MCQA-Eval: Efficient Confidence Evaluation in NLG with Gold-Standard Correctness Labels

Anonymous ACL submission

Abstract

Large Language Models (LLMs) require robust confidence estimation, particularly in critical domains like healthcare and law where unreliable outputs can lead to significant consequences. Despite much recent work in confidence estimation, current evaluation frameworks rely on *correctness functions*—various heuristics that are often noisy, expensive, and possibly introduce systematic biases. These methodological weaknesses tend to distort evaluation metrics and thus the comparative ranking of confidence measures. We introduce MCQA-Eval, an evaluation framework for assessing confidence measures in Natural Language Generation (NLG) that eliminates dependence on an explicit correctness function by leveraging gold-standard correctness labels from multiple-choice datasets. MCQA-Eval enables systematic comparison of both internal state-based white-box (e.g. logit-based) and consistency-based black-box confidence measures, providing a unified evaluation methodology across different approaches. Through extensive experiments on multiple LLMs and widely used QA datasets, we report that MCQA-Eval provides efficient and more reliable assessments of confidence estimation methods than existing approaches.[1]

1 Introduction

Large Language Models (LLMs) demonstrate strong performance across natural language processing tasks, yet their architectural complexity and limited interpretability can produce unreliable outputs. This presents significant challenges in critical domains such as healthcare, where output errors carry serious consequences. Confidence estimation methods have emerged to quantify output reliability. The field connects closely with uncertainty quantification in natural language generation, as both address output trustworthiness. Current approaches divide into consistency-based methods, which analyze agreement across multiple outputs, and internal-states methods that leverage model-specific features like output probabilities. Despite advances in these approaches, developing robust evaluation frameworks remains a central challenge.

Current evaluation frameworks for NLG confidence measures rely on correctness labels to compute metrics such as AUROC and AUARC. These frameworks follow a three-step process: generating model predictions, labeling correctness via a function $f(\cdot)$, and calculating metrics. This label-dependent approach faces several constraints. While human evaluation provides reliable correctness ground truth, it cannot scale to large datasets. Metrics based on reference matching, such as BLEU and ROUGE, fail to recognize semantically equivalent responses phrased differently. LLM-based evaluators offer greater capability but remain noisy and may introduce systematic biases, such as favoring responses generated by themselves or similar LMs (Panickssery et al., 2024), or preferring longer responses (Lin et al., 2022). Moreover, running such evaluators could be expensive.

Flaws in the correctness function $f(\cdot)$ propagate through the evaluation pipeline, affecting metrics like AUROC. This sensitivity becomes particularly problematic when comparing confidence estimation methods with similar performance. Such limitations underscore the need for evaluation frameworks that establish correctness more reliably.

In this paper, we propose MCQA-Eval, a simple, efficient yet effective evaluation framework that eliminates the dependence on unreliable correctness functions. *The key insight is to leverage multiple-choice question-answering (QA) datasets, which inherently provide gold-standard answer choices at no cost.* With these definitive labels, our framework bypasses the ambiguity of determining correctness via correctness function $f(\cdot)$

[1] Code and data will be released upon publication.

and ensures an objective assessment of confidence estimation methods. Rather than replacing existing evaluation pipelines, our framework complements them, offering an additional lens to assess the discriminative power of confidence estimation methods. Fig. 1 shows how our proposal (green) and the existing evaluation pipeline (blue) differ, yet complement each other. Our contributions are summarized as follows:

- We demonstrate that commonly used evaluation methods for NLG confidence measures are sensitive to noise in correctness labels, which can lead to misleading conclusions about evaluation metrics and rankings of different confidence estimation approaches.
- We propose a simple yet effective method that utilizes multiple-choice QA datasets to evaluate confidence measures, supporting both internal-states-based white-box and consistency-based black-box methods.
- Extensive experiments across recent LLMs and QA datasets verify that `MCQA-Eval` produces stable evaluations broadly consistent with existing methods, while eliminating the need for expensive correctness functions.

2 Related Work

Confidence Estimation Confidence estimation is fundamental to machine learning, providing mechanisms to assess model reliability and guide decision-making across tasks. Early confidence estimation research concentrated on classification settings, where confidence scores enabled Selective Classification (Geifman and El-Yaniv, 2017; El-Yaniv et al., 2010; Feng et al., 2022)—allowing models to abstain from low-quality predictions. The rapid advancement of NLG and LLMs has brought renewed attention to confidence estimation. While NLG poses unique challenges due to semantic invariance and vast output spaces (Kuhn et al., 2023), recent works have advanced the field by measuring similarities among sampled responses (Lin et al., 2024b) and deriving measures from LMs' internal states (Malinin and Gales; Lin et al., 2024a; Azaria and Mitchell, 2023).

A related aspect is calibration. While extensively considered in classification (Zhang et al., 2020; Kull et al., 2019; Ma and Blaschko, 2021), it has received lass attention in NLG. Since the distribution of confidence scores could vary significantly across different methods due to their underlying principles (Geng et al., 2023; Da et al., 2024), calibrated confidence measures align better with human intuition for probabilities and are more interpretable (Guo et al., 2017; Cosmides and Tooby, 1996). While this paper focuses on evaluating confidence estimation methods, the same framework could be applied to evaluate future NLG calibration methods. We demonstrate this by including results using common calibration metrics like Expected Calibration Error (ECE).

Evaluation of Confidence Measures While confidence estimation has received considerable attention, the evaluation of confidence measures remains under-explored. Many evaluation methods have been adapted from the classification literature, including Expected Calibration Error (ECE) (Guo et al., 2017; Xiong et al., 2024) and Area Under the Receiver Operating Characteristic Curve (AUROC) (Kuhn et al., 2023). These metrics assess the relationship between confidence scores and prediction accuracy, typically requiring high-quality correctness labels for the evaluated responses.

However, obtaining reliable correctness labels in NLG is challenging due to factors such as semantic variability and ambiguity in open-ended tasks (Novikova et al., 2017). Unlike classification where correctness is well-defined, NLG correctness is often determined through human annotation, LLM-based judges, or similarity-based comparisons between the generated and reference answers. These approaches are costly and often unreliable, as correctness judgments can be subjective and inconsistent (Gatt and Krahmer, 2018).

Recent works have attempted to mitigate these limitations. To allow for non-binary correctness measures, Rank Calibration Error (RCE) (Huang et al., 2024) and AUARC (Nadeem et al., 2009; Lin et al., 2024b) were introduced, both of which leverage continuous correctness scores. Other approaches focus on improving correctness scores themselves. For example, Lin et al. (2024a) aggregates predictions from multiple LLM-based judges and takes a consensus to enhance reliability.

Unlike these methods, our proposed framework completely circumvents the need for correctness labels, making it more robust and scalable for evaluating confidence measures in NLG.

Applications of Confidence Measures Confidence measures play a crucial role in several downstream research areas in NLG, particularly in conformalized NLG and selective generation or gen-

eration with abstention. Stemming from Conformal Prediction (Papadopoulos et al., 2007), in the context of NLG, conformalized methods typically aim to create a set of generation that satisfies a particular user-defined quality goal (e.g. "correct answers") (Quach et al., 2023; Gui et al., 2024; Lee et al., 2024; Yadkori et al., 2024), or providing factual guarantees basing on parts of the generation (Cherian et al., 2024; Mohri and Hashimoto, 2024). Selective generation or generation with abstention, on the other hand, deals with broader considerations that involve refraining from generating if the confidence score is low, with goals like improving the accuracy on the non-rejected portion (Ren et al., 2023b; Cole et al., 2023). Good confidence measures that can distinguish high and low-quality generations are key ingredients to all these research directions, and our paper aims to provide a better evaluation framework for researchers to identify such confidence measures.

3 Confidence Estimation for NLG

First, 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. For free-form NLG datasets, we typically have reference answers $A = a_1, \ldots, a_m$ alongside $\mathbf{x}$. 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:

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

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.

3.1 Confidence Estimation Methods

Existing confidence estimation methods can be broadly divided into two categories: Consistency-based black-box methods and internal state-based white-box methods[2].

Black-Box Methods leverage response consistency across LLM generations (Lin et al., 2024b; Manakul et al., 2023). Higher consistency among generated responses indicates higher confidence in $\mathbf{s}$. These methods first compute pairwise response similarities, then derive confidence from the similarity matrix. For similarity computation, existing methods use Jaccard similarity, NLI models (He et al., 2021), and BERTScore (Zhang* et al., 2020) for similarity computation.

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 (Lin et al., 2024a), 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 verbalized confidence (`P(true)`) (Xiong et al., 2024) elicits explicit "True" or "False" predictions. 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. Note that uncertainty quantification in NLG is a closely related research direction, yet differs in a key way: uncertainty characterizes the predictive distribution rather than a specific $\mathbf{s}$. For more details of this distinction, see Lin et al. (2024b).

3.2 Existing 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). In other words, confidence measures guide selection towards predictions that are likely to be correct.

This idea extends naturally to NLG, where confidence measures are used to guide selective generation or generation with abstention. Assuming a given *correctness function* (Huang et al., 2024) $f(\mathbf{s}; \mathbf{x}) \in \{0, 1\}$, which tells us whether a response

[2]We consider logits as an internal states here.

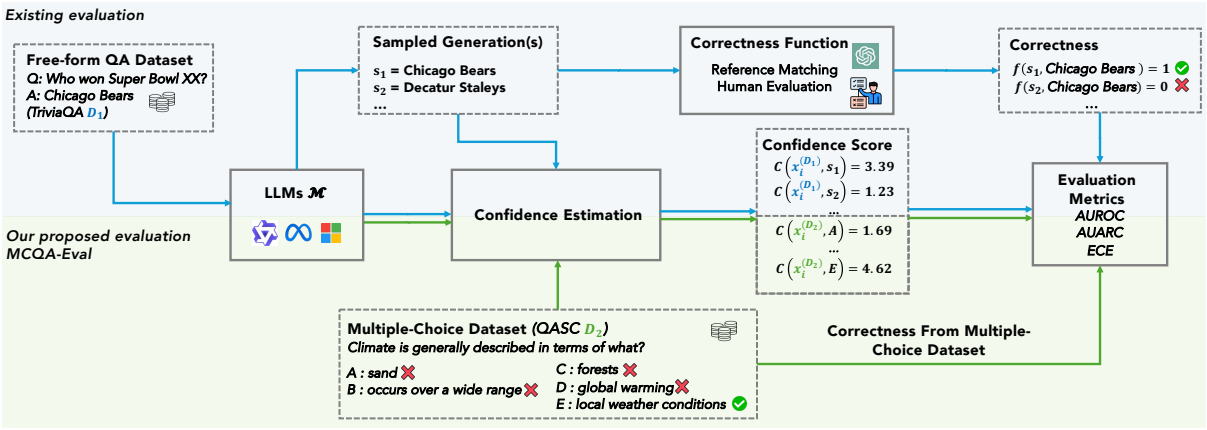


Figure 1: Illustration of the existing evaluation framework (blue) vs our proposed `MCQA-Eval` (green). Unlike existing frameworks, we avoid the costly and unreliable correctness function module by using multiple-choice datasets. This requires slight modification to the confidence estimation steps, which is elaborated in Section 4.

is good or correct[3], several evaluation metrics are used to assess confidence measures for NLG:

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

$$\int_{-\infty}^{\infty} \text{TPR}(t)\, d\text{FPR}(t), \tag{2}$$

where $\text{TPR}(t)$ ($\text{FPR}(t)$) 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):

$$\int_{-\infty}^{\infty} \text{Accuracy}(t)\, d\text{Coverage}(t), \tag{3}$$

where $\text{Accuracy}(t) = \mathbb{E}\{f(\mathbf{s})|C(\mathbf{s}) > t\}$ and $\text{Coverage}(t) = \mathbb{P}\{C(\mathbf{s}) > t\}$. 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):

$$\mathbb{E}\Big[|\mathbb{E}[f(\mathbf{s})|C(\mathbf{s})] - C(\mathbf{s})|\Big]. \tag{4}$$

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

[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.

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

$$\mathbb{E}_C\Big[|\mathbb{P}_{C'}\{reg(C') \geq reg(C)\} - \mathbb{P}_{C'}\{C' \leq C\}|\Big] \tag{5}$$

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.

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, in NLG, correctness is inherently difficult to determine, unless $\mathbf{s}$ exactly matches one of the reference answers, which is rare except for simple tasks. Currently, correctness is often assessed using human evaluation or similarity-based methods:

Human Evaluation This remains arguably the most reliable approach. Human evaluation is either used on smaller datasets (Ren et al., 2023b) or to validate automated correctness functions (Kuhn et al., 2023; Lin et al., 2024b,a), but is expensive and unscalable for large-scale dataset evaluation.

Similarity-Based Methods In practice, correctness is 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 $\{0, 1\}$:

$$f(\mathbf{s}, \mathbf{x}) = \begin{cases} 1, & \text{if } sim(\mathbf{s}, A) > \tau \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Specifically, there are two common approaches for computing similarity. **Reference Matching** relies on lexical-based similarity metrics such as ROUGE and BLEU (Hu and Zhou, 2024; Aynetdinov and Akbik, 2024; Kuhn et al., 2023), which often fail to recognize semantically equivalent answers which are phrased differently. **LLM Judgment** uses a LLM as an evaluator (Ren et al., 2023b; Li et al., 2024; Tan et al., 2024) and is more flexible. However, such methods are computationally expensive and are still not fully reliable. Recent studies indicate that machine-based correctness evaluation sometimes only has an accuracy of 85% on popular datasets (Kuhn et al., 2023; Lin et al., 2024a).

3.3 Limitations of Existing Methods

Flaws in the correctness function inevitably affect downstream evaluation metrics such as AUROC and thus our conclusions about different confidence measures. In this section, we illustrate the limitations of current confidence evaluation methods from two angles: the impact of threshold sensitivity and the inherent noise of similarity measures.

Case Study 1: Threshold Sensitivity 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. (6). 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. 2 suggests that $\texttt{Ecc}(C)$, for example, could either rank at the top or the bottom depending on τ.

Case Study 2: Similarity Noise Correctness labels, whether derived from human evaluation, LLM-based scoring, or reference matching, are inherently noisy. For instance, within LLM-based judgments, correctness labels can fluctuate due to factors such as prompt variations and how the LLM judges were designed and trained. Echoing prior observations, Fig. 3 shows examples where LLM judgments could either differ between different LLM judges or between different calls to the same LLM judge. Lin et al. (2024a) proposes to

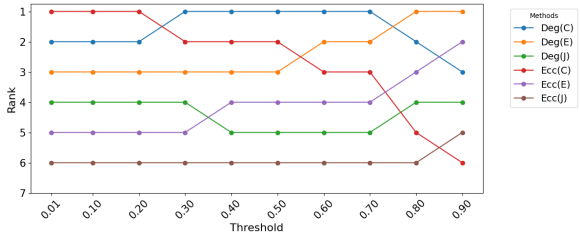


Figure 2: The AUROC ranking of black-box confidence measures (on LLaMA2-13B and CoQA) is sensitive to the threshold τ.

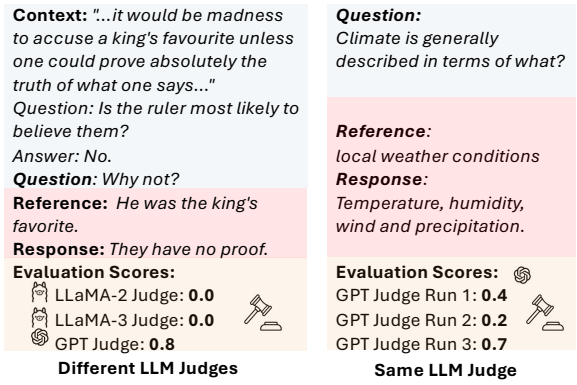


Figure 3: Using LLM judges as f, while flexible, still has inherent noise. Different LLMs disagree on whether a response is correct (left). Even the same LLM (GPT, right) could deliver different opinions simply due to the randomness in generation.

set the correctness function f as the consensus of multiple LLMs, which improves the reliability of the correctness of responses LLMs agree on. However, simply ignoring the disagreement could also introduce systematic selection bias.

To systematically analyze this effect, we simulate correctness label noise with Gaussian noise and analyze its effects. We modify the correctness function as:

$$\tilde{f}(\mathbf{s}; \mathbf{x}) = Sigmoid(logit(f(\mathbf{s}; \mathbf{x})) + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \quad (7)$$

As shown in Table 1, increasing noise levels can lead to significant instability in ranking different confidence measures. Note that our simulation likely *underestimates* the issue, because the noise in Eq. (7) is unbiased and does not reflect systematic bias that may favor certain confidence measures (Lin et al., 2022).

While it might be theoretically possible to estimate the noise level and its propagation to errors on metrics like AUROC, this requires strong assumptions (e.g. Eq. (7)), extensive human evaluation, and replication across LLM judges and datasets. This fragility in existing evaluation methods moti-

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.

vates our framework, which eliminates dependence on uncertain correctness functions.

4 MCQA-Eval: A framework for Assessing Confidence Estimation

At a high level, existing evaluation frameworks for $C_{\mathcal{M}}$ includes three main steps (blue path in Fig. 1):

1. Generate $\mathbf{s}$ from $\mathcal{M}$ given the input $\mathbf{x}_i$.
2. Determine the correctness label of $\mathbf{s}$ using the function $f(\cdot, \mathbf{x})$.
3. Compute evaluation metrics such as AUROC. A higher metric value indicates that $C_{\mathcal{M}}$ is a "better" confidence estimation.

The main limitation of this general pipeline lies in f in step 2. Existing evaluation frameworks all implicitly assume step 1—that the confidence measure $C_{\mathcal{M}}$ must apply to generated sequences $\mathbf{s}$. While this might hold for consistency-based uncertainty measures, where response divergence indicates uncertainty, it does not extend to confidence measures. In other words, we could relax step 1 in order to improve step 2.

Our main proposal in this paper is to adapt multiple-choice datasets to evaluate confidence measures designed for free-form NLG. Unlike free-form NLG datasets, multiple-choice datasets provide inherent correctness values for options, eliminating the need for an explicit correctness function. If we simply "pretend" that these options are free-form generations from the base LM, we can directly evaluate the confidence measure quality. As Fig. 1 shows, the approach differs from existing evaluation pipelines only in applying confidence estimation methods to multiple-choice options.

Consider the QASC (Khot et al., 2020) dataset as an example, each problem comes with a question $\mathbf{x}$ and a few choices, $o_1, \ldots, o_K$. Unlike what such datasets were designed for, we re-format the input prompt as a free-form NLG question, as illustrated in Fig. 4, as if the base LLM generated each option itself, in different runs. In what follows, we first explain explain slight nuances in applying internal state-based white-box confidence measures as well as consistency-based black-box ones.

Logit or Internal State-Based Measures typically examine the internals of a LM when it generates a particular response. The nature of the free-form generation task allows us to simply plug-in the option o_i into the corresponding location of the prompt, and extract similar information that allows us to evaluate the confidence[4].

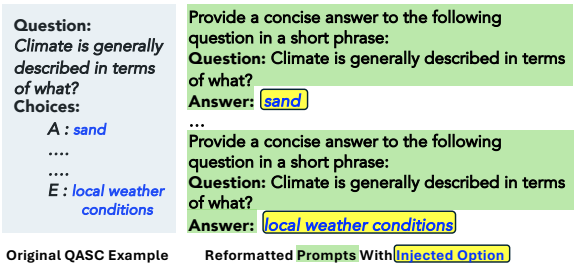


Figure 4: We reformat each option from the multiple-choice question (left), by injecting the option to a free-form QA prompt. One could typically apply any confidence estimation method by treating this option as if it was generated by the base LM. For black-box confidence measures that require additional responses, we only feed the prompt to the base LM.

Algorithm 1 Consistency-based Confidence Estimation for Any Sequences

Input: $\mathbf{x}$, $\mathcal{M}$, candidate sequences $A = \{a_1, \ldots, a_K\}$
Output: $\{C_{\mathcal{M}}(\mathbf{x}, a_1), \ldots, C_{\mathcal{M}}(\mathbf{x}, a_K)\}$
Generate $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ using $\mathcal{M}$ for question $\mathbf{x}$
Compute pairwise similarity matrix M of S.
for each $a_i \in A$ **do**
    Compute a new similarity matrix M_i of $S \cup \{a_i\}$, reusing M.
    Compute confidence score $C_{\mathcal{M}}(\mathbf{x}, a_i)$ using M_i.
end for
return $\{C_{\mathcal{M}}(\mathbf{x}, a_1), \ldots, C_{\mathcal{M}}(\mathbf{x}, a_K)\}$

Consistency-based Confidence Measures Unlike logit-based or internal-state-based measures, 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. Consider methods from Lin et al. (2024b) as an example. To preserve the integrity of the predictive distribution, we first sample n responses from $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$ as usual, 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.

[4]In fact, this was the practice to compute SL for actual generations. For example, https://github.com/lorenzkuhn/semantic_uncertainty/blob/main/code/get_likelihoods.py and https://huggingface.co/docs/transformers/perplexity.

Remarks Our proposal relaxes step 1 at the beginning of this section, allowing for $\mathbf{s}^* = o_i$ not sampled from $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$. This is not to be misunderstood as a proposal to *replace* the current pipeline (Section 3.2)—rather, it is *complementary*. The rationale is that if a good confidence measure predicts the correctness well, it should perform well in *both* evaluation frameworks. In fact, any $o_i \in \Sigma^*$ that does not violate the generation configuration, has a non-zero probability to be sampled from $\mathcal{P}(S; \mathbf{x}, \mathcal{M})$, and a robust confidence measure should be expected to model it well.

5 Experiments

We demonstrate the advantages 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., 2024). These models were specifically pretrained on question-answering tasks, which minimizes irrelevant responses. We include various model sizes for a comprehensive analysis.

Datasets We select five multiple-choice datasets with varying levels of complexity from different domains, including 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 dataset consists of independent questions with a set of answer options, where exactly one option is correct. Evaluating different LLMs on datasets from different domains and diverse levels of difficulty allows for a more comprehensive assessment of model performance across a wide range of scenarios. Table 2 provides an overview of these datasets and the number of questions we use, with detailed descriptions in Appendix A.1. The exact prompt formulation for each dataset is provided in Appendix A.2.

Confidence Estimation Methods We compare six black-box and six white-box methods. The selected methods represent commonly used confidence estimation baselines. The six **black-box** measures evaluated in our experiments are:

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

Table 2: Overview of datasets used in this paper.

- `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.

Unlike black-box methods, **white-box** measures directly use the multiple-choice options as evaluation responses. We implement six white-box confidence estimation baselines, as introduced in Section 3.1: `SL`, `Perplexity`, `TokenSAR`, `CSL`, its variant `CSL-Next`, and `P(true)`.

Metrics Following previous works, we use AUROC as our primary metric[5]. Our framework can also be applied to evaluate confidence calibration. We include additional results in Appendix B, reporting RCE and calibration ECE metrics.

Additional details of our experiment can be found in Appendix A.

5.2 Experimental Findings

This section summarizes our main experimental findings, with detailed results in Appendix B.

Comparison With Existing Evaluation Methods We first compare our evaluation method with the existing pipeline (Baseline) using the QASC dataset. For the baseline, we use gpt-4o-mini to obtain correctness labels (by comparing the generation with the correct option). Table 3 shows that varying the threshold significantly impacts the ranking of both black-box and white-box confidence estimation methods. Additionally, 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.

[5]Responses sampled for black-box methods' are excluded from AUROC calculations due to uncertain correctness.

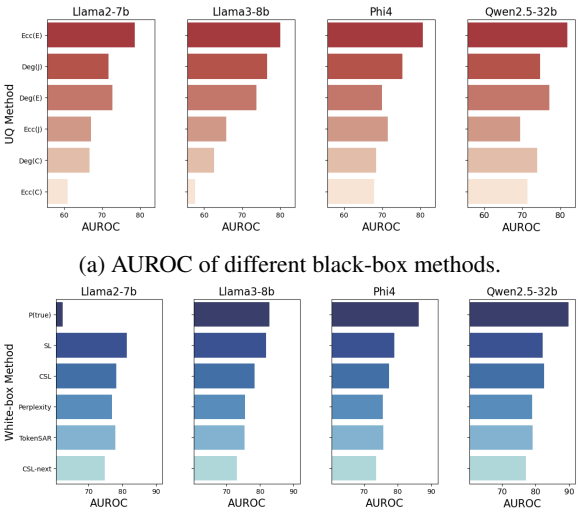
	τ	Ranking 1	2	3	4	5	6
		Black-box					
	0.5	Deg(C)	Deg(E)	Ecc(E)	Ecc(C)	Deg(J)	Ecc(J)
	0.7	Deg(C)	Deg(E)	Ecc(C)	Ecc(E)	Deg(J)	Ecc(J)
Baseline	0.9	Ecc(E)	Deg(E)	Deg(C)	Deg(J)	Ecc(J)	Ecc(C)
`MCQA-Eval`	N/A	Ecc(E)	Deg(E)	Deg(J)	Deg(C)	Ecc(J)	Ecc(C)
		White-box					
	0.5	TokenSAR	Perplexity	SL	CSL	CSL-Next	P(true)
	0.7	TokenSAR	Perplexity	CSL	CSL-Next	SL	P(true)
Baseline	0.9	SL	TokenSAR	Perplexity	CSL	CSL-Next	P(true)
`MCQA-Eval`	N/A	SL	TokenSAR	Perplexity	CSL	CSL-Next	P(true)

Table 3: We analyze how existing LLM-based evaluation methods rank black-box and white-box approaches by varying τ from 0.9 to 0.5. `MCQA-Eval` aligns with the rankings at $\tau = 0.9$, yet requires no overhead for the correctness function.

On the other hand, `MCQA-Eval` aligns with baseline ranking at $\tau = 0.9$. While it is unclear in this case which τ reflects the "most reliable" ranking, this experiment suggests that `MCQA-Eval`'s conclusion is consistent with existing pipeline. However, unlike the Baseline, it does not require the costly correctness function(s), thereby reducing computational costs and enabling scalable evaluation.

Comparison Across LLMs We compare confidence measures across LLMs on the same dataset via `MCQA-Eval`. As shown in Fig. 5 (with additional results available in the Appendix B), larger LLMs tend to achieve better performance across different confidence estimation methods, reflecting their broader pretraining exposure. The relatively ranking of various confidence measures stay mostly stable. Interestingly, unlike some prior results (Lin et al., 2024b; Vashurin et al., 2024), `P(true)` performs very well except for Llama2-7b. We hypothesize that this is due to improvement in recent LLMs' abilities in general, which is similar to the conjecture in (Vashurin et al., 2024). This hypothesis is partially supported by the fact that `P(true)` performs increasingly well as the base LM becomes more sophisticated.

Comparison Across Datasets We compare confidence measures across datasets of varying difficulty with `MCQA-Eval`. Fig. 6 illustrates these results for Phi4-14B. In general, different confidence measures exhibit larger performance gap on simpler datasets such as C-QA, and smaller on more professional datasets like MedQA. The general poor performance on harder datasets could be attributed to limited capability of the base LM (in generating additional responses for black-box measures, or in supplying the base logits for white-box measures. For black-box measures, similarity metrics like NLI and Jaccard may also provide limited distinguishing power. We recommend selecting datasets with difficulty levels that align with the capabilities of the language model, making performance differences more discernible.



(a) AUROC of different black-box methods.

(b) AUROC of different white-box methods.

Figure 5: (a) and (b) show the performance of 4 different LLMs and 12 different confidence estimation methods on the C-QA dataset. A higher AUROC indicates better performance.

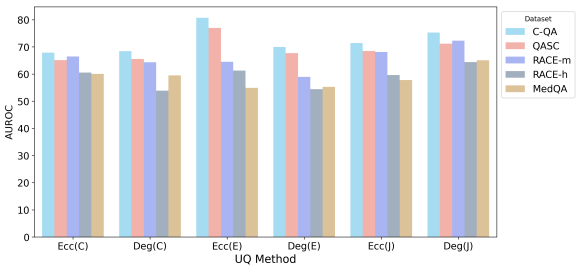


Figure 6: The performance of Phi-4 using black-box methods across different datasets.

6 Conclusion

In this paper, we propose `MCQA-Eval`, a simple framework using multiple-choice QA datasets to evaluate confidence measures for natural language generation. We first highlight the unreliability of widely used *correctness functions* in existing evaluation frameworks. To address this, we propose an alternative approach that reformulates multiple-choice questions into a free-form QA prompt, enabling a more efficient evaluation with higher-quality correctness labels. Experiments across diverse datasets and state-of-the-art LLMs demonstrate that `MCQA-Eval` produces consistent results aligned with prior research findings, while eliminating dependence on costly correctness functions.

Limitations

While our proposed evaluation framework avoids the use of correctness functions, offering speed and reliability, it also has its limitations. As noted in Section 4, `MCQA-Eval` should not serve as the *only* evaluation method. Bypassing response generation provides no guarantee that injected options resemble what would otherwise be generated by the base LM. If the goal is to evaluate confidence measures *for a specific LM and its generation*, then this generation step by definition should not be skipped. Further, certain trained confidence measures (e.g. linear probes on the LM's internal states) might not generalize as well to injected options, and may perform systematically worse in `MCQA-Eval` than in the current framework (although one may argue that generalizability should be part of the evaluation to begin with). Finally, `MCQA-Eval` 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, et al. 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.

John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73.

Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. 2024. Llm uncertainty quantification through directional entailment graph and claim level response augmentation. *arXiv preprint arXiv:2407.00994*.

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.

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

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.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

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*.

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.

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*.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

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*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

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. 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*.

Xingchen Ma and Matthew B. Blaschko. 2021. Meta-cal: Well-controlled post-hoc calibration by ranking. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7235–7245. PMLR.

Andrey Malinin and Mark Gales. 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.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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*.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. 2007. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395. IEEE.

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.

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.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 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, et al. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627*.

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, et al. 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, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. 2020. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11117–11128. PMLR.

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 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 5 given choices. The dataset consists of 1,221 test questions.

- **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.

A.2 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.3 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.4 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 4 and 5 show the AUROC and AUARC for black-box and white-box confidence measures, respectively. Similarly, Tables 6 and 7 present RCE and ECE results. Note that all ECE 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 histogram binning method (Zadrozny and Elkan, 2001) on all methods.

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

Table 4: 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

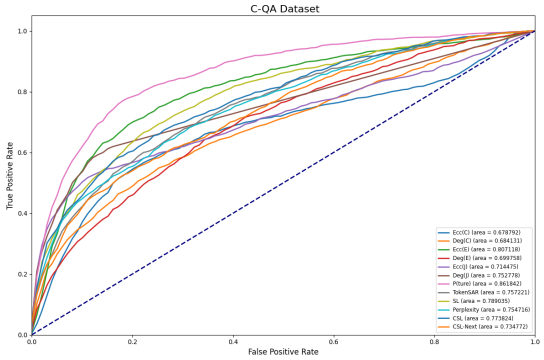
Table 5: AUROC and AUARC for white-box methods, across different models and datasets

B.2 Additional Visualizations for ROC Curves

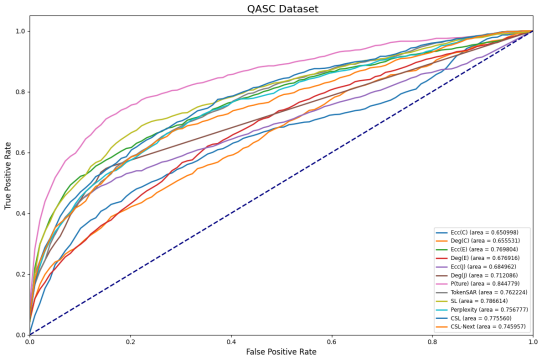
Fig. 7 presents the ROC curves for Phi4-14B. `P(true)` achieves much better performance than other confidence measures on the more challenging datasets, likely because Phi4-14B is a relatively advanced model. On the easier datasets, where we could observe a bigger performance gap between different confidence 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 `MCQA-Eval`.

C AI Assistant Usage

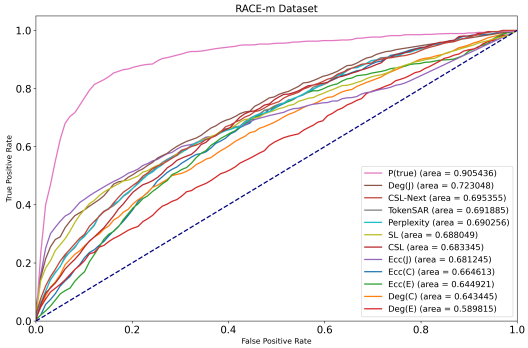
We used GPT for grammar checking and Copilot as an assistive tool.



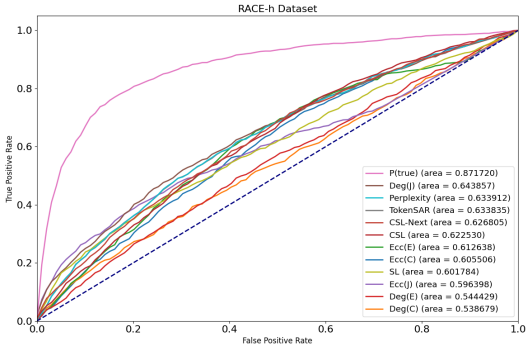
(a) C-QA Dataset



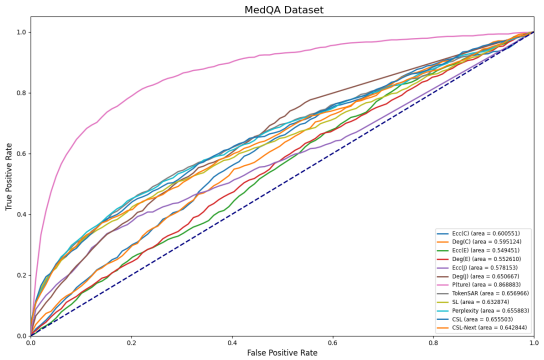
(b) QASC Dataset



(c) RACE-m Dataset



(d) RACE-h Dataset



(e) MedQA Dataset

Figure 7: 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	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

Table 6: RCE and (calibrated) ECE for black-box methods, across different models and datasets

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

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