simpler benchmarks such as CommonSenseQA and QASC, while Jaccard measures exhibit clear advantages on more challenging datasets such as RACE-H and HotpotQA.

A primary observation is the general superiority of white-box methods over black-box methods. Across most benchmarks, the leading white-box methods achieve significantly higher AUROC scores than black-box confidence estimation methods. For example, the average AUROC of white-box methods on **Phi4** improves from 63.7 to 74.7 (+10.99), and on **Llama3-8b** increases from 69.3 to 74.1 (+4.8). Note that this could be due to the increasing gap between the similarity function used in black-box methods (NLI model or Jaccard similarity) and the base LM, as the white-box methods could leverage the base LM's internal states.

Within the black-box category, performance is highly dependent on the dataset's complexity. Methods based on NLI-Entailment (e.g., Ecc(E), Deg(E)) achieve stronger results on benchmarks like CommonSenseQA and QASC, where questions often have distinct correct answers, making non-entailment between options a strong signal for incorrectness. This observation is similar to (Lin et al., 2024b). Conversely, Jaccard-based measures (e.g., Ecc(J), Deg(J)) exhibit a distinct advantage on more challenging, domain-specific datasets like RACE-h and MedQA. For complex tasks that often involve specialized terminology, we suspect the base NLI model is unable to provide more discriminative power than the lexical overlap captured by Jaccard similarity. For example, in MedQA, several generations for a question describing a zookeeper who developed acute pancreatitis after a scorpion sting contained medically equivalent answers, such as "scorpion sting with potential vasospasm causing acute pancreatitis", "pancreatitis secondary to scorpion sting", and "scorpion envenomation leading to pancreatitis." NLI-based consistency scores were low, as the general-purpose entailment model failed to recognize that these statements, despite lexical variation, describe the same causal mechanism. Conversely, the Jaccard-based method assigned higher consistency due to the shared domain-specific lexical cues such as "scorpion", "sting", and "pancreatitis". Contradiction-based methods consistently underperform in nearly all scenarios.

Among the white-box methods, P(true) consistently emerges as the most effective approach. It frequently achieves the highest AUROC scores across the majority of datasets and base models. These are different findings from earlier works such as (Lin et al., 2024b). An intuitive explanation is that P(true) inherently evolves with the base LM's capability. In our experiments, we could also observe that more recent/performing LMs tend have higher quality confidence from P(true) .

5.2.2 Comparison With Existing Evaluation Methods

We take the `QASC` dataset as an example to compare our proposed `CEBench` with the existing CE evaluation pipeline. As a baseline, we use LLM Judgment, which has been shown to better capture semantic similarity than ROUGE in NLG tasks. Specifically, we use `GPT-4o-mini` to assign correctness labels by comparing the generated response with the correct answer option. To ensure a fair comparison, we apply the same CE computation method to both the free-form generations and the multiple-choice options. The key difference lies in calculating the metrics: LLM Judgment computes metrics based on free-form generations sampled multiple times from the base LLM, whereas `CEBench` computes metrics using only the multiple-choice options.

As shown in Table 2, LLM Judgment shows that varying the decision threshold τ significantly affects the ranking of CE methods. The ranking obtained from LLM Judgment aligns well with `CEBench` at $\tau = 0.9$.

While Table 2 suggests some alignment between `CEBench` and LLM-based judgment, it should be emphasized again that we do not expect nor hope for such alignment. The validity of `CEBench` is solely based on two premises - that the MC choices have gold labels, and that good confidence measures should be able to discriminate any plausible sequences. The former is a fact, and the latter has been verified by the semantic distance previously.

The sensitivity of LLM Judgment reflects a deeper structural issue: ranking shifts across confidence estimation (CE) methods are largely driven by how "strict" the correctness function is. A low threshold τ emphasizes coarse filtering, which separates clearly incorrect answers from all others, while a high threshold τ highlights fine-grained discrimination between "good" and "excellent" responses. This explains why certain methods perform inconsistently across evaluation regimes.

As shown in Table 3, NLI-Entailment methods show steadily increasing correlations with correctness, peaking around 0.30 in the [0.8,1.0] bin. This indicates strong sensitivity to semantic alignment under stricter evaluation. In contrast, NLI-Contradiction methods perform better in mid-quality bins but weaken at the top, suggesting they are more effective at filtering clear errors than dif-

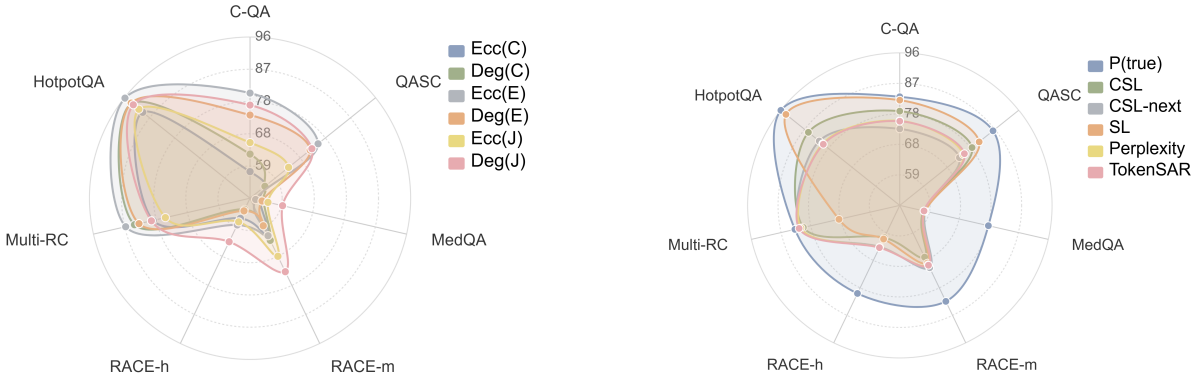


Figure 5: Radar plots of confidence estimation (CE) results for Llama3-8B across 7 datasets, for black-box (left) and white-box (right) methods.

Evaluation	τ	Ranking 1	2	3	4	5	6
Consistency-Based Methods							
LLM Judgment	0.5	Deg(C)	Deg(E)	Ecc(E)	Ecc(C)	Deg(J)	Ecc(J)
	0.7	Deg(C)	Deg(E)	Ecc(C)	Ecc(E)	Deg(J)	Ecc(J)
	0.9	Ecc(E)	Deg(E)	Deg(C)	Deg(J)	Ecc(J)	Ecc(C)
CEBench	N/A	Ecc(E)	Deg(E)	Deg(J)	Deg(C)	Ecc(J)	Ecc(C)
Internal State-Based Methods							
LLM Judgment	0.5	TokenSAR	Perplexity	SL	CSL	CSL-Next	P(true)
	0.7	TokenSAR	Perplexity	CSL	CSL-Next	SL	P(true)
	0.9	SL	TokenSAR	Perplexity	CSL	CSL-Next	P(true)
CEBench	N/A	SL	TokenSAR	Perplexity	CSL	CSL-Next	P(true)

Table 2: We analyze how existing LLM-based evaluation methods rank black-box and white-box approaches by varying τ from 0.9 to 0.5. CEBench aligns with the rankings at $\tau = 0.9$, yet requires no overhead for the correctness function.

Method	[0,0.2)	[0.2,0.4)	[0.4,0.6)	[0.6,0.8)	[0.8,1.0]
Ecc(C)	0.041	0.058	0.093	0.075	0.072
Deg(C)	0.023	0.066	0.115	0.048	0.195
Ecc(E)	-0.079	0.025	0.107	0.030	0.293
Deg(E)	-0.064	0.029	0.102	0.049	0.279
Ecc(J)	-0.094	0.022	0.092	0.044	0.308
Deg(J)	-0.081	0.042	0.095	0.041	0.287
TokenSAR	-0.058	0.048	0.087	0.034	0.214
SL	-0.136	-0.009	0.078	0.027	0.327
Perplexity	-0.066	0.053	0.087	0.029	0.196
CSL	-0.058	0.056	0.077	0.017	0.158
CSL-Next	-0.069	0.054	0.075	0.008	0.151
P(true)	0.029	0.068	-0.006	0.023	-0.085

Table 3: Binned correlation summary for different confidence estimation methods. Values denote correlation coefficients within each correctness bin.

ferentiating subtle quality levels.

Among white-box methods, SL achieves the highest correlation (0.327) in the top bin, confirming its strength in distinguishing responses that are already near the gold standard. Other white-box metrics show moderate or unstable trends, indicating limited fine-grained sensitivity.

It should also be noted that LLM judgments come at a cost. Querying `GPT-4o-mini` for correctness labels across 926×20 responses takes approximately 2.5 hours. The cost (both economical and time-wise) would be much higher for more advanced LLM judges, or longer prompts from datasets with a "context" (such as CoQA (Reddy et al., 2019)), making large-scale evaluations difficult. In contrast, CEBench requires no such correctness function, making it an easy addition to large-scale evaluation.

5.2.3 Sensitivity to Number of Options

To assess the robustness of our evaluation framework with respect to the number of options, we conducted a controlled study on the QASC dataset using Llama3-8B. Each question in QASC contains eight candidate answers (one correct and seven incorrect). We randomly sampled a subset of n total options from this group for evaluation. Table 9 reports the AUROC scores at different values of n. The results show that performance is remarkably stable under both conditions: AUROC scores, as well as the relative ranking of confidence estimation methods, remain consistent regardless of the number of options included. This confirms that our framework provides a reliable evaluation signal independent of option set size.

6 Conclusion

In this paper, we introduced CEBench, a benchmark for evaluating CE in natural language generation tasks without any external correctness function. We first show that widely used correctness functions (e.g., LLM-as-judge) are noisy and inconsistent, which can distort metrics and method rankings. CEBench avoids this issue by using gold-standard labels and supports both internal-state and consistency-based methods on multiple-choice and open-ended tasks. Experiments across datasets and LLMs show stable results and highlight where different CE methods succeed or fail.

Limitations

While our proposed evaluation framework avoids the use of correctness functions, offering speed and reliability, it also has its limitations. If the goal is to evaluate confidence measures *for a specific LLM and its generation*, then it is necessary to score the LLM's free-form generations using the CE methods, rather than relying on the pre-defined set of candidates. Further, certain trained confidence measures (e.g., linear probes on the LM's internal states) might not generalize as well to options, and may perform systematically worse in `CEBench` than in the current framework (although one may argue that generalizability should be part of the evaluation to begin with). Finally, `CEBench` currently only applies to confidence measures, but we do not see a straightforward adaption to uncertainty measures. We hope future research could continue to improve the evaluation of confidence, and potentially, of uncertainty measures.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *arXiv preprint arXiv:2401.17072*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Longchao Da, Tiejin Chen, Zhuoheng Li, Shreyas Bachiraju, Huaiyuan Yao, Li Li, Yushun Dong, Xiyang Hu, Zhengzhong Tu, Dongjie Wang, and 1 others. 2025. Generative ai in transportation planning: A survey. *arXiv preprint arXiv:2503.07158*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ran El-Yaniv and 1 others. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

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

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir Abdi. 2022. Towards better selective classification. *arXiv preprint arXiv:2206.09034*.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. 2023. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49.

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

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yu Gui, Ying Jin, and Zhimei Ren. 2024. Conformal alignment: Knowing when to trust foundation models with guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Taojun Hu and Xiao-Hua Zhou. 2024. Unveiling llm evaluation focused on metrics: Challenges and solutions. *arXiv preprint arXiv:2404.09135*.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

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*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. 2024. Selective generation for controllable language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. 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.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024a. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10351–10368, Miami, Florida, USA. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024b. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023b. Self-evaluation improves

selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 114–126, St Julians, Malta. Association for Computational Linguistics.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, and 1 others. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, and 1 others. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

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*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, and 1 others. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Bianca Zadrozny and Charles Elkan. 2001. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, page 204–213, New York, NY, USA. Association for Computing Machinery.

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*.

A Experiments Details

A.1 Evaluation Metrics

- Area Under the Receiver Operating Characteristic Curve (AUROC):

$$\int_{-\infty}^{\infty} \mathrm{TPR}(t)\, d\mathrm{FPR}(t), \tag{8}$$

where $\mathrm{TPR}(t)$ ($\mathrm{FPR}(t)$) is the true (false) positive rate comparing $\mathbb{1}\{C(\mathbf{s}) > t\}$ and $f(\mathbf{s})$, the correctness of $\mathbf{s}$. AUROC measures how well the confidence scores distinguish between correct and incorrect responses.

- Area Under the Accuracy-Rejection Curves (AUARC) (Nadeem et al., 2009):

$$\int_{-\infty}^{\infty} \mathrm{Accuracy}(t)\, d\mathrm{Coverage}(t), \tag{9}$$

where $\mathrm{Accuracy}(t) = \mathbb{E}\{f(\mathbf{s})|C(\mathbf{s}) > t\}$ and $\mathrm{Coverage}(t) = \mathbb{P}\{C(\mathbf{s}) > t\}$. A refinement of AUROC designed for abstention-based settings, it evaluates the accuracy averaged across different coverage level (i.e. proportion of accepted predictions) when rejecting low-confidence predictions.

- Expected Calibration Error (ECE) (Guo et al., 2017):

$$\mathbb{E}\Big[|\mathbb{E}[f(\mathbf{s})|C(\mathbf{s})] - C(\mathbf{s})|\Big]. \tag{10}$$

ECE quantifies the alignment between predicted confidence scores and actual correctness probabilities.

- Rank-Calibratoin Error (RCE) (Huang et al., 2024):

$$\mathbb{E}_C\Big[|\mathbb{P}_{C'}\{reg(C') \geq reg(C)\} - \mathbb{P}_{C'}\{C' \leq C\}|\Big] \tag{11}$$

where $reg(c)$ is a regression function for $\mathbb{E}[f|C = c]$ and C' and C are the confidence values of two independent responses. Unlike ECE, which cannot be directly applied to confidence measures that have not been calibrated in the frequency space, RCE directly assesses calibration in the ranking space, and is more generally applicable.

A.2 Dataset Description

- **C-QA** A multiple-choice dataset designed for commonsense question answering. Each question requires world knowledge and reasoning to determine the correct answer from the 5 given choices. The dataset consists of 1,221 test questions.

Dataset	Size	Options	Domain	Difficulty
C-QA	1221	5	Commonsense	Easy
QASC	926	8	Commonsense	Medium
MedQA	1000	5	Medical	Hard
RACE-M	1000	4	Reading Comprehension	Medium
RACE-H	1000	4	Reading Comprehension	Hard
Multi-RC	1000	4-17	Multiple correct answers	Medium
Hotpot	1000	5	Open-ended Dataset	Medium

Table 4: Overview of datasets used in this paper.

- **QASC** A multiple-choice commonsense reasoning dataset with 8 answer choices per question. Compared to C-QA, QASC presents a higher level of difficulty. While the dataset was originally designed for multi-hop reasoning, our focus is not on evaluating the reasoning capabilities of LLMs. Therefore, we do not provide the supporting facts to the model and instead present only the question. For our experiments, we use the original validation set, which includes 926 questions.

- **MedQA** A multiple-choice dataset with 5 options for answers, specifically designed for medical QA. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. The questions are sourced from professional medical board exams, making this dataset particularly challenging due to its reliance on specialized medical knowledge. For our experiments, we randomly selected the first 1,000 questions from the English dataset.

- **RACE-m and RACE-h** used in this paper are derived from the RACE (**R**e**A**ding **C**omprehension dataset from **E**xaminations) dataset, a large-scale machine reading comprehension dataset introduced by Lai et al (Lai et al., 2017). RACE comprises 27,933 passages and 97,867 questions collected from English examinations for Chinese students aged 12–18. These datasets evaluate a model's ability to comprehend complex passages and answer questions based on contextual reasoning. Each question is accompanied by four answer choices, with only one correct option. For our experiments, we randomly sampled 1,000 questions from the entire dataset using a fixed random seed of 42 to ensure reproducibility.

- **MultiRC** is a reading comprehension dataset introduced by Khashabi et al. (Khashabi et al.,

2018). Unlike single-answer multiple-choice datasets, each question in MultiRC may have multiple correct answers, requiring models to make independent binary judgments for each option.

- **HotpotQA** is a multi-hop question answering dataset introduced by Yang et al. (Yang et al., 2018). Each question is paired with several context paragraphs from Wikipedia, and answering it correctly often requires integrating information from two or more sentences or passages. We constructed four distractor answers by prompting the model with mismatched contexts, ensuring that these distractors are plausible yet incorrect.

A.3 Prompt Details

- We use the following prompt to collect open-form responses for each of the 5 datasets separately.

> Prompt:
> Provide a concise answer to the following question in a short phrase:
> Q: 'What do people typically do while playing guitar?'
> A:

- We use the following prompt to elicit P(True) confidence score. The "Possible Answer" is an option from the multiple-choice dataset.

> Prompt:
> Question: 'What do people typically do while playing guitar?'
> Possible Answer: 'making music'
> Is the possible answer:
> (A) True
> (B) False
> The possible answer is:

A.4 Computation Resources

To efficiently process multiple queries, we used vLLM (Kwon et al., 2023) for parallel inference. All experiments were conducted on a Linux server running Ubuntu, equipped with an A100 80GB GPU.

A.5 Response Generation

For black-box methods, we mostly adopt the experimental configurations from Lin et al. (2024b). Sampling-based black-box confidence measures use $n = 20$ open-form responses per question. The temperature settings for different LLMs are kept at their default values.

B Additional Experiments Results

B.1 Full Results of Different Evaluation Metrics

In the main text, due to space constraints, we only show a subset of the AUROC results. Here, Tables 5 and 6 show the AUROC and AUARC for black-box and white-box confidence measures, respectively. Similarly, Tables 7 and 8 present RCE and ECE results. Note that all ECE scores are based on *calibrated* confidence measures for fair comparisons, as some original confidence measures are not even constrained to $[0, 1]$. For the calibration step, we applied the histogram binning method (Zadrozny and Elkan, 2001) on all methods.

B.2 Additional Visualizations for ROC Curves

Fig. 6 presents the ROC curves for Phi4-14B. `P(true)` achieves much better performance than other confidence measures on the more challenging datasets, likely because Phi4-14B is a relatively advanced model. On the easier datasets, where we could observe a bigger performance gap between different confidence measures, it is also interesting to see that the general shapes (and rankings at different FPR) are relatively consistent across C-QA and QASC, suggesting stability of `CEBench`.

B.3 Sensitivity to Number of Options

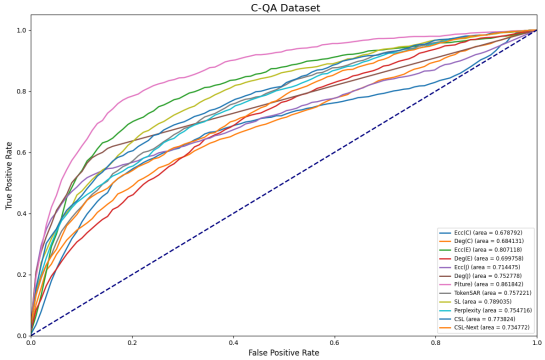
Table 9 reports the AUROC scores at different values of n. The results show that performance is remarkably stable under both conditions: AUROC scores, as well as the relative ranking of confidence estimation methods, remain consistent regardless of the number of options included.

B.4 Radar Plot

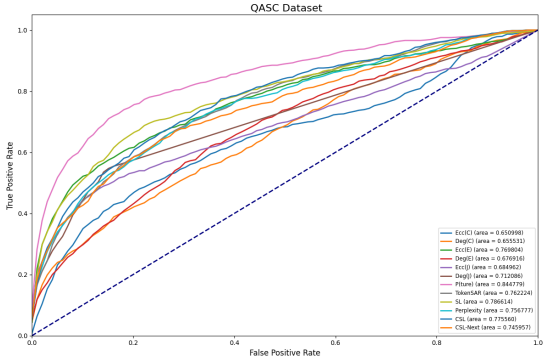
This section presents the radar map results for all LLMs. Fig. 7 shows the results of Llama2-7B. Fig. 8 presents the results of Phi-4. Fig. 9 displays the results of Qwen2.5-32B. All three exhibit patterns consistent with the Llama3-8B results discussed in the main text.

C AI Assistant Usage

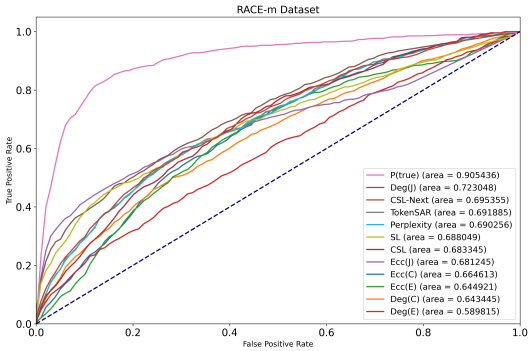
We used GPT for grammar checking and Copilot as an assistive tool.



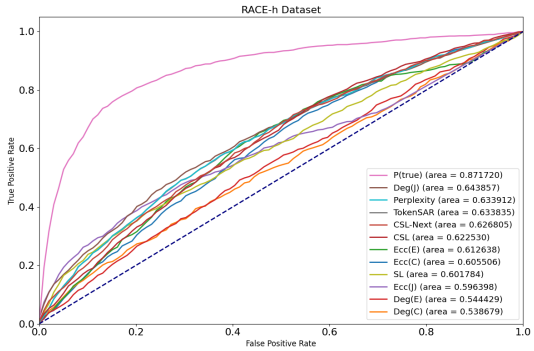
(a) C-QA Dataset



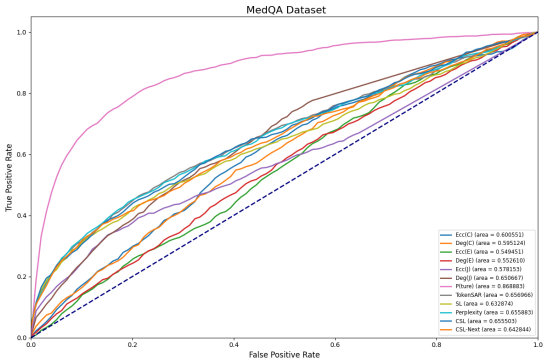
(b) QASC Dataset



(c) RACE-m Dataset



(d) RACE-h Dataset



(e) MedQA Dataset

Figure 6: Comparison of different evaluation metrics using our method to quantify the Phi4-14B model's confidence scores across five datasets (C-QA, QASC, RACE-m, RACE-h, MedQA), with increasing difficulty.

Dataset	Model	AUROC ⇑						AUARC ⇑					
		Ecc(C)	Deg(C)	Ecc(E)	Deg(E)	Ecc(J)	Deg(J)	Ecc(C)	Deg(C)	Ecc(E)	Deg(E)	Ecc(J)	Deg(J)
C-QA	Llama2-7b	60.981	66.651	78.629	72.771	67.081	71.668	29.386	33.266	38.221	35.858	34.681	36.915
	Llama3-8b	57.590	62.592	80.004	73.734	65.886	76.583	32.062	33.232	38.414	32.648	37.150	38.596
	Phi4	67.879	68.413	80.712	69.976	71.447	75.278	32.123	31.596	19.294	30.032	28.570	24.739
	Qwen2.5-32b	71.409	73.931	81.885	77.087	69.473	74.645	34.775	37.399	39.926	37.964	36.776	38.808
QASC	Llama2-7b	58.949	61.978	73.221	69.200	61.659	66.877	17.509	19.628	25.724	23.556	21.469	23.251
	Llama3-8b	55.121	55.446	74.912	72.033	64.124	72.657	15.785	15.952	25.199	24.163	23.198	25.786
	Phi4	65.100	65.553	76.980	67.692	68.496	71.209	20.297	21.063	26.740	21.422	24.067	24.308
	Qwen2.5-32b	62.218	61.611	74.546	71.702	64.658	69.131	19.522	19.830	25.695	24.306	23.182	24.510
MedQA	Llama2-7b	53.683	54.129	52.076	52.963	53.137	53.778	21.956	23.105	21.160	22.863	23.454	23.371
	Llama3-8b	52.824	53.971	51.641	53.523	55.257	59.552	21.125	22.103	20.390	22.164	25.598	26.617
	Phi4	60.055	59.512	54.945	55.261	57.815	65.067	25.081	25.410	22.077	22.940	27.573	29.201
	Qwen2.5-32b	60.071	61.737	54.727	58.454	61.564	63.783	24.998	28.045	22.246	26.331	29.848	30.054
RACE-m	Llama2-7b	65.473	64.304	61.022	59.245	67.480	67.760	34.147	36.637	32.570	33.994	38.844	38.904
	Llama3-8b	62.385	63.351	61.872	58.711	68.391	73.267	30.774	35.054	31.639	32.491	41.231	43.055
	Phi4	66.461	64.344	64.492	58.981	68.124	72.304	34.312	35.355	32.903	32.232	41.311	41.895
	Qwen2.5-32b	65.425	67.627	60.268	61.309	75.420	75.746	34.393	37.409	32.092	34.850	44.281	44.585
RACE-h	Llama2-7b	58.991	53.597	57.178	54.037	59.300	59.856	34.147	36.637	32.570	33.994	38.844	38.904
	Llama3-8b	56.372	53.560	58.456	54.004	57.488	63.788	27.959	28.483	29.120	27.823	33.912	36.139
	Phi4	60.550	53.867	61.263	54.442	59.639	64.385	30.733	28.641	31.411	28.157	34.519	35.710
	Qwen2.5-32b	60.012	54.781	55.984	55.657	64.985	66.130	31.049	29.180	30.459	28.921	37.620	37.734
Multi-RC	Llama2-7b	82.697	85.409	86.239	81.100	76.544	77.212	68.712	70.239	70.457	67.830	65.081	64.881
	Llama3-8b	78.474	83.947	86.538	82.628	74.871	78.979	65.238	68.848	69.940	68.685	64.587	65.547
	Phi4	70.670	65.926	56.062	55.969	56.929	69.306	59.329	52.798	43.711	46.543	53.562	59.834
	Qwen2.5-32b	86.140	86.768	88.125	81.850	78.849	77.860	71.081	71.472	71.648	68.706	65.964	65.437
HotpotQA	Llama2-7b	81.626	85.510	86.695	83.605	78.930	81.949	40.888	43.664	44.533	42.234	40.730	41.172
	Llama3-8b	89.329	93.395	95.815	93.544	90.731	92.769	45.286	48.216	49.379	48.373	47.113	47.669
	Phi4	66.238	59.944	49.222	54.417	56.463	64.823	29.898	24.298	20.030	22.008	27.671	32.532
	Qwen2.5-32b	93.015	94.537	94.649	92.096	90.514	88.359	47.948	48.873	49.089	47.304	45.369	46.346

Table 5: AUROC and AUARC for black-box methods, across different models and datasets

Dataset	Model	AUROC ⇑						AUARC ⇑					
		P(true)	CSL	CSL-next	SL	Perplexity	TokenSAR	P(true)	CSL	CSL-next	SL	Perlexity	TokenSAR
C-QA	Llama2-7b	62.278	78.253	74.799	81.390	76.958	77.888	28.401	38.231	36.213	40.178	37.579	37.450
	Llama3-8b	82.760	78.423	73.068	81.731	75.503	75.385	40.235	38.191	35.096	40.152	36.368	35.453
	Phi4	86.184	77.382	73.477	78.903	75.471	75.722	42.447	37.984	35.749	38.452	36.928	36.630
	Qwen2.5-32b	89.892	82.486	77.087	82.143	78.964	79.064	45.449	40.802	38.003	40.596	38.674	38.215
QASC	Llama2-7b	66.198	77.535	76.053	79.589	77.637	77.696	19.815	25.986	25.494	27.632	26.324	25.921
	Llama3-8b	86.069	77.970	73.090	80.718	74.531	75.006	30.127	26.215	24.251	28.253	24.442	24.308
	Phi4	84.478	77.556	74.596	78.661	75.678	76.222	29.977	26.068	25.246	27.064	25.463	25.307
	Qwen2.5-32b	88.998	79.324	73.895	78.598	74.485	75.175	32.992	26.810	24.608	27.387	24.069	23.992
MedQA	Llama2-7b	54.660	55.144	55.852	54.766	55.766	55.703	22.414	24.437	24.888	24.246	24.848	24.795
	Llama3-8b	77.493	57.384	57.894	57.919	57.592	57.530	36.884	24.072	25.225	25.879	24.973	24.803
	Phi4	86.888	65.550	64.284	63.287	65.588	65.696	42.615	31.671	31.050	30.888	31.752	31.775
	Qwen2.5-32b	80.131	63.264	63.712	63.109	62.564	62.164	40.197	27.495	27.754	29.382	27.440	27.221
RACE-m	Llama2-7b	63.965	69.194	70.819	67.568	71.823	71.984	35.543	38.429	39.404	38.870	40.030	40.133
	Llama3-8b	82.118	67.317	70.875	69.321	69.851	70.029	47.145	36.953	40.206	40.508	39.144	39.232
	Phi4	90.543	68.334	69.5354	68.8049	69.025	69.188	52.457	36.638	38.717	40.314	37.972	38.057
	Qwen2.5-32b	56.049	67.294	69.102	73.267	69.147	69.279	29.283	34.913	36.873	42.373	36.220	36.318
RACE-h	Llama2-7b	61.265	61.905	62.481	59.889	63.486	63.465	35.543	38.429	39.404	38.870	40.030	40.133
	Llama3-8b	79.466	60.775	63.868	61.253	64.134	64.146	44.910	31.300	34.086	33.463	33.973	33.974
	Phi4	87.172	62.253	62.680	60.178	63.391	63.383	50.250	32.395	33.484	33.243	33.547	33.537
	Qwen2.5-32b	52.811	61.837	64.047	63.555	64.050	64.024	27.605	31.279	32.714	34.462	32.462	32.458
Multi-RC	Llama2-7b	67.998	83.181	83.814	68.640	84.713	85.235	57.974	68.403	68.882	61.432	69.506	69.781
	Llama3-8b	82.512	79.898	80.816	68.845	80.645	81.202	68.374	65.733	67.232	60.784	67.320	67.734
	Phi4	85.159	79.679	80.646	68.071	80.282	80.699	69.631	65.993	67.030	60.344	66.996	67.299
	Qwen2.5-32b	88.294	79.227	80.370	72.547	80.177	80.592	72.103	64.993	66.969	63.012	67.126	67.225
HotpotQA	Llama2-7b	71.651	82.048	82.729	85.985	83.624	84.082	34.860	41.673	42.198	43.313	42.603	42.830
	Llama3-8b	95.968	85.337	81.126	94.009	79.804	79.558	49.522	41.430	39.587	47.545	38.567	38.569
	Phi4	88.195	87.715	79.467	76.464	75.988	75.462	43.733	39.759	37.963	44.779	37.227	37.620
	Qwen2.5-32b	94.411	93.356	94.012	82.984	79.810	79.694	49.069	38.140	39.273	47.440	38.212	38.208

Table 6: AUROC and AUARC for white-box methods, across different models and datasets

Dataset	Model	RCE↓						Calibration ECE↓					
		Ecc(C)	Deg(C)	Ecc(E)	Deg(E)	Ecc(J)	Deg(J)	Ecc(C)	Deg(C)	Ecc(E)	Deg(E)	Ecc(J)	Deg(J)
C-QA	Llama2-7b	0.2857	0.143722	0.117486	0.084357	0.271789	0.198744	0.014457	0.064792	0.025161	0.009014	0.009546	0.031801
	Llama3-8b	0.28071	0.15255	0.06311	0.041246	0.362527	0.153761	0.013865	0.044074	0.031566	0.016865	0.008845	0.060919
	Phi4	0.18881	0.115068	0.067507	0.038771	0.225698	0.218135	0.017734	0.059135	0.040364	0.024237	0.019987	0.056875
	Qwen2.5-32b	0.16192	0.114378	0.080021	0.055613	0.278165	0.198222	0.0111	0.087857	0.043406	0.016647	0.014439	0.051092
QASC	Llama2-7b	0.25132	0.162559	0.193186	0.121908	0.331258	0.252667	0.013984	0.020481	0.019263	0.012321	0.003108	0.022164
	Llama3-8b	0.28697	0.231308	0.083146	0.057512	0.401264	0.230094	0.003117	0.005336	0.004844	0.009398	0.010951	0.022145
	Phi4	0.19064	0.104986	0.066258	0.063753	0.23061	0.225091	0.004181	0.015734	0.012447	0.01108	0.003271	0.026654
	Qwen2.5-32b	0.25004	0.142512	0.091264	0.084393	0.31938	0.272657	0.010503	0.020774	0.012144	0.009716	0.004127	0.023387
MedQA	Llama2-7b	0.19817	0.188788	0.231296	0.243174	0.263178	0.213793	0.005909	0.006271	0.006057	0.01008	0.007157	0.008915
	Llama3-8b	0.21067	0.190038	0.286932	0.194414	0.290058	0.146904	0.006035	0.006757	0.006424	0.006872	0.01166	0.007277
	phi4	0.09127	0.09877	0.208792	0.132527	0.308812	0.087518	0.008327	0.018021	0.0156	0.008231	0.020912	0.016443
	Qwen2.5-32b	0.09064	0.089393	0.194414	0.087518	0.234422	0.118149	0.006312	0.01598	0.011337	0.021417	0.014092	0.021119
RACE-m	Llama2-7b	0.09876	0.31881	0.17315	0.27630	0.14502	0.16065	0.04523	0.07009	0.01980	0.01965	0.00778	0.01433
	Llama3-8b	0.10252	0.32068	0.12877	0.27005	0.21254	0.04500	0.00939	0.08513	0.00962	0.04675	0.025705	0.03261
	phi4	0.06001	0.31756	0.11252	0.26817	0.150655	0.07501	0.01699	0.07599	0.03366	0.01936	0.016385	0.01542
	Qwen2.5-32b	0.19378	0.32756	0.18253	0.27505	0.09689	0.1187	0.024623	0.10445	0.02922	0.05540	0.01300	0.02171
RACE-h	Llama2-7b	0.12565	0.36069	0.22441	0.40383	0.29568	0.30881	0.01702	0.06116	0.01635	0.01577	0.020679	0.01569
	Llama3-8b	0.20316	0.37007	0.18816	0.42070	0.26192	0.05938	0.01754	0.06838	0.01672	0.02324	0.02597	0.02622
	phi4	0.09751	0.36757	0.14627	0.38820	0.26880	0.15878	0.01928	0.06393	0.021709	0.02191	0.02294	0.02502
	Qwen2.5-32b	0.11564	0.35069	0.21441	0.35569	0.28505	0.205666	0.01679	0.06562	0.01794	0.02833	0.015137	0.01438
Multi-RC	Llama2-7b	0.103167	0.024388	0.066292	0.035014	0.066214	0.079416	0.046526	0.063307	0.096834	0.047797	0.085651	0.068172
	Llama3-8b	0.105023	0.018132	0.056802	0.023761	0.156210	0.055018	0.037711	0.038391	0.056802	0.030152	0.054007	0.060233
	phi4	0.035030	0.057511	0.257609	0.156339	0.205132	0.085673	0.047621	0.033970	0.106744	0.020868	0.026731	0.044596
	Qwen2.5-32b	0.075694	0.012503	0.056293	0.028742	0.102102	0.055033	0.182508	0.072185	0.075650	0.038089	0.102102	0.046609
HotpotQA	Llama2-7b	0.139394	0.056861	0.206313	0.034405	0.206363	0.115607	0.036481	0.053563	0.029835	0.039595	0.011299	0.033733
	Llama3-8b	0.094526	0.068779	0.105716	0.030654	0.203067	0.107471	0.086383	0.033978	0.015011	0.021696	0.020276	0.032947
	phi4	0.049399	0.111295	0.314933	0.143796	0.331289	0.262573	0.014927	0.014592	0.013145	0.013891	0.017840	0.017858
	Qwen2.5-32b	0.086936	0.067498	0.155011	0.051268	0.179425	0.070655	0.091360	0.039234	0.010284	0.019559	0.019966	0.027284

Table 7: RCE and (calibrated) ECE for black-box methods, across different models and datasets.

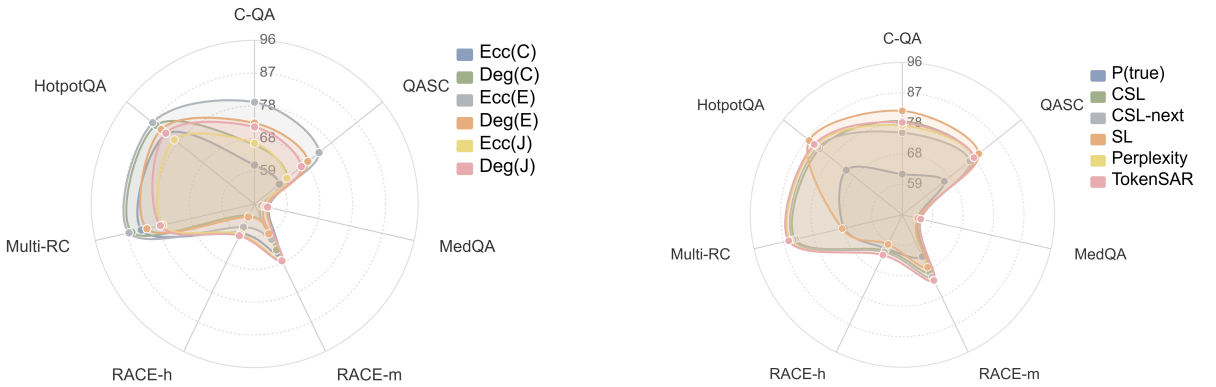


Figure 7: Radar plots of confidence estimation (CE) results for Llama2-7B across 7 datasets, for black-box (left) and white-box (right) methods.

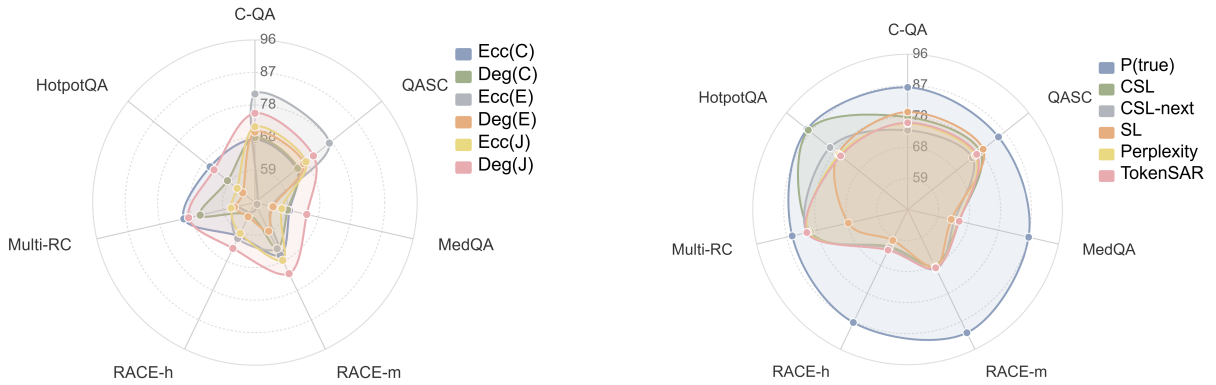


Figure 8: Radar plots of confidence estimation (CE) results for Phi4 across 7 datasets, for black-box (left) and white-box (right) methods.

Dataset	Model	RCE↓						Calibration ECE↓					
		P(true)	CSL	CSL-next	SL	SL(norm)	TokenSAR	P(true)	CSL	CSL-next	SL	SL(norm)	TokenSAR
C-QA	Llama2-7b	0.084386	0.0506	0.041895	0.041267	0.038126	0.034997	0.0102	0.035637	0.041958	0.023881	0.04454	0.027278
	Llama3-8b	0.040614	0.03563	0.068102	0.031902	0.057489	0.038742	0.01871	0.034739	0.050008	0.022294	0.04291	0.026352
	Phi4	0.043731	0.04626	0.046892	0.043771	0.041858	0.03501	0.0583	0.034232	0.055943	0.019535	0.04302	0.030655
	Qwen2.5-32b	0.058105	0.02999	0.044359	0.032513	0.044363	0.059406	0.0369	0.022175	0.046935	0.021905	0.03671	0.021438
QASC	Llama2-7b	0.077448	0.04685	0.078796	0.051258	0.043136	0.045007	0.01119	0.024505	0.037871	0.023245	0.0326	0.023127
	Llama3-8b	0.030627	0.04811	0.117522	0.050664	0.08503	0.043753	0.00894	0.020665	0.038958	0.025687	0.03274	0.020785
	Phi4	0.082518	0.04437	0.116905	0.066942	0.088115	0.049376	0.02122	0.021401	0.0415	0.028242	0.02548	0.033083
	Qwen2.5-32b	0.11997	0.04878	0.062505	0.081237	0.073773	0.041861	0.03096	0.014358	0.040047	0.025111	0.02665	0.023483
MedQA	Llama2-7b	0.181911	0.19254	0.19879	0.191288	0.228796	0.238798	0.00606	0.015623	0.007533	0.00791	0.00669	0.007449
	Llama3-8b	0.028131	0.08939	0.121274	0.207542	0.163158	0.178161	0.0166	0.012721	0.008949	0.03	0.00861	0.010613
	phi4	0.05126	0.09127	0.115648	0.176285	0.119399	0.116273	0.02853	0.046391	0.05184	0.058272	0.05787	0.05535
	Qwen2.5-32b	0.078141	0.06126	0.07314	0.128151	0.088143	0.075015	0.03067	0.020881	0.033491	0.047763	0.03295	0.032673
RACE-m	Llama2-7b	0.16253	0.26130	0.22254	0.13502	0.24317	0.24567	0.00741	0.01935	0.03113	0.01820	0.061396	0.062452
	Llama3-8b	0.05938	0.18003	0.09814	0.12752	0.10252	0.12189	0.05006	0.05534	0.04303	0.04812	0.01986	0.02156
	phi4	0.09689	0.15753	0.09314	0.09689	0.13127	0.13565	0.02585	0.04808	0.02775	0.032335	0.01727	0.01938
	Qwen2.5-32b	0.16940	0.17566	0.17691	0.17691	0.24567	0.25255	0.00695	0.04720	0.05091	0.07564	0.07986	0.08066
RACE-h	Llama2-7b	0.17566	0.31818	0.32318	0.33256	0.31818	0.32256	0.01748	0.02600	0.01719	0.01613	0.021382	0.021339
	Llama3-8b	0.05563	0.22316	0.12189	0.15565	0.163782	0.149404	0.045399	0.01684	0.031577	0.034098	0.030134	0.030341
	phi4	0.08939	0.19566	0.150030	0.13315	0.19316	0.19566	0.019294	0.035576	0.02874	0.03037	0.02238	0.040637
	Qwen2.5-32b	0.24754	0.22254	0.21754	0.21316	0.29505	0.30006	0.016826	0.02004	0.02105	0.022801	0.03156	0.04110
Multi-RC	Llama2-7b	0.047553	0.027493	0.019998	0.148808	0.016250	0.019998	0.027857	0.029057	0.032357	0.122430	0.034530	0.028392
	Llama3-8b	0.028115	0.023746	0.030619	0.105655	0.023761	0.038762	0.029688	0.029818	0.060284	0.122580	0.038933	0.033965
	phi4	0.016250	0.020013	0.026249	0.118199	0.032506	0.023777	0.023944	0.048699	0.066653	0.116689	0.056379	0.044237
	Qwen2.5-32b	0.028136	0.023761	0.019998	0.061248	0.019998	0.032506	0.041890	0.039398	0.065292	0.080128	0.053568	0.043404
HotpotQA	Llama2-7b	0.119451	0.068157	0.086918	0.036881	0.086278	0.081369	0.019784	0.055529	0.057178	0.027370	0.061784	0.059436
	Llama3-8b	0.054423	0.053141	0.104933	0.055020	0.081345	0.050025	0.010844	0.022715	0.057353	0.017567	0.048256	0.033327
	phi4	0.037533	0.075621	0.068105	0.031860	0.075010	0.076914	0.030763	0.031029	0.052968	0.027885	0.057284	0.055820
	Qwen2.5-32b	0.121838	0.036275	0.089724	0.021883	0.065023	0.055031	0.014455	0.030119	0.050357	0.011667	0.021745	0.023468

Table 8: RCE and (calibrated) ECE for white-box methods, across different models and datasets.

Method	n=2	n=3	n=4	n=5	n=6	n=7	n=8
P(true)	85.60	85.52	85.21	85.08	85.19	85.26	85.14
SL	81.85	81.03	80.98	81.04	80.66	80.73	80.85
CSL	78.71	78.30	78.43	78.11	77.83	77.83	78.01
TokenSAR	75.86	75.46	75.46	75.25	74.94	74.97	75.10
Ecc(E)	75.47	75.37	75.31	75.26	75.08	74.92	74.95
Perplexity	75.12	74.99	75.11	74.69	74.51	74.40	74.62
CSL-next	73.81	73.61	73.85	73.29	73.11	72.94	73.19
Deg(J)	73.23	73.15	72.91	72.81	72.63	72.58	72.64
Deg(E)	72.03	71.83	72.00	72.18	72.21	71.94	72.01
Ecc(J)	63.40	63.99	63.85	64.14	63.86	64.16	64.00
Deg(C)	55.31	55.31	55.55	55.41	55.51	55.66	55.52
Ecc(C)	55.04	54.41	55.30	55.12	55.13	55.19	55.19

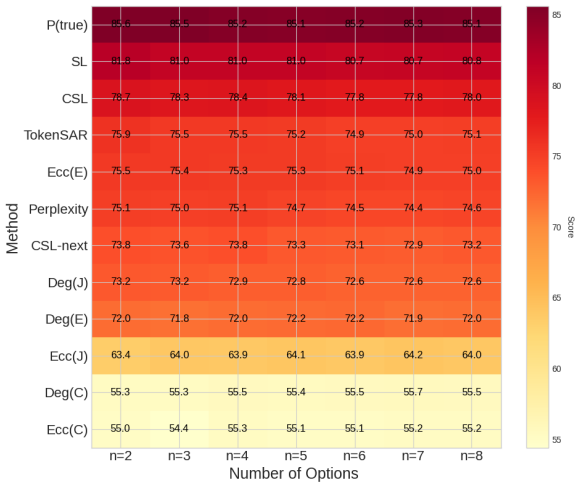


Table 9: AUROC across different numbers of options on the QASC dataset using the Llama3-8B model. The left panel shows the numerical results, while the right panel presents the corresponding heatmap. Changing the number of options has only a minor effect on AUROC values and almost no impact on the ranking. CEBench remains robust to variations in the number of options.

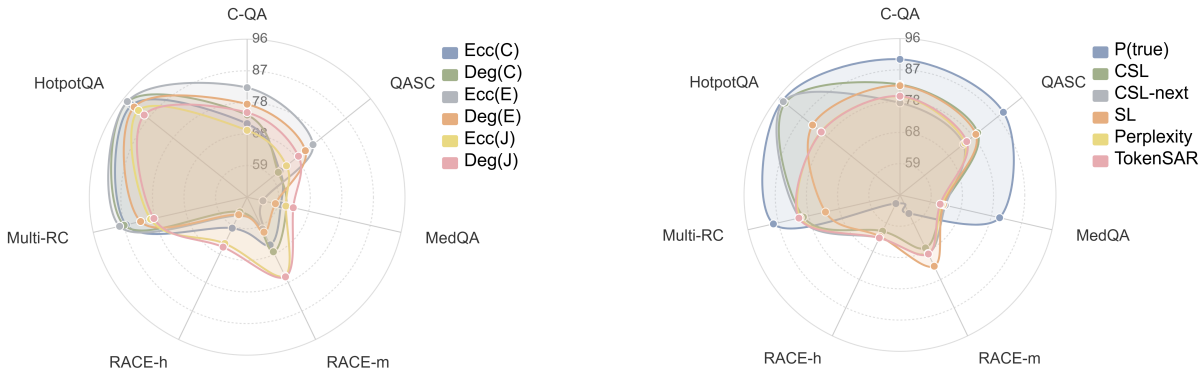


Figure 9: Radar plots of confidence estimation (CE) results for Qwen2.5-32b across 7 datasets, for black-box (left) and white-box (right) methods.