

Calibration Is Not Enough: Evaluating Confidence Estimation Under Language Variations

Yuxi Xia^{1,2}, Dennis Ulmer³, Terra Blevins⁴, Yihong Liu⁵,
Hinrich Schütze⁵, Benjamin Roth^{1,2}

¹Faculty of Computer Science, UniVie Doctoral School Computer Science,

²Faculty of Philological and Cultural Studies, University of Vienna, Austria

³ILLC, University of Amsterdam, Netherlands

⁴Khoury College of Computer Sciences, Northeastern University, USA

⁵LMU Munich, Munich Center for Machine Learning (MCML), Germany

Correspondence: yuxi.xia@univie.ac.at

Abstract

Confidence estimation (CE) indicates how reliable the answers of large language models are and impacts user trust and decision-making. Existing evaluations mainly concern the alignment between confidence and correctness, but ignore the variability of language: confidence estimates should remain consistent under semantically equivalent prompts or answer variations, while changing when answer meaning differs, as this may indicate a change in correctness. Therefore, we introduce a novel evaluation framework based on three complementary properties: **robustness** to prompt perturbations, **stability** across semantically equivalent answers, and **sensitivity** to semantically different answers. We show that these metrics are largely independent from existing CE metrics, and that common CE methods often fail on them: while most methods achieve high robustness and stability, they struggle to distinguish semantically different answers, potentially because they do not effectively leverage generation-side information. Overall, our framework exposes overlooked limitations of current CE evaluations and provides guidance for selecting confidence estimators for real-world applications.

1 Introduction

Large Language Models (LLMs) are effective at answering fact-based questions across domains such as medicine, law, and education (Alqahtani et al., 2023; Alfertshofer et al., 2024; Xiao et al., 2025). In these high-stakes settings, it is paramount to improve user trust and support safer human-AI interactions (Kadavath et al., 2022; Zhang et al., 2024) to avoid harm and achieve positive outcomes. To this end, *confidence estimation* (CE; Si et al., 2023; Tian et al., 2023; Chen and Mueller, 2024) can help by indicating how certain and therefore reliable a model’s answer is.

Existing works on CE (Guo et al., 2017; Osawa et al., 2019; Ovadia et al., 2019; Groot and Valdene-



Figure 1: Example for different failure cases of confidence estimators under language variation in either the input prompt or the generated response of the LLM.

gro Toro, 2024; Ni et al., 2025; Xiong et al., 2024) mainly focus on evaluating calibration (whether stated confidences match expected model accuracy), or discrimination (whether it ranks right answers higher than wrong ones). This overlooks a key property of language (models): Language is extremely variable, and while the surface form, i.e. a particular choice of words, might change, the meaning of an expression can stay untouched (Baan et al., 2023). However, a method may be well-calibrated yet produce confidence values that fluctuate under minor prompt variations (Sclar et al., 2024; Xia et al., 2025a), or fail to adapt to answers with different surface forms and / or meanings, as we exemplify in Figure 1.

This paper presents a novel evaluation paradigm for LLM CE methods that takes language variation into account, including three novel dimensions: **Robustness (P-RB)** measures whether CE methods can ignore semantically irrelevant changes to the prompt and output the same confidence score. **Stability (A-STB)** requires CE methods to assign consistent confidences to semantically equivalent an-

swers, and **sensitivity (A-SST)** measures whether confidence estimates change accordingly when the answer meaning changes. In summary, we make four main contributions: (1) we propose three novel and complementary metrics that go beyond calibration to evaluate CE methods under realistic language variation; (2) we conduct a large-scale evaluation of 10 diverse CE methods across 11 LLMs from five model families; (3) we show empirically that while most confidence estimators are robust and stable, they exhibit a striking lack of sensitivity, potentially because generation-side information is not effectively utilized; (4) we provide practical guidance on selecting confidence estimators for target applications and open-source the code.¹

2 Related Work

Confidence Estimation in LLMs. CE has emerged as a crucial research direction for improving user trust and model interpretability (Tian et al., 2023; Chen et al., 2024b). Currently, there are 5 common approaches (Geng et al., 2024; Shorinwa et al., 2025; Vashurin et al., 2025; Testoni and Calixto, 2026) to estimate LLM confidence: (1) *logit-based*, such as token-level probability (Zhu et al., 2023; Huang et al., 2025); (2) *linguistic confidence* (Lin et al., 2022; Xiong et al., 2024; Ulmer et al., 2025), which elicits verbalized confidence directly from the model; (3) *auxiliary-based* methods (Ulmer et al., 2024; Xia et al., 2025a; Zhu et al., 2026), which employ an auxiliary model trained to predict the confidence of a target LLM based on answer content and predicted correctness; (4) *internal states-based*, Sriramanan et al. (2024) derive model uncertainty from attention heads, other methods (Liu et al., 2024; Burns et al., 2023; Kadavath et al., 2022) train probes which comprise smaller, often linear classifiers on internal model representations. Lastly, (5) *consistency-based methods* (Lin et al., 2024; Manakul et al., 2023; Xiong et al., 2024) rely on multiple stochastic generations obtained by prompting the LLM with loose temperatures and diverse prompts. Confidence is then estimated based on the similarity or frequency of the generated outputs. Since our proposed metrics themselves rely on comparing the confidence of multiple different responses under perturbation, consistency-based methods are not applicable in our framework.

¹Code is available at https://github.com/Yuuxii/Evaluate_confidence_estimation_under_language_variations.

Evaluation of Confidence Estimators. LLMs are typically evaluated using task-level metrics (Le Bronnec et al., 2024; Moe et al., 2025; Xia et al., 2025b), which are not suitable for CE methods. Instead, calibration metrics, including ECE (Guo et al., 2017) and its variants (Kumar et al., 2019; Nixon et al., 2019; Kirchenbauer et al., 2022; Błasiok and Nakkiran, 2024), and Brier score (Brier, 1950), are used to measure the alignment between predicted confidence and actual correctness (Groot and Valdenegro Toro, 2024; Zhou et al., 2024; Stengel-Eskin and Durme, 2023). Ranking-based metrics such as AUROC and AUPRC (Bradley, 1997) further evaluate whether CE methods can distinguish correct from incorrect responses (Vazhentsev et al., 2022; Tian et al., 2023; Ulmer et al., 2024), in-distribution from out-of-distribution inputs (Ovadia et al., 2019; Charpentier et al., 2020; Ulmer et al., 2022), or whether loss correlates with decreased confidence (Ulmer et al., 2022). However, existing metrics do not assess the robustness of CE methods to prompt perturbations, as studied for LLM outputs, for instance by ProSA (Zhuo et al., 2024), or their ability to distinguish between (dis)similar responses.

3 Background

Notation. We use $\mathbf{x}$ to denote an input sequence, $\mathbf{t}$ for a language model prompt, and $\mathbf{y}$ for a model response. We write $\mathbf{y}^*$ for the intended or correct response, and in cases where there are multiple correct responses, we denote the corresponding set by $\mathcal{Y}^*$. Then, we write a confidence estimator as a function f that predicts a confidence value c conditioned on $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{t}$, written $c = f(\mathbf{y}, \mathbf{x}, \mathbf{t})$.²

3.1 Preliminaries

One common metric for evaluating CE is the expected calibration error (Guo et al., 2017):

$$\text{ECE} = \mathbb{E}[|p(y = y^* | c) - c|], \quad (1)$$

where the expectation is typically approximated by binning N test predictions into M buckets of equal size by their predicted confidence. The ECE has been criticized for not being a *proper scoring rule*, i.e. trivial predictors other than the true distribution can minimize it (Liu et al., 2023). An actual proper

²So e.g. for sequence likelihood, f involves passing $\mathbf{x}$, $\mathbf{t}$ to the LLM to obtain token probabilities. For e.g. an auxiliary model, f , can be directly equated with the secondary model.

	ECE ↓	Brier ↓	AUROC ↑	A-STB ↑	A-SST ↑
f_{oracle}	0.00	0.00	1.00	1.00	1.00
f_{constant}	0.01	0.17	0.83	1.00	0.00
f_{random}	0.33	0.33	0.67	0.72	0.18
f_{prior}	0.00	0.25	0.50	1.00	0.00

Table 1: Simulation results. Two of our own metrics are stability (A-STB) and sensitivity (A-SST).

scoring rule is the Brier score (Brier, 1950):

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (c_i - \mathbf{1}(y_i = y_i^*))^2. \quad (2)$$

Both ECE and Brier score evaluate whether the confidence score aligns with (expected) correctness. However, one might also be interested in whether confidence helps us to distinguish between correct and incorrect predictions. This is commonly evaluated through the area under the receiver-operating characteristic curve (AUROC; Bradley, 1997), where error detection is framed as a binary classification task, and the ratio of the true to false positive rate is measured under all possible thresholds. However, none of these metrics incorporates the variability of natural language.

3.2 An Illustrative Example

To intuitively demonstrate why global calibration metrics fail to catch localized insensitivity to language, we construct a controlled simulation evaluating a baseline that completely ignores answer text. For each $\mathbf{x}_i$ out of N samples, we obtain M generations $\mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(M)}$ each. We sample “difficulty” $d_i \sim \mathcal{U}[0, 1]$ for every $\mathbf{x}_i$ and simulate answer correctness $\text{corr}_i^{(m)}$ of each $\mathbf{y}_i^{(m)}$ by $\text{corr}_i^{(m)} \sim \text{Bernoulli}(1 - d_i)$, such that higher d_i tend to result in a smaller chance to be correct. We now define confidence estimators, namely an oracle f_{oracle} which returns the actual correctness: $f_{\text{oracle}}(\mathbf{y}_i^{(m)}) = \text{corr}_i^{(m)}$; a constant estimator $f_{\text{constant}}(\mathbf{y}_i) = 1 - d_i$ predicts the difficulty per input, ignoring the generation and its correctness. Similarly, a prior estimator stoically estimates the expected difficulty across the whole dataset s.t. $f_{\text{prior}}(\mathbf{y}_i) = \mathbb{E}_{d_i \sim \mathcal{U}[0, 1]}[1 - d_i] = 0.5$. Lastly, a random estimator $f_{\text{random}}(\mathbf{y}_i) \sim \text{Bernoulli}(1 - d_i)$ independently samples a confidence of 1 or 0 from the same distribution as $\text{corr}_i^{(m)}$.

Results. We run a simulation with $N = 100,000$ and $M = 10$ and show the results in Table 1.³ Unsurprisingly, f_{oracle} achieves perfect results, and the f_{random} scores above-chance AUROC, but with the worst calibration. The f_{constant} scores perfect ECE, a relatively low Brier score of 0.17 and an AUROC of 0.83. f_{prior} shows random-chance AUROC due to only predicting a single value, but reaches perfect ECE and a Brier score of 0.25, demonstrating that ECE is not a proper scoring rule. Although these predictors are deliberately simplified, the implications for LLM CE are staggering: A random baseline that predicts 0 and 1 according to answer difficulty alone can achieve above-random AUROC, but more importantly, f_{constant} , which predicts question difficulty alone and ignores any information from answers, can perform well in both calibration and discrimination. A CE method that behaves randomly or that predicts the same confidence, both independently of the actual answer, is uninformative in a practical setting. There are also two extremes that motivate our novel metrics: Stability (STB) measures whether an estimator predicts the same confidence for semantically equivalent answers, while sensitivity (SST) captures whether an estimator changes its prediction upon encountering a different answer.

4 Methodology

We investigate whether estimated confidences are reliable based on three qualities: *robustness* ensures that the estimated confidence is robust towards irrelevant prompt perturbations; *stability* and *sensitivity* quantify how much the CE methods respond to semantically (dis)similar answers. The computation of all metrics is visualized in Figure 2.

4.1 Robustness Under Prompt Perturbation

Motivation. Prompt formatting is known to influence LLM outputs (Min et al., 2022; Wang et al., 2023; Sclar et al., 2024; Feng et al., 2024), and Xia et al. (2025a) shows that it can affect the performance of CE methods. In practice, prompts are often modified across applications and users. However, while works on verbalized confidence (e.g. Tian et al., 2023; Xiong et al., 2024) have varied prompt formats, the robustness of other confidence estimators has not been systematically examined. We propose that an ideal CE method should provide

³For the metrics introduced in Section 4, we assume that $\mathbf{y}_i^{(m)}$ with the same $\text{corr}_i^{(m)}$ have the same meaning.

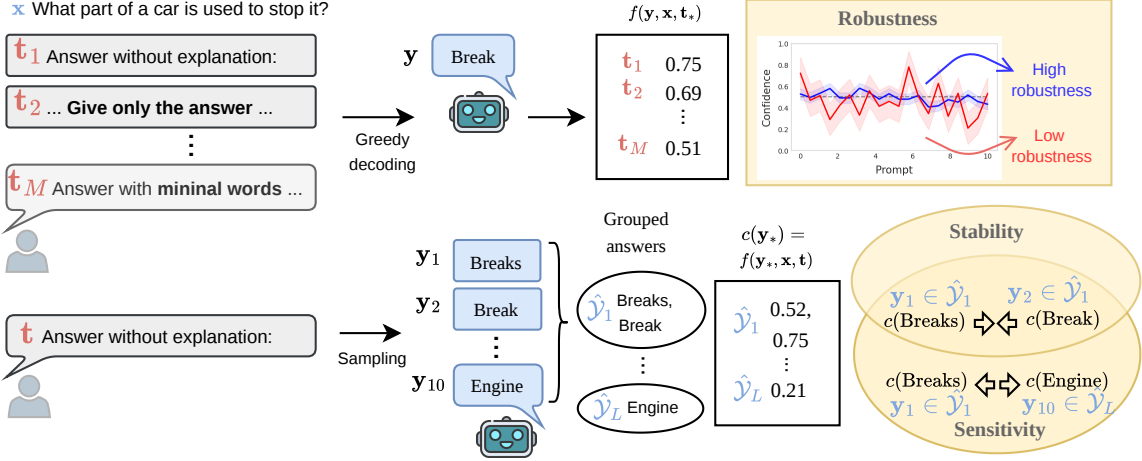


Figure 2: Design of the proposed evaluation metrics: *Robustness* measures the score variation under semantically irrelevant prompt perturbations. *Stability* measures confidence consistency across equivalent answers. *Sensitivity* evaluates whether CE methods assign distinct scores to answers with different meanings than to equivalent ones.

consistent confidences when prompts are altered, but responses stay the same: To isolate prompt variance from generation noise, Robustness is measured by holding the generated response constant while perturbing the instruction format.

Definition 4.1 (Robustness). Let $\mathcal{T} = \{t_1, \dots, t_M\}$ be a set of M semantically equivalent prompts. A confidence estimator f is *robust* to prompt perturbations iff

$$\forall t, t' \in \mathcal{T}, \quad |f(y, x, t) - f(y, x, t')| \leq \varepsilon.$$

Robustness Metric (P-RB $\uparrow$). In practice, we measure robustness given a prompt set $\mathcal{T} = \{t_1, \dots, t_M\}$, assuming a “default” prompt t_1 . We use a semantic equivalence operation $E(\cdot, \cdot)$ (Kuhn et al., 2023)⁴ to identify the set of all answers y_m that are equivalent to the answer y_1 using t_1 . Thus, given $\mathcal{Y} = \{y^{(m)}\}_{m=1}^M$, we define an index set $\hat{\mathcal{I}} = \{m \mid E(y_1, y_m)\}$. Robustness is then computed by measuring the standard deviation $\text{Std}(\cdot)$ of scores across prompts producing equivalent answers and averaging across N samples:

$$\text{P-RB} = 1 - \frac{1}{N} \sum_{i=1}^N \text{Std} \left(\left\{ f(y_i^{(m)}, x_i, t_m) \right\}_{m \in \hat{\mathcal{I}}_i} \right), \quad (3)$$

with $\text{P-RB} > (2 - \varepsilon)/2$ needed for Definition 4.1 to hold (see Appendix A).

⁴While Kuhn et al. (2023) use a bi-directional entailment classifier to determine semantic equivalence, we use LLM-as-a-judge with the prompt in Appendix C.4.

4.2 Stability and Sensitivity Under Answer Variation

Motivation. Setting aside different prompts, LLMs already produce a variety of responses when using stochastic decoding. A confidence estimator should incorporate answer semantics: If the semantics remain unchanged, so should its expected correctness, and thus its confidence. If the answer differs in meaning, the confidence should be adjusted accordingly. As we saw in Section 3.2, a constant predictor can be well-calibrated, but is unreliable by ignoring semantic changes. Conversely, a random predictor is sensitive (since the confidence is random each time), but lacks stability when meaning stays constant.

To define these notions more formally, let us again have a set of M generated responses $\mathcal{Y}$ for a given input x and a (single) prompt t . Then, given some predicted answer $y_m \in \mathcal{Y}$, we can define $\hat{\mathcal{Y}} = \{y' \in \mathcal{Y} \mid E(y_m, y')\}$. Now we can outline two new desiderata in the face of answer variation: To evaluate how an estimator handles generation-side variance alone, we freeze the prompt and analyze stochastically sampled outputs.

Definition 4.2 (Stability). Let $y^{(m)}$ be one out of M responses given a prompt t , and $\hat{\mathcal{Y}} \subseteq \mathcal{Y}$ a set of semantically equivalent responses to y_m . Then, a confidence estimator f is *stable* iff

$$\forall y' \in \hat{\mathcal{Y}}: \quad |f(y^{(m)}, x, t) - f(y', x, t)| \leq \varepsilon.$$

From this, we define the complementary notion

of sensitivity, namely that semantically different answers should receive different scores:

Definition 4.3 (Sensitivity). Let $\mathcal{E}(\mathcal{Y}) = \{\hat{\mathcal{Y}}_l\}_{l=1}^L$ be a finite partition of $\mathcal{Y}$ s.t. $\bigcup_{l=1}^L \hat{\mathcal{Y}}_l = \mathcal{Y}$ and $\forall \hat{\mathcal{Y}}_l, \hat{\mathcal{Y}}_{l'} \in \mathcal{E}, l \neq l' : \hat{\mathcal{Y}}_l \cap \hat{\mathcal{Y}}_{l'} = \emptyset$. Then, given an input $\mathbf{x}$, prompt $\mathbf{t}$ a set of semantically equivalent responses $\hat{\mathcal{Y}}_l \in \mathcal{E}(\mathcal{Y})$, a confidence estimator f is *sensitive* iff

$$\forall \hat{\mathcal{Y}}_l, \hat{\mathcal{Y}}_{l'} \in \mathcal{E}(\mathcal{Y}), l \neq l', \mathbf{y} \in \hat{\mathcal{Y}}_l, \mathbf{y}' \in \hat{\mathcal{Y}}_{l'} : |f(\mathbf{y}, \mathbf{x}, \mathbf{t}) - f(\mathbf{y}', \mathbf{x}, \mathbf{t})| > \varepsilon.$$

Stability Metric (A-STB $\uparrow$). Practically, we do not check all sets $\hat{\mathcal{Y}}_l$, but just restrict ourselves to the most meaningful ones: For each sample, the largest set of answers is denoted as $\hat{\mathcal{Y}}^{\max}$, while the set with the fewest ones is $\hat{\mathcal{Y}}^{\min}$. We then again compute the standard deviation over confidence scores among semantically equivalent answers in $\hat{\mathcal{Y}}^{\max}$ over N data points:

$$\text{A-STB} = 1 - \frac{1}{N} \sum_{i=1}^N \text{Std}(\{f(\mathbf{y}_j, \mathbf{x}_i, \mathbf{t})\}_{\mathbf{y}_j \in \hat{\mathcal{Y}}_i^{\max}}). \quad (4)$$

Applying the same argument as for robustness, we show in [Appendix A](#) that $\text{A-STB} > (2 - \varepsilon)/2$ must hold for [Definition 4.2](#).

Sensitivity Metric (A-SST $\uparrow$). To quantify how effectively a confidence estimator distinguishes semantically different answers relative to semantically equivalent ones, we compute

$$\text{A-SST} = \frac{1}{N} \sum_{i=1}^N |\Delta(\mathcal{Y}_i^{\max}, \hat{\mathcal{Y}}_i^{\min}) - \Delta(\mathcal{Y}_i^{\max}, \hat{\mathcal{Y}}_i^{\max})|, \quad (5)$$

where, using $c(\mathbf{y}) \equiv f(\mathbf{y}, \mathbf{x}, \mathbf{t})$ as a shorthand,

$$\Delta(\hat{\mathcal{Y}}, \hat{\mathcal{Y}}') = \frac{1}{|\hat{\mathcal{Y}}||\hat{\mathcal{Y}}'|} \sum_{\mathbf{y} \in \hat{\mathcal{Y}}} \sum_{\mathbf{y}' \in \hat{\mathcal{Y}}'} |c(\mathbf{y}) - c(\mathbf{y}')|. \quad (6)$$

When the confidence estimator is both stable and sensitive according to [Definitions 4.2](#) and [4.3](#), then this metric is bounded by $\text{A-SST} > \varepsilon/N \sum_{i=1}^N |\hat{\mathcal{Y}}_i^{\max}|^{-1}$, as we derive in [Appendix A](#). A high value indicates that the score differences between semantically different answers are larger than among equivalent answers, and the confidence estimator is therefore sensitive. It should be noted that our metrics do not require any information about the correctness of a response, in contrast

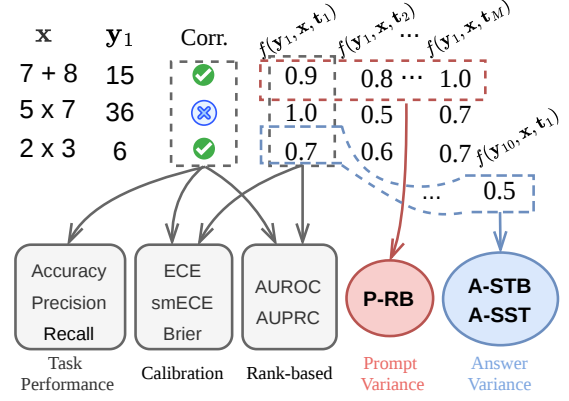


Figure 3: Comparison of the information CE evaluation metrics require. While others need correctness labels, P-RB, A-STB and A-SST solely rely on multiple answers under different prompts or stochastic decoding.

to other common metrics in [Figure 3](#). This is especially attractive in cases where automated correctness judgments are expensive, difficult or noisy, e.g. ambiguous questions ([Min et al., 2020](#)) or time-sensitive questions ([Pletenev et al., 2025](#)).

5 Experiments

In our experiments section, we aim to answer the following three research questions:

RQ1: How do our proposed metrics relate to established metrics for the evaluation of confidence estimators? Do they measure something different?

RQ2: Prior research has shown that calibration in LLMs improves with scale ([Minderer et al., 2021](#); [Dan and Roth, 2021](#); [Kadavath et al., 2022](#); [Zhu et al., 2023](#); [Xiong et al., 2024](#); [Tao et al., 2025](#)). Is this also true for our metrics?

RQ3: Combining all metrics, which confidence estimators perform best across different model families and datasets?

5.1 Setup

Our experiments are conducted on an aggregation of four question answering datasets and 11 LLMs across different model families and sizes.

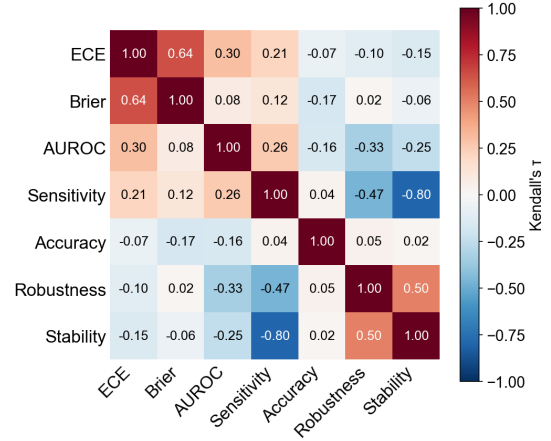
Evaluated Methods. We benchmark 10 CE methods across 4 categories: (1) Output logit-based: Sequence likelihood (*Seq. Likelihood*) aggregates the token probabilities of the generated answer. *Platt Scaling* ([Platt, 1999](#)) learns a scale and shift scalar to calibrate sequence likelihood (in our case on the training set). *Boosted Prob.* by

Dinh and Niehues (2025) identifies dominant token probabilities in the output sequence. $P(\text{True})$ prompts the LLMs to judge if the response is true, and uses the normalized probability of the “True” token (Kadavath et al., 2022). (2) Linguistic features: *Verbalized Conf.* (Tian et al., 2023; Xiong et al., 2024) prompts the LLM to state how certain it is about its answer. (3) Auxiliary models: Calib-1-Focal (*Calib1*) trains an auxiliary classifier (Ulmer et al., 2024; Xia et al., 2025a), uses response correctness as training targets. (4) Internal states-based: *Attention Score* and *Hidden Score* (Sriramanan et al., 2024) are derived from the mean log-determinant of attention heads and hidden representations, respectively; $P(IK)$ (Kadavath et al., 2022) and *SAPLMA* (Azaria and Mitchell, 2023) train probe classifiers on last-layer hidden states from different token positions. Implementation details are in Appendix C.6. All the confidence scores estimated by different methods are scaled to 0-1 for evaluation.

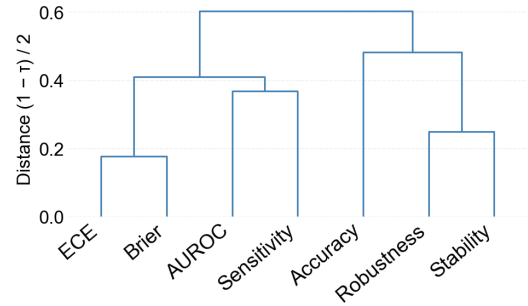
Datasets. We test on four open-ended question-answering benchmarks covering diverse topics. They include 2,000 randomly sampled training examples each from Natural Questions (NQ; Kwiatkowski et al., 2019), SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). All methods are evaluated on 1,000 randomly sampled test questions each.

LLMs. We evaluate 11 LLMs spanning five major families, including Mistral (Mistral AI, 2024, 2025), Llama 3 (Meta AI, 2024), Qwen 2.5 (Yang et al., 2024b), OLMo 2 (Team OLMo et al., 2024) and GPT-4o (OpenAI, 2024) models, whose known sizes range from 7B to 123B and include both instruction and reasoning models. More details are given in Appendix C.5.

Experimental Design. For measuring prompt robustness, we use greedy decoding to ensure that any variability is solely driven by prompt perturbations, for which we test the 10 variants shown in Table 22 (Appendix C.3). To avoid introducing ambiguity, we keep the question fixed and perturb only the instructions. Variants include forms such as “answer the question” and “provide an answer to the question.” Each prompt also enforces conciseness, with phrases like “give ONLY the answer,” “no other words or explanation,” etc. For stability and sensitivity, we set the model temperature to 0.7 to generate 10 sampled answers.



(a) Kendall’s τ between metrics. Statistic test results are shown in Appendix B.5.



(b) Dendrogram of hierarchical clusters.

Figure 4: Similarity of CE evaluation metrics. (a) Heatmap of Kendall’s τ between metrics, including accuracy. (b) Hierarchical clustering based on τ values.

We use GPT-4o as a judge model (see the judge prompt in Table 23 and Table 24 in appendix) to determine the semantic equivalence between LLM responses and target answers for correctness evaluation, as well as the semantic equivalence among sampled responses when computing stability and sensitivity metrics.

5.2 Comparing Evaluation Metrics

We first compare the similarity of CE evaluation metrics by pooling results across models and confidence estimators, computing Kendall’s τ (Kendall, 1938) correlation coefficient between them, and then averaging the result across our four datasets. We then run hierarchical clustering over the results, treating correlation (as $(1 - \tau)/2$) as a distance measure to identify closely related metrics.

Results. Figure 4a shows only mild correlation between our and established metrics: We find a $\tau = 0.26$ for sensitivity and AUROC, which is much lower than e.g. $\tau = 0.64$ between Brier score and ECE. In contrast, our metrics’ correla-

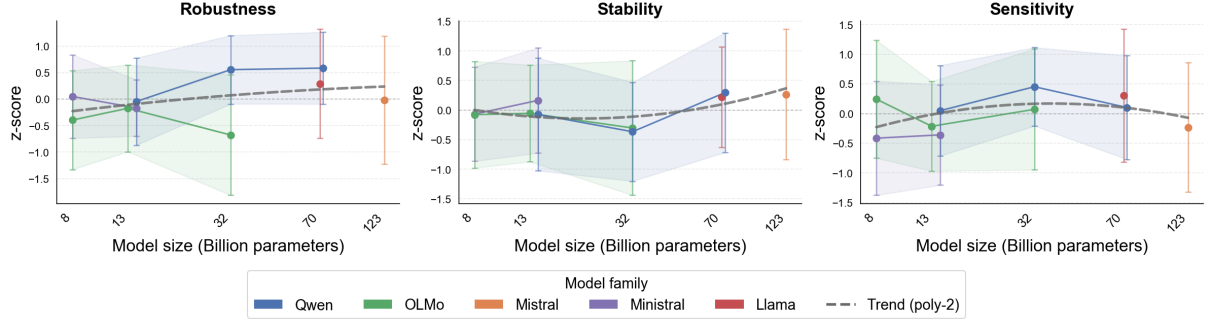


Figure 5: Normalized metric value distributions by model family and size. Whiskers and shaded regions indicate the standard deviation over observed metric values across datasets and CE methods. Models of the same family are connected and appear in the same color. Trend lines are second-degree polynomials fitted on all observations.

		SciQ						PopQA					
Method		ECE↓	Brier↓	AUROC↑	P-RB↑	A-STB↑	A-SST↑	ECE↓	Brier↓	AUROC↑	P-RB↑	A-STB↑	A-SST↑
Logit-based	Seq. Likelihood	0.08	0.15	0.68	0.94	0.92	0.17	0.39	0.31	0.84	0.95	0.94	0.20
	Boosted Prob.	0.18	0.18	0.73	0.99	1.00	0.01	0.58	0.51	0.86	0.98	0.97	0.07
	Platt Scaling	0.10	0.16	0.68	0.99	0.98	0.04	0.28	0.25	0.84	0.99	0.99	0.04
	P(True)	0.15	0.17	0.73	0.95	0.97	0.18	0.33	0.32	0.80	0.93	0.94	0.25
Internal States	Attention Score	0.02	0.15	0.55	0.99	0.99	0.02	0.05	0.20	0.61	0.99	0.98	0.05
	Hidden Score	0.04	0.15	0.59	0.98	0.98	0.04	0.05	0.20	0.57	0.99	0.99	0.03
	SAPLMA	0.15	0.17	0.75	0.89	0.92	0.21	0.14	0.15	0.88	0.92	0.95	0.10
	P(IK)	0.06	0.15	0.68	0.98	1.00	0.00	0.08	0.16	0.82	0.97	1.00	0.00
Verb. Ling.	Verbalized Conf.	0.15	0.17	0.64	0.99	1.00	0.03	0.61	0.59	0.66	0.98	0.98	0.07
Aux. Ling.	Calib1	0.10	0.16	0.60	0.99	0.97	0.07	0.09	0.17	0.78	0.99	0.97	0.07

Table 2: Calibration and our metrics averaged across all models. Arrows indicate metric direction (↓ lower is better, ↑ higher is better); the best value per column is bold. Results of other datasets are shown in Table 3 in Appendix B.1.

tion coefficients with each other range from -0.80 to 0.50 . Interestingly, no metric has a meaningful correlation with accuracy. Using the hierarchical clustering in Figure 4b, we can further observe that robustness and stability are clustered together, albeit their similarity is still trumped by ECE and Brier score; sensitivity remains grouped with AUROC. In Figure 8 in Appendix B.4, we plot metric distributions, showing strongly skewed results for robustness and stability on all datasets, contrasted by an overall lack of sensitivity, all the while distributions appear very distinct from ECE, AUROC, and Brier score. Addressing RQ1, we take this as evidence that our metrics measure novel and so-far overlooked properties of confidence estimators.

5.3 Scaling Laws for Confidence Estimates

Prior literature has repeatedly shown that larger models tend to be better calibrated (Ahuja et al., 2022; Zhu et al., 2023; Xiong et al., 2024; Tao et al., 2025). To test whether the same is true for our metrics, we plot the distribution of met-

ric scores per model across datasets and CE methods, using parameter size in billions as a reference point.⁵ Since metric score distributions differ between confidence estimators and datasets, we apply z -score normalization across measurements for the same CE method and dataset from the same model family. This isolates the general improvement of metric scores, even if dataset difficulty and the performance of specific CE methods differ.

Results. We show trends for our metrics in Figure 5, demonstrating a mild improvement with model size.⁶ This trend by estimator is often inconsistent depending on the type of method and model, as we show in more detailed graphs in Figure 10 in Appendix B.7. For RQ2, scaling may improve confidence quality with respect to language vari-

⁵We remove Calib1 since it uses an auxiliary model.

⁶Figure 9 (Appendix B.7) shows the plots for ECE, Brier score, and AUROC, where we mostly corroborate existing findings. The weaker improvements with scale do not contradict prior work, since often only specific estimators such as token/sequence likelihood (Zhu et al., 2023; Tao et al., 2025) or verb. confidence (Kadavath et al., 2022) are considered.

ation; however, the substantial differences across estimators and models highlight the need for careful evaluation in practical settings.

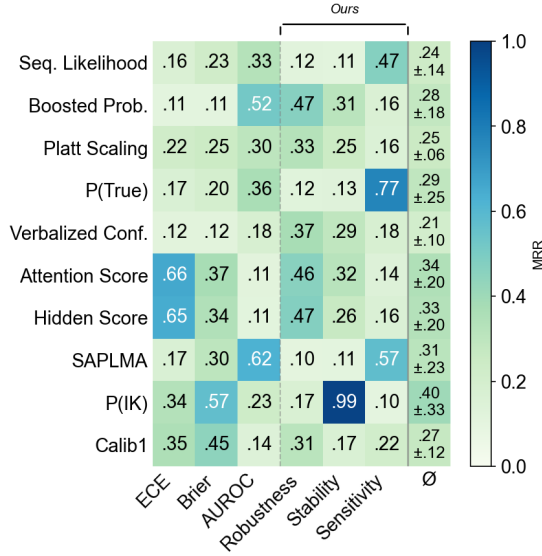


Figure 6: Mean reciprocal rank of CE methods per evaluation metric, averaged over datasets. Ø is another average across all metrics, including standard deviation.

5.4 Comparing Confidence Estimators

To test whether any confidence estimators reign supreme across all models and datasets, we first analyze the evaluation results per metric and dataset, averaged across models. For each CE method, we then compute the mean reciprocal rank (MRR), where 1 means that it outperformed all others.

Results. Table 2 shows the result for SciQ and PopQA, with the remaining dataset results shown in Table 3 in Appendix B.1. Overall, no method outperforms any other on all metrics. Well-calibrated methods also seem to struggle with sensitivity and vice versa. At the same time, while all confidence estimators score high on robustness (P-RB) and stability (A-STB), most fail on sensitivity (A-SST). By overall MRR, as shown in Figure 6 across all datasets and in Figure 11 in Appendix B.9 per dataset, $P(IK)$ achieves the best overall MRR, but with low ranking on robustness and sensitivity. Regarding **RQ3**, this highlights that the choice of the best CE method will be highly dependent on the specific dataset, model, and other constraints, such as different priorities of certain metrics.

5.5 Why Is Sensitivity So Low?

In the previous section, we found that almost all CE methods scored high on robustness and stabil-

ity, but very low on sensitivity. This means that in practice, confidence estimates fail to adapt to semantically *different* responses. This is worrying, since a different response likely also implies a change in output quality or correctness, and therefore a miscalibrated confidence. To investigate this phenomenon, we display specific responses in detail (see Tables 11 to 13 in Appendix B.10). We hypothesize that CE methods leveraging information closer to the response generation exhibit higher sensitivity to answer changes. Estimators trained solely on the question’s final token (like $P(IK)$) are inherently blind to the generated text, resulting in low sensitivity; in contrast, methods evaluating the response’s final token (like SAPLMA) achieve much higher sensitivity. Similarly, methods that directly rely on output logits without post-hoc scaling, such as sequence likelihood and $P(True)$, tend to be more sensitive to changes in the generated answer, as shown by the rankings in Figure 6. This aligns with results by Eusebi et al. (2026), showing that correctness information is incrementally encoded during generations.

6 Discussion

Since our results in Section 5.4 found no universally superior confidence estimator, the choice of method depends on the target application, as different settings require trade-offs between robustness, stability, and sensitivity. For instance, user-facing QA systems prioritize robustness to maintain consistent abstention behavior under evolving prompts. In contrast, multi-candidate reranking settings, such as LLM-as-a-judge (Wagner et al., 2024; Jung et al., 2025) or reward models (Yang et al., 2024a), rely more on stability and sensitivity to assess competing responses. Conditional computation and model cascades (Gupta et al., 2024) emphasize sensitivity to detect low-confidence predictions for escalation to stronger models. Multi-agent collaboration (Chen et al., 2024a; Yang et al., 2024c) requires balancing all three properties, while in controlled environments with fixed prompts and limited answer variation, traditional calibration metrics are sufficient. Importantly, however, the observed lack of sensitivity across all confidence estimators implies that methods should better exploit relevant information across generation time steps (Eusebi et al., 2026) and model components (Shelmanov et al., 2025).

7 Conclusion

We introduced three novel complementary metrics (robustness, stability and sensitivity) to evaluate whether confidence estimates remain reliable under language variation. Our results show that these metrics are largely independent of traditional CE metrics for calibration, suggesting that existing evaluations capture only a limited view of confidence quality. While most CE methods achieve high calibration, robustness and stability, they remain insensitive to changes in answer meaning. This raises concerns for downstream applications such as multi-candidate reranking. We further show that scaling alone does not resolve this issue, and our findings suggest that incorporating more answer-side information may improve sensitivity.

As no CE method consistently dominates, we highlight fundamental trade-offs between these metrics, suggesting that the “best” confidence estimator depends strongly on the target application.

Limitations

While our study provides a comprehensive evaluation of LLM confidence estimation across robustness, stability, sensitivity, calibration, and discrimination, several limitations remain.

First, our analysis is limited to English and fact-based question answering tasks; extending to multilingual or open-ended domains (e.g., reasoning, dialogue, multiple-choice, or creative generation) may reveal different findings. Second, although we designed a diverse set of prompting strategies, they do not exhaustively capture all linguistic and contextual variations that may influence confidence expression. In particular, it remains challenging to construct more aggressive prompt perturbations that preserve identical semantic meaning and avoid ambiguity across all questions and datasets. Third, part of our evaluation relies on GPT-4o as a judge model to assess answer correctness and semantic equivalence. While this enables scalable annotation, it may introduce model-specific biases (Bavaresco et al., 2025). Fourth, we do not include consistency-based confidence estimation methods, since our metrics themselves rely on multiple samples to be computed, and are therefore incompatible with self-consistency. Consistency-based methods are not naturally superior on our metrics; for instance, they could easily assign the same confidence score to two semantically different answers appearing at the same frequency in the sampling

pool. In the end, our contribution is not comprehensively benchmark all existing CE methods, but instead to highlight overlooked aspects regarding language variation using our proposed metrics.

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF)[10.47379/VRG19008] “Knowledge infused Deep Learning for Natural Language Processing”, and is supported by the Dutch National Science Foundation (NWO Vici VI.C.212.053). We acknowledge the support of the OpenAI Researcher Access Program for providing API credits used in this research.

References

- Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323.
- Michael Alfertshofer, Cosima C Hoch, Paul F Funk, Katharina Hollmann, Barbara Wollenberg, Samuel Knoedler, and Leonard Knoedler. 2024. [Sailing the seven seas: a multinational comparison of chatgpt’s performance on medical licensing examinations](#). *Annals of Biomedical Engineering*, 52(6):1542–1545.
- Tariq Alqahtani, Hisham A Badreldin, Mohammed Al-rashed, Abdulrahman I Alshaya, Sahar S Alghamdi, Khalid Bin Saleh, Shuroug A Alowais, Omar A Alshaya, Ishrat Rahman, Majed S Al Yami, et al. 2023. [The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research](#). *Research in social and administrative pharmacy*, 19(8):1236–1242.
- 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.
- Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. [Uncertainty in natural language generation: From theory to applications](#). *arXiv preprint arXiv:2307.15703*.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2025. LLMs instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

- Linguistics (Volume 2: Short Papers)*, pages 238–255.
- Jarosław Błasiok and Preetum Nakkiran. 2024. [Smooth ECE: Principled reliability diagrams via kernel smoothing](#). In *The Twelfth International Conference on Learning Representations*.
- Andrew P. Bradley. 1997. [The use of the area under the roc curve in the evaluation of machine learning algorithms](#). *Pattern Recognition*, 30(7):1145–1159.
- Glenn W. Brier. 1950. [Verification of forecasts expressed in terms of probability](#). *Monthly Weather Review*, 78(1):1 – 3.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *The Eleventh International Conference on Learning Representations*.
- Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2020. [Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts](#). *Advances in neural information processing systems*, 33:1356–1367.
- Jiuhai Chen and Jonas Mueller. 2024. [Quantifying uncertainty in answers from any language model and enhancing their trustworthiness](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.
- Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024a. [Reconcile: Round-table conference improves reasoning via consensus among diverse llms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.
- Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024b. [Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.
- Soham Dan and Dan Roth. 2021. [On the effects of transformer size on in-and out-of-domain calibration](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101.
- Tu Anh Dinh and Jan Niehues. 2025. [Are generative models underconfident? better quality estimation with boosted model probability](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3382, Suzhou, China. Association for Computational Linguistics.
- Aliai Eusebi, Alexander Herzog, Xiaoyu Liang, Marie Vasek, Enrico Mariconti, and Lorenzo Cavallaro. 2026. [Reading calibrated uncertainty from language model trajectories](#). *Preprint*, arXiv:2605.22864.
- Wade Fagen-Ulmschneider. 2023. [Perception of probability words](#). Manuscript, University of Illinois Urbana-Champaign.
- Zijian Feng, Hanzhang Zhou, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. [Unveiling and manipulating prompt influence in large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.
- Tobias Groot and Matias Valdenegro Toro. 2024. [Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models](#). In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.
- 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, ICML’17*, page 1321–1330. JMLR.org.
- Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkritum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. [Language model cascades: Token-level uncertainty and beyond](#). *arXiv preprint arXiv:2404.10136*.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. [Look before you leap: An exploratory study of uncertainty analysis for large language models](#). *IEEE Transactions on Software Engineering*, 51(2):413–429.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.
- Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. [Trust or escalate: Llm judges with provable guarantees for human agreement](#). In *International Conference on Learning Representations*, volume 2025, pages 3101–3125.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna

- Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.
- John Kirchenbauer, Jacob Oaks, and Eric Heim. 2022. [What Is Your Metric Telling You? Evaluating Classifier Calibration under Context-specific Definitions of Reliability](#). *ArXiv preprint*, abs/2205.11454.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.
- Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. [Verified Uncertainty Calibration](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3787–3798.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Florian Le Bronnec, Alexandre Verine, Benjamin Nègrevergne, Yann Chevalere, and Alexandre Alauzen. 2024. [Exploring precision and recall to assess the quality and diversity of LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11418–11441, Bangkok, Thailand. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *Transactions on Machine Learning Research*.
- Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. 2023. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42):1–63.
- Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. [Litcab: Lightweight language model calibration over short- and long-form responses](#). In *Proceedings of the Twelfth International Conference on Learning Representations*. ICLR.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). *Preprint*, arXiv:2303.08896.
- Meta AI. 2024. [Introducing meta llama 3: The most capable openly available llm to date](#). Accessed: 2025-05-01.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694.
- Mistral AI. 2024. [Mistral large instruct 2411](#). Accessed: 2025-05-01.
- Mistral AI. 2025. [Minstral-3-8b-reasoning-2512](#). <https://huggingface.co/mistralai/Minstral-3-8B-Reasoning-2512>. Hugging Face model card, accessed 2026-05-22.
- Lwin Moe, Uyen Trang Nguyen, and Boi Trinh Luu. 2025. [Mitigating class imbalance in fact-checking datasets through llm-based synthetic data generation](#). In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation, MAD’ 25*, page 73–80, New York, NY, USA. Association for Computing Machinery.
- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299.

- Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2025. Are large language models more honest in their probabilistic or verbalized confidence? In *Information Retrieval*, pages 124–135, Singapore. Springer Nature Singapore.
- Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–41.
- OpenAI. 2024. [Hello gpt-4o](#). Accessed: 2025-05-01.
- Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. 2019. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.
- Sergey Pletenev, Maria Marina, Nikolay Ivanov, Daria Galimzianova, Nikita Krayko, Mikhail Salnikov, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. [Will it still be true tomorrow? multilingual evergreen question classification to improve trustworthy QA](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8603–8620, Suzhou, China. Association for Computational Linguistics.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Dähm, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, et al. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35700–35719.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. [A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions](#). *ACM Comput. Surv.*, 58(3).
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. [Prompting gpt-3 to be reliable](#). Preprint, arXiv:2210.09150.
- Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. [Llm-check: Investigating detection of hallucinations in large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.
- Elias Stengel-Eskin and Benjamin Van Durme. 2023. [Calibrated interpretation: Confidence estimation in semantic parsing](#). Preprint, arXiv:2211.07443.
- Linwei Tao, Yi-Fan Yeh, Mingjing Dong, Tao Huang, Philip Torr, and Chang Xu. 2025. Revisiting uncertainty estimation and calibration of large language models. *arXiv preprint arXiv:2505.23854*.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. [2 olmo 2 furious](#). *arXiv preprint arXiv:2501.00656*.
- Alberto Testoni and Iacer Calixto. 2026. [Mind the gap: Benchmarking LLM uncertainty and calibration with specialty-aware clinical QA and reasoning-based behavioural features](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2364–2382, Rabat, Morocco. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. [Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdo Yun, and Seong Oh. 2024. [Calibrating large language models using their generations only](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.
- Dennis Ulmer, Alexandra Lorson, Ivan Titov, and Christian Hardmeier. 2025. Anthropomorphic uncertainty: What verbalized uncertainty in language models is missing. *arXiv preprint arXiv:2507.10587*.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. [Benchmarking uncertainty quantification methods for large language models with lm-polygraph](#). *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.
- vLLM Contributors. 2023. vllm: A high-throughput and memory-efficient llm serving library. <https://github.com/vllm-project/vllm>. Accessed: 2025-01.
- Nico Wagner, Michael Desmond, Rahul Nair, Zahra Ashktorab, Elizabeth M Daly, Qian Pan, Martín Santillán Cooper, James M Johnson, and Werner Geyer. 2024. Black-box uncertainty quantification method for llm-as-a-judge. *arXiv preprint arXiv:2410.11594*.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Yuxi Xia, Pedro Henrique Luz De Araujo, Klim Zaporozhets, and Benjamin Roth. 2025a. [Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3740–3761, Vienna, Austria. Association for Computational Linguistics.
- Yuxi Xia, Klim Zaporozhets, and Benjamin Roth. 2025b. [Learn to pick the winner: Black-box ensembling for textual and visual question answering](#). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 12–26, Hannover, Germany. HsH Applied Academics.
- Dongfu Xiao, Chen Gao, Zhengquan Luo, Chi Liu, and Sheng Shen. 2025. [Can llms assist computer education? an empirical case study of deepseek](#). *Preprint*, arXiv:2504.00421.
- 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 *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Adam X Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. 2024a. Bayesian reward models for llm alignment. *arXiv preprint arXiv:2402.13210*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024b. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. 2024c. Confidence calibration and rationalization for llms via multi-agent deliberation. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. [Calibrating the confidence of large language models by eliciting fidelity](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.
- Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. [Relying on the unreliable: The impact of language models’ reluctance to express uncertainty](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.
- Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795.

Mingcheng Zhu, Yu Liu, and Tingting Zhu. 2026. Towards generation-efficient uncertainty estimation in large language models. *arXiv preprint arXiv:2605.06053*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [ProSA: Assessing and understanding the prompt sensitivity of LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Theoretical Results

In this section, we directly relate the definitions we lay out in [Section 4](#) to our proposed metrics. To this end, we formally prove lower bounds that our metrics have to exceed in order for the definitions to hold. We start with [Definition 4.1](#) and the prompt robustness metric in [Equation \(3\)](#).

Theorem A.1. *Given a set of M semantically equivalent prompts $\mathcal{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_M\}$ and a robust confidence estimator f according to [Definition 4.1](#) s.t.*

$$\forall \mathbf{t}, \mathbf{t}' \in \mathcal{T}, \quad |f(\mathbf{y}, \mathbf{x}, \mathbf{t}) - f(\mathbf{y}, \mathbf{x}, \mathbf{t}')| \leq \varepsilon, \quad (7)$$

then it must hold for the robustness metric $\hat{\sigma}_r$ defined in [Equation \(3\)](#) that

$$P\text{-}RB > \frac{2 - \varepsilon}{2}. \quad (8)$$

Proof. In the following, we will use the shorthand c for $f(\mathbf{y}, \mathbf{x}, \mathbf{t})$. Because of [Definition 4.1](#), we know that any two c, c' cannot lie more than ε apart. Since standard deviation and variance are metrics of dispersion, they are invariant to translation, meaning that they will stay constant if the data is shifted by a constant amount. Therefore, let us assume that $c \in [0, \varepsilon]$. Since we would like to find the maximum dispersion possible under [Equation \(7\)](#), let us further assume that k points are equal to 0, and $M' - k$ equal ε , where $M' \leq M$ refers to number of generations that are equivalent to the default generation $\mathbf{y}_1$, including $\mathbf{y}_1$. Therefore, their mean becomes $\mu = \frac{M' - k}{M'} \varepsilon$. Correspondingly, we can evaluate their variance as

$$\sigma^2 = \frac{1}{M'} [k(0 - \mu)^2 + (M' - k)(\varepsilon - \mu)^2] \quad (9)$$

$$= \frac{1}{M'} \left[k \frac{(M' - k)^2 \varepsilon^2}{M'^2} + (M' - k) \frac{k^2 \varepsilon^2}{M'^2} \right] \quad (10)$$

$$= \frac{\varepsilon^2}{M'^3} [k(M' - k)^2 + (M' - k)k^2] \quad (11)$$

$$= \frac{\varepsilon^2}{M'^3} k(M' - k)M'. \quad (12)$$

Next, we would like to identify the k that maximizes $k(M' - k)M'$, we can easily be identified as $k = \frac{M'}{2}$. Reinserting it into the previous expression yields

$$= \frac{\varepsilon^2}{M'^3} \frac{M'}{2} \left(M' - \frac{M'}{2} \right) M' \quad (13)$$

$$= \frac{\varepsilon^2}{M'^3} \frac{M'^3}{4} = \frac{\varepsilon^2}{4} \leftrightarrow \sigma = \frac{\varepsilon}{2}. \quad (14)$$

By plugging this result into the definition of robustness in [Definition 4.1](#), we can show that since the biggest standard deviation for which [Equation \(7\)](#) still holds is $\frac{\varepsilon}{2}$ leads robustness to amount to

$$\text{PR-B} > 1 - \frac{\varepsilon}{2} = \frac{2 - \varepsilon}{2}. \quad (15)$$

□

Since [Definition 4.2](#) and the stability metric in [Equation \(4\)](#) follow the same structure, the matching theorem follows naturally.

Theorem A.2. *Given a set of M responses given a prompt $\mathbf{t}$, $\hat{\mathcal{Y}} \subseteq \mathcal{Y}$ a set of semantically equivalent responses to $\mathbf{y}_m$ and a stable* confidence estimator f according to [Definition 4.2](#) s.t.*

$$\forall \mathbf{y}' \in \hat{\mathcal{Y}} : |f(\mathbf{y}^{(m)}, \mathbf{x}, \mathbf{t}) - f(\mathbf{y}', \mathbf{x}, \mathbf{t})| \leq \varepsilon.$$

Then it must hold for the stability metric $\hat{\sigma}_s$ defined in [Equation \(3\)](#) that

$$A\text{-STB} > \frac{2 - \varepsilon}{2}. \quad (16)$$

Proof. The proof follows the same argument as [Theorem A.1](#) without loss of generality. □

Lastly, we would like to prove a similar bound for the sensitivity metric in [Equation \(5\)](#), which we do below.

Theorem A.3. *Let $\mathcal{Y}$ denote a set of M generated responses for a given input $\mathbf{x}$ and a (single) prompt $\mathbf{t}$ s.t. $\mathcal{Y} = \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}\}$. Also let $\mathcal{E}(\mathcal{Y}) = \{\hat{\mathcal{Y}}_l\}_{l=1}^L$ be a finite partition of $\mathcal{Y}$ s.t. $\bigcup_{l=1}^L \hat{\mathcal{Y}}_l = \mathcal{Y}$ and $\forall \hat{\mathcal{Y}}_l, \hat{\mathcal{Y}}_{l'} \in \mathcal{E}, l \neq l' : \hat{\mathcal{Y}}_l \cap \hat{\mathcal{Y}}_{l'} = \emptyset$ and let f be a sensitive confidence estimator according to [Definition 4.3](#) s.t.*

$$\begin{aligned} \forall \hat{\mathcal{Y}}_l, \hat{\mathcal{Y}}_{l'} \in \mathcal{E}(\mathcal{Y}), l \neq l', \mathbf{y} \in \hat{\mathcal{Y}}_l, \mathbf{y}' \in \hat{\mathcal{Y}}_{l'} : \\ |f(\mathbf{y}, \mathbf{x}, \mathbf{t}) - f(\mathbf{y}', \mathbf{x}, \mathbf{t})| > \varepsilon, \end{aligned} \quad (17)$$

and that f is stable according to [Definition 4.2](#) s.t.

$$\forall \mathbf{y}' \in \hat{\mathcal{Y}} : |f(\mathbf{y}^{(m)}, \mathbf{x}, \mathbf{t}) - f(\mathbf{y}', \mathbf{x}, \mathbf{t})| \leq \varepsilon.$$

Then it must hold for the sensitivity metric $A\text{-SST}$ defined in [Equation \(5\)](#) that

$$A\text{-SST} > \frac{\varepsilon}{N} \sum_{i=1}^N |\hat{\mathcal{Y}}_i^{\max}|^{-1}. \quad (18)$$

Proof. We start by restating the definition of sensitivity from [Equation \(5\)](#):

$$A\text{-SST} = \frac{1}{N} \sum_{i=1}^N |\Delta(\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\min}) - \Delta(\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\max})| \quad (19)$$

where $\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\min} \in \mathcal{E}(\mathcal{Y})$ are the largest and smallest sets within the partition respectively, with i corresponding to a specific input $\mathbf{x}_i$ in the dataset. We treat each of the terms in the difference separately next. Firstly, we have

$$\Delta(\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\min}) \quad (20)$$

$$= \frac{1}{|\hat{\mathcal{Y}}_i^{\max}| |\hat{\mathcal{Y}}_i^{\min}|} \sum_{\mathbf{y} \in \hat{\mathcal{Y}}_i^{\max}} \sum_{\mathbf{y}' \in \hat{\mathcal{Y}}_i^{\min}} |c(\mathbf{y}) - c(\mathbf{y}')| \quad (21)$$

$$> \frac{1}{|\hat{\mathcal{Y}}_i^{\max}| |\hat{\mathcal{Y}}_i^{\min}|} \sum_{\mathbf{y} \in \hat{\mathcal{Y}}_i^{\max}} \sum_{\mathbf{y}' \in \hat{\mathcal{Y}}_i^{\min}} \varepsilon = \varepsilon, \quad (22)$$

where the last step follows from [Definition 4.3](#). We can also simplify $\Delta(\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\max})$ by noticing that from the $|\hat{\mathcal{Y}}_i^{\max}|^2$ differences considered, $|\hat{\mathcal{Y}}_i^{\max}|$ of them will amount to 0 when an element is compared to itself. Using the definition of stability in [Definition 4.2](#), we also know that the remaining $|\hat{\mathcal{Y}}_i^{\max}|^2 - |\hat{\mathcal{Y}}_i^{\max}|$ comparisons amount to at most ε each. Therefore,

$$\Delta(\hat{\mathcal{Y}}_i^{\max}, \hat{\mathcal{Y}}_i^{\max}) \leq \frac{1}{|\hat{\mathcal{Y}}_i^{\max}|^2} (|\hat{\mathcal{Y}}_i^{\max}|^2 - |\hat{\mathcal{Y}}_i^{\max}|) \varepsilon \quad (23)$$

$$= \varepsilon - \frac{1}{|\hat{\mathcal{Y}}_i^{\max}|} \varepsilon = \frac{|\hat{\mathcal{Y}}_i^{\max}| - 1}{|\hat{\mathcal{Y}}_i^{\max}|} \varepsilon. \quad (24)$$

Putting the two parts together, we obtain

$$A\text{-SST} > \frac{1}{N} \sum_{i=1}^N \left| \varepsilon - \frac{|\hat{\mathcal{Y}}_i^{\max}| - 1}{|\hat{\mathcal{Y}}_i^{\max}|} \varepsilon \right| \quad (25)$$

$$= \frac{1}{N} \sum_{i=1}^N \left| \frac{1}{|\hat{\mathcal{Y}}_i^{\max}|} \varepsilon \right| \quad (26)$$

$$= \frac{\varepsilon}{N} \sum_{i=1}^N |\hat{\mathcal{Y}}_i^{\max}|^{-1}. \quad (27)$$

□

B Additional results

B.1 Additional Dataset Results of Metric Values

Similar to Table 2, Table 3 presents the results on the NQ and TriviaQA datasets averaged across all models. Overall, we observe trends highly consistent with those of our proposed metrics on SciQ and PopQA.

B.2 Judge Reliability

To validate judge reliability, we conducted a small-scale study involving two human annotators (Human 1 with an AI background and Human 2 without an AI background) and GPT-5.5 as an independent judge. We evaluated 200 sampled correctness and semantic-partition judgments, with 50 examples from each dataset. All judges are provided with the same prompts as in the paper.

The results are shown in Table 4. For correctness judgments, GPT-4o shows strong agreement with both human annotators and GPT-5.5. Agreement across all four judges is also high, with a Krippendorff’s alpha of 0.814. GPT-4o’s agreement with the human-consensus labels is nearly identical to that of GPT-5.5, suggesting that the correctness labels are not artifacts of the original GPT-4o judge.

For semantic partitioning, we conducted an acceptability audit to determine whether the independent validators approved the partitions generated by GPT-4o. The results are shown in Table 5. Approval rates are high for both human annotators and GPT-5.5.

B.3 Accuracy of Models

We present the detailed model accuracy on each benchmark dataset in Figure 7. Overall, all models achieve the strongest performance on SciQ and the weakest on PopQA, with the performance gap between the two datasets reaching nearly 50%. This highlights the substantial diversity and varying difficulty of the evaluated benchmarks.

B.4 Distribution of Metric Values

Figure 8 shows the distribution of metric values across models and CE methods for each dataset. The results demonstrate that CE performance varies substantially across datasets, with no single dataset consistently yielding higher metric values across all metrics. Moreover, each metric exhibits distinct distributional characteristics: robustness and stability are consistently skewed toward high values,

whereas sensitivity remains uniformly low across datasets. At the same time, the distributions of our proposed metrics differ markedly from those of ECE, AUROC, and Brier score, further suggesting that they capture complementary aspects of confidence quality.

B.5 Statistic Test

We report repeated-measures-aware Kendall’s τ correlations between the evaluation metrics in Table 6. Each cell presents the mean dataset-wise tau with a cluster-bootstrap 95% confidence interval. Stars indicate Benjamini-Hochberg-corrected permutation p-values: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

The corrected analysis supports our original conclusion: ECE and Brier are strongly related, whereas P-RB, A-STB, and A-SST capture complementary dimensions.

B.6 Retained Fraction of Prompts for Prompt Robustness

We report the average retained fraction $|I|/M$ in Table 7, where I is the subset of prompts producing answers semantically equivalent to the answer generated from the default prompt. The higher-accuracy datasets, SciQ and TriviaQA, have the highest retained fractions. Nevertheless, all datasets also retain substantial proportions of prompts.

B.7 Detailed Scaling Results

We present more detailed scaling results in Figure 10. For linguistic CE methods, such as verbalized confidence, scaling generally improves robustness and stability, while sensitivity exhibits inconsistent trends. In contrast, the trends for the other CE method families are substantially less consistent across metrics and model scales. Overall, the results suggest that the effect of scaling strongly depends on both the CE method type and the underlying model family.

B.8 Relationship between A-SST and answer correctness.

We additionally examine whether A-SST reflects differences in answer correctness. For each example, we compare representative answers from the largest and smallest semantic groups, label the correctness of each answer, and separate cases based on whether the two answers have the same or different correctness labels.

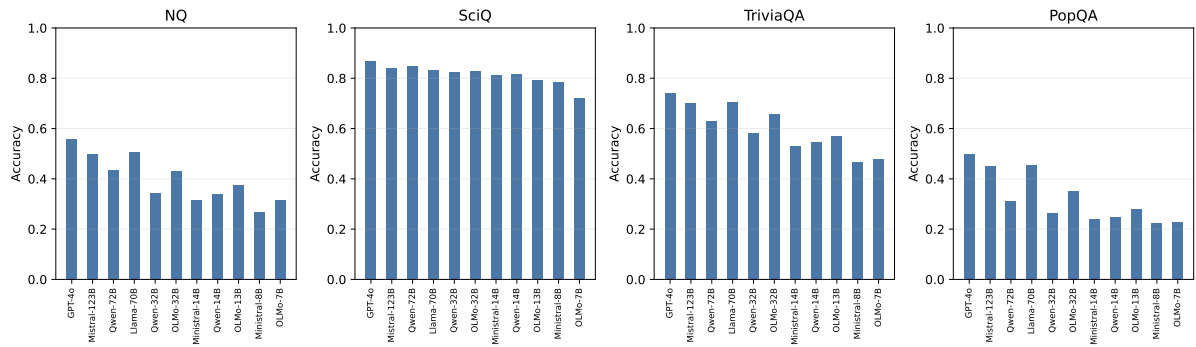


Figure 7: Accuracy of models on different datasets.

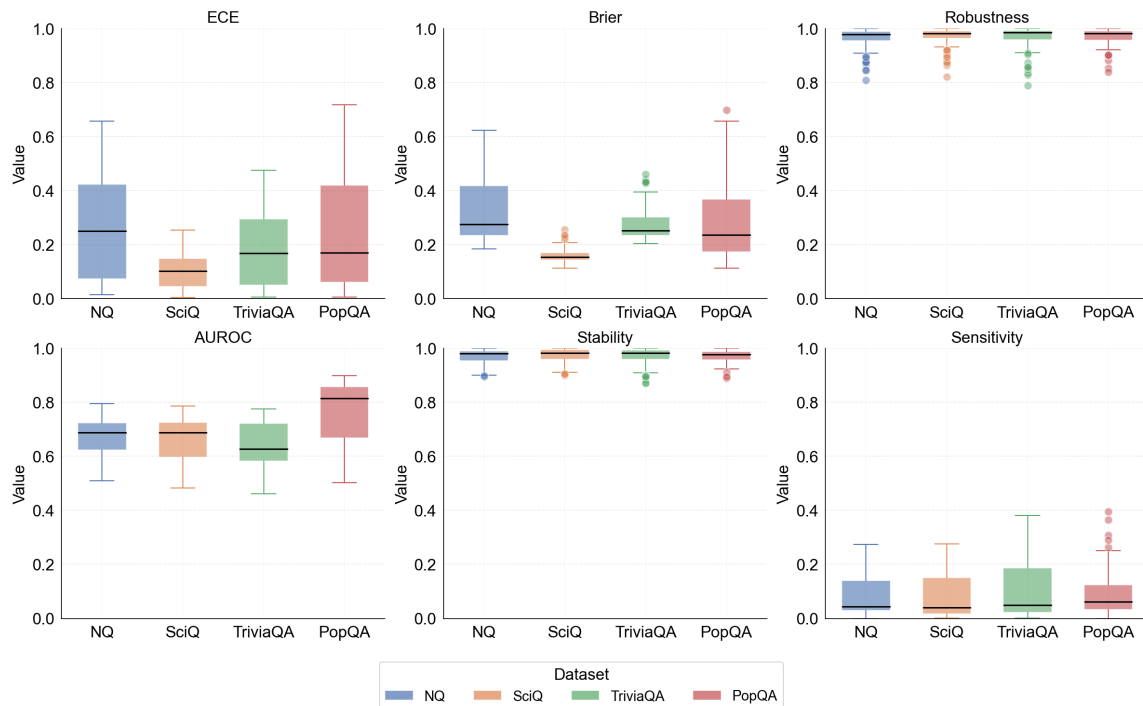


Figure 8: Distribution of metric across models and CE methods, shown per-dataset. Whiskers show the interquartile range and dots signify outliers.

		NQ					TriviaQA						
Method		ECE↓	Brier↓	AUROC↑	P-RB↑	A-STB↑	A-SST↑	ECE↓	Brier↓	AUROC↑	P-RB↑	A-STB↑	A-SST↑
Logit-based	Seq. Likelihood	0.40	0.37	0.72	0.95	0.94	0.16	0.28	0.28	0.70	0.95	0.92	0.20
	Boosted Prob.	0.57	0.54	0.71	0.99	0.99	0.03	0.38	0.37	0.71	0.99	0.99	0.03
	Platt Scaling	0.23	0.27	0.72	0.99	0.99	0.03	0.10	0.24	0.70	0.99	0.98	0.05
	P(True)	0.35	0.35	0.72	0.93	0.95	0.20	0.27	0.28	0.70	0.95	0.96	0.28
Aux. Ling. Internal States	Attention Score	0.04	0.23	0.58	0.99	0.99	0.03	0.03	0.23	0.56	0.99	0.99	0.03
	Hidden Score	0.05	0.22	0.61	0.98	0.98	0.04	0.02	0.24	0.52	0.99	0.99	0.02
	SAPLMA	0.26	0.27	0.75	0.87	0.91	0.18	0.25	0.28	0.67	0.86	0.89	0.23
	P(IK)	0.20	0.27	0.64	0.96	1.00	0.00	0.10	0.23	0.64	0.96	1.00	0.00
	Verbalized Conf.	0.55	0.53	0.60	0.99	0.99	0.05	0.37	0.37	0.60	0.99	0.99	0.08
	Calib1	0.06	0.22	0.66	0.98	0.97	0.07	0.07	0.23	0.60	0.99	0.98	0.06

Table 3: Calibration and our metrics averaged across all models. Arrows indicate metric direction (↓ lower is better, ↑ higher is better); the best value per column is bold.

Comparison	Agreement	Cohen’s κ
<i>Human agreement</i>		
Human 1 vs. Human 2	0.930	0.839
<i>Model-human agreement</i>		
GPT-4o vs. Human 1	0.930	0.842
GPT-4o vs. Human 2	0.900	0.774
GPT-5.5 vs. Human 1	0.930	0.832
GPT-5.5 vs. Human 2	0.910	0.784
<i>Model-model agreement</i>		
GPT-4o vs. GPT-5.5	0.920	0.812
<i>Aggregate agreement</i>		
GPT-4o vs. human consensus	0.946	–
GPT-5.5 vs. human consensus	0.952	–
All judges	0.855	–

Table 4: Agreement on response-correctness judgments among human and LLM evaluators. Human consensus is defined over items for which the two human annotators agreed. All-judge agreement is the proportion of items on which all four evaluators assigned the same label. Cohen’s κ is reported only for pairwise comparisons over the full sample.

Averaged across datasets, A-SST increases most clearly for answer-sensitive methods when the semantic groups also differ in correctness, as shown in Table 8.

The largest increases are observed for P(True) ($\Delta = 0.116$) and SAPLMA ($\Delta = 0.100$), both of which use information from the generated answer. In contrast, P(IK), which depends only on the question, shows no change ($\Delta = 0.000$).

These results provide supporting evidence for the intended interpretation of A-SST: for some answer-sensitive methods, confidence changes more when semantic answer differences coincide with differences in correctness.

Validator	Approval Rate
Human 1	0.975
Human 2	0.965
GPT-5.5	0.905
All validators	0.884

Table 5: Validator approval rates for the generated answer partitions.

B.9 Detailed Reciprocal Rank of CE Methods

The overall mean reciprocal rank (MRR) of different CE methods for each dataset is shown in Figure 11. We observe that CE method performance varies across datasets, and that no single method consistently performs best across all datasets and metrics.

Compared to ECE, Brier score, and AUROC, the cross-dataset variation is smaller for robustness, stability, and sensitivity. We hypothesize that this is because traditional calibration and discrimination metrics are more directly tied to correctness, and are therefore more strongly affected by dataset difficulty.

B.10 Case Study

Robustness. We show two robustness examples in Table 9 and Table 10. The results indicate that, across all prompts (t_1 – t_{10}), the Qwen-14B and Mistral models generate semantically equivalent answers. Although the confidence scores produced by different CE methods vary, most methods remain robust to prompt perturbations. In contrast, P(True) and Verbalized Conf. are noticeably less robust in these two cases ($P\text{-RB} = 0.5$), consistent with the trend observed in the main paper. We hypothesize that methods relying more heavily on

Metric	ECE	Brier	AUROC	P-RB	A-STB	A-SST	Acc.
ECE	1						
Brier	0.644*** [0.590, 0.696]	1					
AUROC	0.303*** [0.234, 0.368]	0.084* [0.019, 0.157]	1				
P-RB	-0.102** [-0.168, -0.032]	0.016 [-0.049, 0.078]	-0.330*** [-0.390, -0.270]	1			
A-STB	-0.148*** [-0.205, -0.090]	-0.061 [-0.123, 0.001]	-0.249*** [-0.309, -0.190]	0.502*** [0.436, 0.569]	1		
A-SST	0.214*** [0.155, 0.273]	0.120*** [0.058, 0.181]	0.262*** [0.200, 0.325]	-0.466*** [-0.531, -0.402]	-0.795*** [-0.824, -0.764]	1	
Acc.	-0.065 [-0.136, 0.004]	-0.169*** [-0.232, -0.104]	-0.162*** [-0.230, -0.093]	0.045 [-0.025, 0.115]	0.023 [-0.044, 0.089]	0.037 [-0.035, 0.109]	1

Table 6: Mean dataset-wise Kendall’s τ correlations among evaluation metrics and model accuracy. Brackets report cluster-bootstrap 95% confidence intervals. Asterisks indicate BH-corrected permutation significance: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

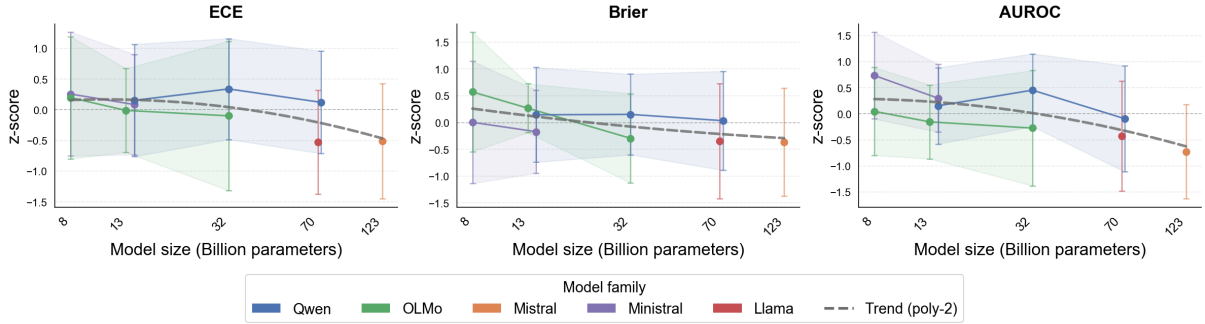


Figure 9: Normalized metric value distributions by model family and size. Whiskers and shaded regions indicate the standard deviation over observed metric values across datasets and CE methods. Models of the same family are connected and appear in the same color. Trend lines are second degree polynomials fitted on all observations.

	NQ	PopQA	SciQ	TriviaQA
$ I /M$	0.698	0.643	0.860	0.877

Table 7: Average fraction of prompts retained $|I|/M$ after semantic-equivalence filtering.

prompt formatting are more impacted by prompt perturbations.

Stability and Sensitivity. We present three qualitative examples illustrating stability and sensitivity in Table 11-Table 13. For stability, SAPLMA is the least stable method across semantically equivalent answers. We hypothesize that this is because SAPLMA is trained on the hidden state of the final generated token; therefore, small lexical changes at the end of the response can substantially alter the estimated confidence. For example, in Table 11, the semantically equivalent answers “Concave lens” and “Concave” end with different tokens, resulting

Method	A-SST _{same}	A-SST _{different}	Δ
Seq. Likelihood	0.183	0.200	0.017
Boosted Prob.	0.029	0.014	-0.015
Platt Scaling	0.039	0.043	0.004
P(True)	0.231	0.347	0.116
Attention Score	0.028	0.033	0.005
Hidden Score	0.040	0.046	0.006
SAPLMA	0.170	0.270	0.100
P(IK)	0.000	0.000	0.000
Verbalized Conf.	0.084	0.089	0.006
Calib1	0.067	0.083	0.016

Table 8: Mean A-SST for pairs of semantic groups with the same A-SST_{same} or different A-SST_{different} answer-correctness labels. The difference is calculated as $\Delta = \text{A-SST}_{\text{different}} - \text{A-SST}_{\text{same}}$.

in confidence scores of 0.07 and 1.00, respectively.

For sensitivity, P(True), which directly leverages generation-side information, performs best, followed by SAPLMA, which focuses on the final generated token. Methods such as sequence like-

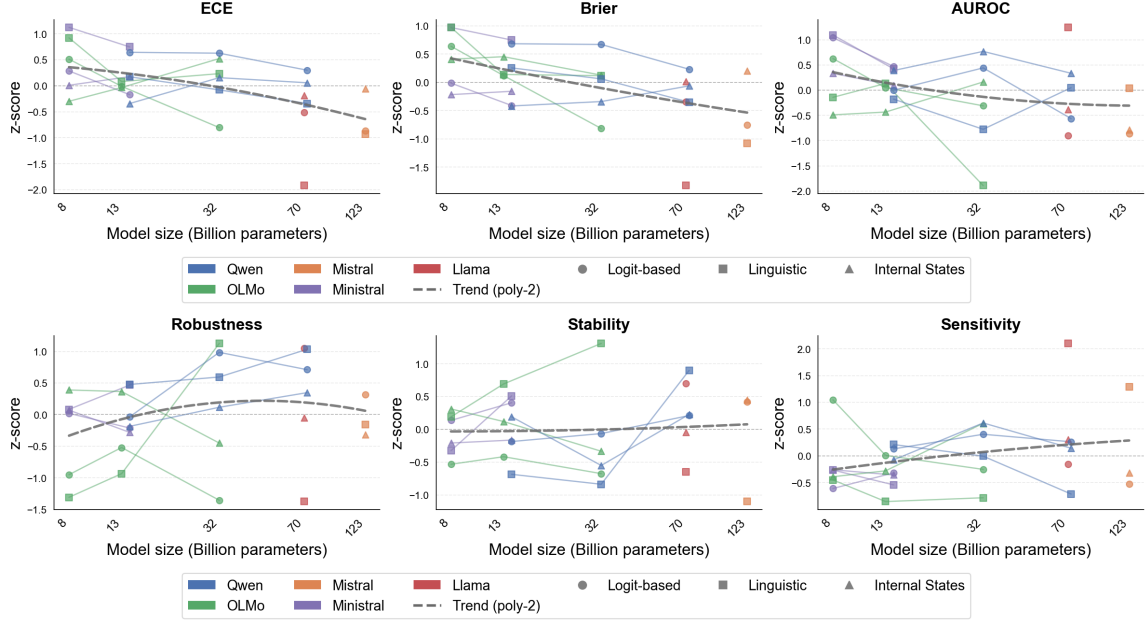


Figure 10: Normalized metric value distributions by model family and size. Confidence estimators from the same estimator family and for the same model family are connected and appear in the same color. Measurements from confidence estimators in the same family are shown with the same marker. Trend lines are second-degree polynomials fitted on all observations.

likelihood also use generation-side information, but since all candidate answers remain highly likely generations, the resulting confidence differences are relatively small. Post-hoc scaling methods, such as Platt scaling, further compress these differences and therefore reduce sensitivity. In contrast, methods that do not use answer-side information, such as P(IK), which is trained only on the hidden states of the final token of the question, are almost completely insensitive to answer changes (A-SST = 0). Although Calib1 conditions on both the question and answer, the short-form nature of our datasets likely reduces the contribution of the answer representation, resulting in moderate but still limited sensitivity.

B.11 Results of GPT-4o

We report the aggregated results for GPT-4o in Table 14. Due to limited access to the commercial model, we evaluate only five CE methods. Nevertheless, the observed trends are largely consistent with the main results reported in the paper. Note that, consistent with the main paper, we also use GPT-4o to judge the correctness of GPT-4o responses. However, since our proposed metrics do not depend on correctness labels, this setup does not affect our findings.

B.12 Robustness: Confidence-elicitation Prompts for Verbalized Confidence

In the main paper, we first use an *answer-elicitation prompt* (**Ans-elicit**) that instructs the LLM to produce an answer to the input question. At this stage, all the methods except verbalized confidence can already estimate confidence scores based on the generated answer. In contrast, verbalized confidence requires an additional *confidence-elicitation prompt* (**Conf-elicit**) to guide the LLM in verbalizing its confidence regarding the produced answer. To further investigate the effect of prompt perturbations on confidence estimation, we conduct additional experiments for verbalized confidence that only perturb the confidence-elicitation prompts instead. This design allows us to explore the robustness of verbalized confidence under two prompt perturbation settings: (1) perturb the answer-elicitation prompts and (2) perturb the confidence-elicitation prompts.

Perturbation of confidence-elicitation prompt.

To evaluate the robustness of verbalized confidence, we design seven variants of the confidence-elicitation prompt (see Table 15), grouped into three types: (i) Scale variants, which elicit confidence using numerical ranges of 01, 0100%, and 010; (ii) Lexical variants, which vary the key term used to request a score between 0 and 1 (“probability”, “certainty” and “confidence”); and (iii) Lin-

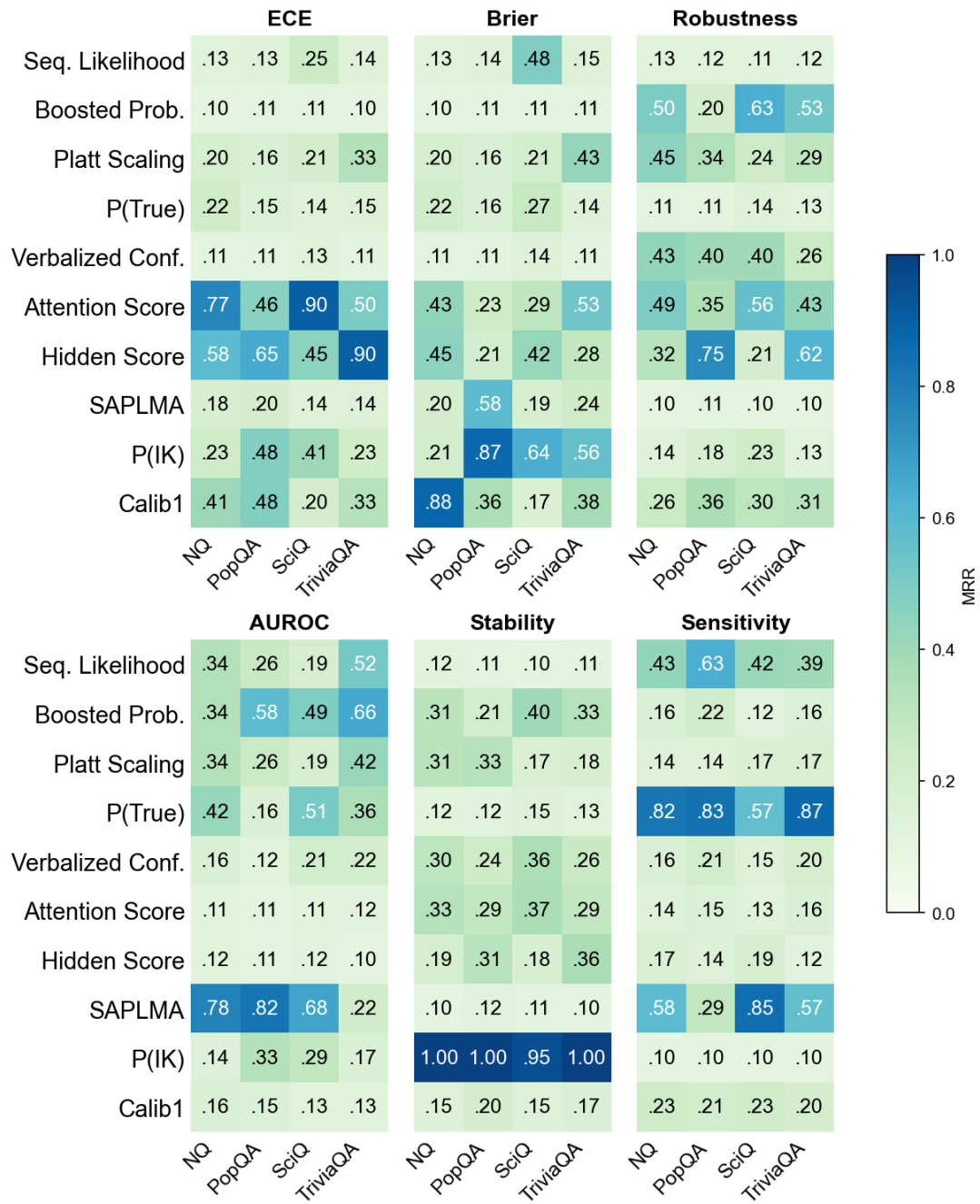


Figure 11: Reciprocal rank of CE methods per evaluation metric over datasets.

The electrode at which oxidation occurs is called? y : the anode. Corr.: 1.00												
CE		t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	P-RB
Qwen-14B	Seq. Likelihood	0.96	0.72	0.98	1.00	1.00	1.00	0.99	0.89	1.00	1.00	0.92
	Boosted Prob.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Platt Scaling	0.75	-	0.75	0.75	0.75	0.75	0.75	0.73	0.75	0.75	0.99
	P(True)	1.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.50
	Attention Score	0.81	0.81	0.80	0.81	0.81	0.81	0.80	0.80	0.81	0.81	1.00
	Hidden Score	0.85	0.86	0.84	0.84	0.83	0.86	0.84	0.85	0.83	0.83	0.99
	SAPLMA	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	0.98	0.99
	P(IK)	0.90	0.89	0.88	0.91	0.91	0.91	0.89	0.89	0.93	0.95	0.98
	Verbalized Conf.	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	1.00
	Calib1	0.77	0.76	0.77	0.76	0.76	0.78	0.77	0.75	0.74	0.76	0.99

Table 9: Case study 1 of prompt robustness. The confidence estimated by different CE methods under 10 perturbed answer-elic prompt prompts ($t_1 - t_{10}$) for the same question and LLM response is shown.

With an atomic weight of 22, what element, named for members of Greek mythology, uses the symbol Ti? y : Titanium ore. Corr.: 1.00												
CE		t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	P-RB
Mistral-123B	Seq. Likelihood	0.99	0.87	0.90	0.92	0.79	1.00	0.91	0.99	0.90	0.84	0.93
	Boosted Prob.	1.00	0.99	1.00	0.99	0.97	1.00	1.00	1.00	0.99	0.99	0.99
	Platt Scaling	0.73	0.70	0.71	0.71	0.68	0.73	0.71	0.73	0.71	0.70	0.99
	P(True)	0.08	0.01	0.05	0.01	0.65	0.08	0.03	0.08	0.02	0.84	0.72
	Attention Score	0.69	0.71	0.73	0.71	0.70	0.69	0.71	0.69	0.71	0.71	0.99
	Hidden Score	0.72	0.76	0.76	0.76	0.76	0.72	0.76	0.73	0.76	0.76	0.98
	SAPLMA	0.98	0.22	0.00	0.30	0.07	0.96	0.02	0.94	0.38	0.13	0.62
	P(IK)	0.70	0.51	0.45	0.70	0.67	0.69	0.52	0.42	0.76	0.57	0.89
	Verbalized Conf.	1.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.50
	Calib1	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	1.00

Table 10: Case study 2 of prompt robustness. The confidence estimated by different CE methods under 10 perturbed prompts ($t_1 - t_{10}$) for the same question and LLM response is shown.

guistic expressions, following Tian et al. (2023), which elicit confidence using predefined verbal scales from “Almost No Chance” to “Almost certain”, and a corresponding multiple-choice format (e.g., “a: Almost No Chance”). Those selected expressions are converted to numerical scores according to the rules in the next paragraph. All confidence outputs are normalized to a 01 scale for comparison.

Mapping of Linguistic Prompts. For the linguistic prompts, we followed Tian et al. (2023) to map the linguistic expressions to numerical scores. The specific mapping are [“Almost No Chance”, “Highly Unlikely”, “Chances are Slight”, “Little Chance”, “Unlikely”, “Probably Not”, “About Even”, “Better than Even”, “Likely”, “Probably”, “Very Good Chance”, “Highly Likely”, “Almost Certain”] to [0.02, 0.05, 0.1, 0.1, 0.2, 0.25, 0.5, 0.6, 0.7, 0.7, 0.8, 0.9, 0.95] according to Fagen-Ulmschneider (2023).

Results and Analysis. As shown in Table 16, when perturb only the answer-elic prompt prompts (Ans-

elicit), verbalized confidence produces more stable confidence estimates (P-RB = 0.986) compared to perturb confidence-elic prompt prompts are used (P-RB = 0.913). We hypothesize that this difference in stability arises because the generated answers remain largely consistent across perturbed answer-elic prompt prompts, leading to nearly identical instructions when subsequently eliciting confidence for those answers. In contrast, perturbations to the confidence-elic prompt prompts directly alter how the model is instructed to express its confidence, thereby introducing greater variability in the resulting confidence scores.

Prompt-wise Confidence Analysis for Verbalized Confidence. To further analyze the reason of the low robustness of verbalized confidence, we conduct a prompt-wise confidence analysis on selected models by examining the pairwise correlations of confidences between conf-elic prompt prompts for individual models (shown by Figure 12). We found that larger models maintain higher stability of confidence, demonstrating their ability to respond to diverse confidence-elic prompt prompts. LLMs remain

What type of lens is thicker at the edges than it is in the middle? y^* : concave lens						
		Convex	Concave lens	Concave		
CE		$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_2$	A-STB	A-SST
Llama-70B	Seq. Likelihood	0.59	0.59	0.82	0.90	0.17
	Boosted Prob.	1.00	1.00	1.00	1.00	0.00
	Platt Scaling	0.67	0.67	0.73	0.98	0.04
	P(True)	0.32	0.00	0.00	1.00	0.32
	Attention Score	0.84	0.84	0.84	1.00	0.00
	Hidden Score	0.90	0.77	0.78	1.00	0.11
	SAPLMA	1.00	0.07	1.00	0.60	0.23
	P(IK)	0.90	0.90	0.90	1.00	0.00
	Verbalized Conf.	1.00	0.00	0.00	1.00	1.00
	Calib1	0.87	0.81	0.78	0.99	0.09

Table 11: Case study 1 of answer stability and sensitivity. The confidence estimated by different CE methods for unique sampled answers to the same question is shown. Purple marks the answer group used for A-STB; purple and green mark the semantic groups used for A-SST.

where was the world chess tournament 2017 held? y^* : Tbilisi, Georgia							
New York City, USA Riyadh, Saudi Arabia New York City Carlsen vs. Caruana 2018							
CE	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	A-STB	A-SST	
Mistral-123B NQ	Seq. Likelihood	0.61	0.69	0.74	0.64	0.94	0.03
	Boosted Prob.	0.99	0.99	0.99	0.98	1.00	0.00
	Platt Scaling	0.60	0.62	0.63	0.61	0.99	0.01
	P(True)	0.05	0.87	0.04	0.01	0.99	0.03
	Attention Score	0.53	0.53	0.53	0.53	1.00	0.00
	Hidden Score	0.47	0.52	0.52	0.47	0.98	0.02
	SAPLMA	0.39	0.01	0.97	0.00	0.74	0.57
	P(IK)	0.79	0.79	0.80	0.80	1.00	0.00
	Verbalized Conf.	1.00	1.00	1.00	0.00	1.00	1.00
	Calib1	0.59	0.59	0.56	0.41	0.99	0.15

Table 12: Case study 2 of answer stability and sensitivity. Purple marks the answer group used for A-STB; purple and green mark the semantic groups used for A-SST.

What is the capital of Cao Wei? y^* : Luoyang						
		洛阳	Luo Yin	Luoyang		
CE		y_1	y_2	y_3	A-STB	A-SST
Qwen-14B	Seq. Likelihood	0.47	0.46	0.61	0.94	0.01
	Boosted Prob.	1.00	0.97	1.00	1.00	0.03
	Platt Scaling	0.49	0.49	0.52	0.99	0.00
	P(True)	1.00	0.00	1.00	1.00	1.00
	Attention Score	0.36	0.24	0.18	0.92	0.03
	Hidden Score	0.39	0.16	0.18	0.91	0.09
	SAPLMA	0.00	0.40	0.92	0.59	0.06
	P(IK)	0.23	0.23	0.23	1.00	0.00
	Verbalized Conf.	0.95	0.85	0.95	1.00	0.10
	Calib1	0.49	0.21	0.57	0.96	0.27

Table 13: Case study 3 of answer stability and sensitivity. Purple marks the answer group used for A-STB; purple and green mark the semantic groups used for A-SST. 洛阳 is the Chinese character of “Luoyang”.

Method	ECE ↓	Brier ↓	AUROC ↑	P-RB ↑	A-STB ↑	A-SST ↑
Seq. Likelihood	0.264	0.275	0.678	0.949	0.916	0.216
Platt Scaling	0.076	0.216	0.649	0.990	0.982	0.047
Verbalized Conf.	0.258	0.262	0.701	0.962	0.985	0.109
P(True)	0.295	0.296	0.733	0.976	0.988	0.226
Calib1	0.104	0.208	0.704	0.982	0.960	0.091

Table 14: GPT-4o results aggregated over all datasets. Due to the limited access of the commercial model, only methods that do not require information beyond the top 20 output logits apply to this model.

Concept	Prompt Name	Prompt
Scale Variants	P(1)	Provide the probability that your answer is correct. Give ONLY the probability between 0.0 and 1.0 , no other words or explanation.
	P(%)	Provide the probability that your answer is correct. Give ONLY the probability between 0% and 100% , no other words or explanation.
	P(10)	Provide the probability that your answer is correct. Give ONLY the probability between 0 and 10 , no other words or explanation.
Lexical Variants	CF(1)	Provide the confidence that your answer is correct. Give ONLY the confidence between 0.0 and 1.0, no other words or explanation.
	CT(1)	Provide the certainty that your answer is correct. Give ONLY the confidence between 0 and 10, no other words or explanation.
Linguistic Expressions	L.	Describe how likely it is that your answer is correct as one of the following expressions : ['Almost No Chance', 'Highly Unlikely', 'Chances are Slight', 'Little Chance', 'Unlikely', 'Probably Not', 'About Even', 'Better than Even', 'Likely', 'Probably', 'Very Good Chance', 'Highly Likely', 'Almost Certain']. Give ONLY the chosen expression, no other words or explanation.
	L. MC	Describe how likely it is that your answer is correct by choosing one of the following options : [a: 'Almost No Chance', b: 'Highly Unlikely', c: 'Chances are Slight', d: 'Little Chance', e: 'Unlikely', f: 'Probably Not', g: 'About Even', h: 'Better than Even', i: 'Likely', j: 'Probably', k: 'Very Good Chance', l: 'Highly Likely', m: 'Almost Certain']. Give ONLY the chosen option, no other words or explanation.

Table 15: Detail of prompts for eliciting confidence from LLMs. These prompts are specifically applied to verbalized confidence, we use these prompts to further validate the robustness of verbalized confidence in [Appendix B.12](#).

Dataset	Ans-elicited	Conf-elicited
NQ	0.985	0.907
PopQA	0.984	0.895
SciQ	0.990	0.928
TriviaQA	0.985	0.921
AVG	0.986	0.913

Table 16: Robustness of verbalized confidence under two prompt perturbation settings: perturbing answer-elicited prompt (Ans-elicited) and perturbing confidence-elicited prompt (Conf-elicited).

stable with the change of scales and lexical words in prompts. Specifically, prompts that perturbed with scale and lexical words change tend to have more correlated verbalized confidence compared to adding linguistic expressions in the prompt. Small models ($\leq 32B$) struggle more in maintaining robustness, especially with the multi-choice linguistic expression prompts.

B.13 Sensitivity: Alternative Sets of Semantically Different Answers

Table 17 reports sensitivity results obtained by using all semantically different answers outside $\hat{y}^{\max}$ (denoted as $\hat{y}^{\neq \max}$), instead of the original setting that compares $\hat{y}^{\max}$ with $\hat{y}^{\min}$. The results remain stable under this alternative definition and do not affect the main findings of the paper.

B.14 Instruction vs Reasoning Models

We compare the averaged results of instruction-tuned and reasoning models across all datasets; the

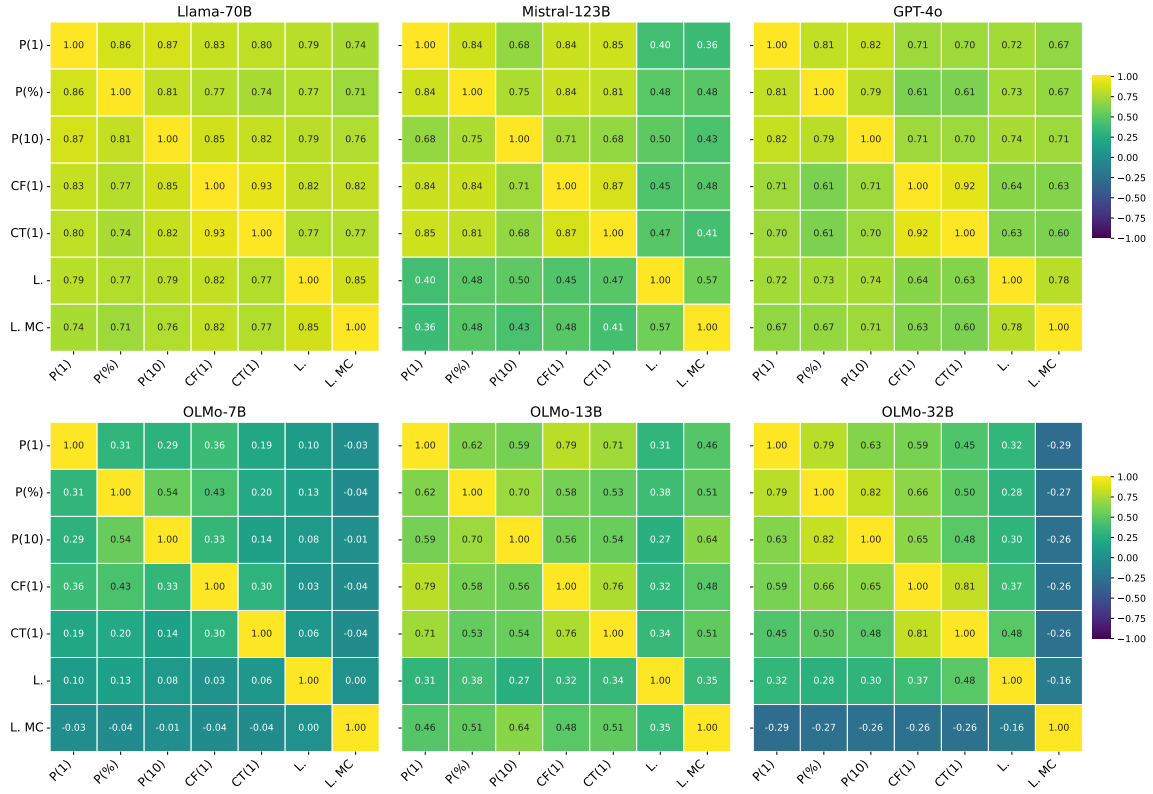


Figure 12: *Robustness*: Pearson correlations of verbalized confidence when using different confidence-elicit prompts in Table 15.

results are shown in Table 18-Table 21. Overall, both model types exhibit highly similar trends. We hypothesize that this is because the evaluated tasks consist of short-form question answering, which requires limited multi-step reasoning and therefore yields only minor differences between instruction-tuned and reasoning models.

Method	$\hat{y}_{\min}$	$\hat{y}_{\neq \max}$
Seq. Likelihood	0.182	0.171
Boosted Prob.	0.037	0.035
Platt Scaling	0.039	0.037
P(True)	0.228	0.227
Attention Score	0.032	0.032
Hidden Score	0.034	0.034
SAPLMA	0.180	0.178
P(IK)	0.000	0.000
Verbalized Conf.	0.057	0.054
Calib1	0.065	0.065

Table 17: Comparison of answer sensitivity settings. The original setting compares the largest semantic group with the smallest semantic group $\hat{y}_{\min}$; the alternative compares the largest group with all remaining semantically different answers $\hat{y}_{\neq \max}$.

Method	Instruction Model						Reasoning Model					
	ECE	Brier	AUROC	P-RB	A-STB	A-SST	ECE	Brier	AUROC	P-RB	A-STB	A-SST
Seq. Likelihood	0.41	0.39	0.71	0.95	0.98	0.17	0.38	0.32	0.75	0.95	0.94	0.15
Boosted Prob.	0.56	0.54	0.71	0.99	1.00	0.03	0.62	0.57	0.72	0.98	0.98	0.04
Platt Scaling	0.22	0.28	0.71	0.99	1.00	0.04	0.27	0.27	0.75	0.99	0.99	0.03
P(True)	0.33	0.35	0.72	0.93	0.97	0.19	0.40	0.37	0.74	0.95	0.94	0.21
Attention Score	0.05	0.23	0.58	0.99	0.99	0.03	0.04	0.20	0.61	0.99	0.99	0.03
Hidden Score	0.05	0.23	0.61	0.98	0.99	0.05	0.04	0.20	0.60	0.99	0.99	0.02
SAPLMA	0.26	0.28	0.74	0.87	0.97	0.20	0.23	0.24	0.77	0.89	0.93	0.15
P(IK)	0.21	0.27	0.63	0.96	1.00	0.00	0.18	0.24	0.69	0.96	1.00	0.00
Verbalized Conf.	0.52	0.51	0.60	0.99	1.00	0.06	0.64	0.62	0.63	0.99	0.98	0.03
Calib1	0.07	0.22	0.64	0.98	0.99	0.07	0.05	0.19	0.71	0.98	0.96	0.05

Table 18: NQ metric values averaged across instruction and reasoning models, respectively. Darker color indicates better performance.

Method	Instruction Model						Reasoning Model					
	ECE	Brier	AUROC	P-RB	A-STB	A-SST	ECE	Brier	AUROC	P-RB	A-STB	A-SST
Seq. Likelihood	0.09	0.15	0.68	0.94	0.97	0.17	0.04	0.15	0.71	0.94	0.94	0.15
Boosted Prob.	0.18	0.18	0.72	1.00	1.00	0.01	0.18	0.18	0.73	0.99	1.00	0.02
Platt Scaling	0.10	0.15	0.68	0.99	0.99	0.04	0.09	0.16	0.71	0.99	0.99	0.04
P(True)	0.16	0.17	0.73	0.95	0.99	0.18	0.11	0.15	0.74	0.97	0.99	0.19
Attention Score	0.02	0.15	0.55	0.99	1.00	0.02	0.02	0.16	0.56	0.99	1.00	0.02
Hidden Score	0.04	0.15	0.58	0.98	1.00	0.04	0.03	0.16	0.61	0.98	0.99	0.04
SAPLMA	0.15	0.17	0.75	0.91	0.98	0.22	0.18	0.19	0.72	0.84	0.95	0.18
P(IK)	0.05	0.14	0.69	0.98	1.00	0.00	0.11	0.17	0.67	0.98	1.00	0.00
Verbalized Conf.	0.14	0.17	0.62	0.99	1.00	0.03	0.16	0.18	0.71	0.99	1.00	0.03
Calib1	0.09	0.16	0.60	0.98	1.00	0.07	0.12	0.17	0.61	0.99	0.98	0.04

Table 19: SciQ metric values averaged across instruction and reasoning models, respectively. Darker color indicates better performance.

Method	Instruction Model						Reasoning Model					
	ECE	Brier	AUROC	P-RB	A-STB	A-SST	ECE	Brier	AUROC	P-RB	A-STB	A-SST
Seq. Likelihood	0.28	0.29	0.69	0.95	0.99	0.21	0.27	0.27	0.76	0.95	0.96	0.17
Boosted Prob.	0.37	0.36	0.70	0.99	1.00	0.03	0.43	0.41	0.76	0.99	0.99	0.05
Platt Scaling	0.08	0.23	0.69	0.99	1.00	0.05	0.18	0.25	0.76	0.99	0.99	0.04
P(True)	0.26	0.28	0.69	0.94	0.99	0.28	0.29	0.30	0.75	0.96	0.99	0.26
Attention Score	0.03	0.23	0.56	0.99	1.00	0.04	0.04	0.25	0.54	0.99	0.99	0.03
Hidden Score	0.02	0.23	0.52	1.00	1.00	0.01	0.03	0.25	0.54	0.99	1.00	0.02
SAPLMA	0.26	0.28	0.66	0.87	0.99	0.24	0.25	0.28	0.71	0.83	0.93	0.18
P(IK)	0.09	0.23	0.63	0.96	1.00	0.00	0.14	0.24	0.71	0.95	1.00	0.00
Verbalized Conf.	0.35	0.35	0.59	0.98	1.00	0.09	0.46	0.45	0.66	0.99	1.00	0.04
Calib1	0.06	0.23	0.59	0.99	1.00	0.05	0.08	0.24	0.63	0.98	0.98	0.06

Table 20: TriviaQA metric values averaged across instruction and reasoning models, respectively. Darker color indicates better performance.

Method	Instruction Model						Reasoning Model					
	ECE	Brier	AUROC	P-RB	A-STB	A-SST	ECE	Brier	AUROC	P-RB	A-STB	A-SST
Seq. Likelihood	0.41	0.33	0.83	0.95	0.99	0.21	0.30	0.21	0.88	0.96	0.97	0.16
Boosted Prob.	0.58	0.53	0.86	0.98	1.00	0.07	0.54	0.43	0.88	0.97	0.98	0.09
Platt Scaling	0.26	0.26	0.83	0.99	1.00	0.04	0.33	0.24	0.88	0.99	0.99	0.03
P(True)	0.31	0.31	0.80	0.92	0.99	0.26	0.42	0.35	0.80	0.95	0.98	0.23
Attention Score	0.05	0.21	0.61	0.99	1.00	0.05	0.04	0.17	0.62	0.99	0.99	0.03
Hidden Score	0.05	0.21	0.57	0.99	1.00	0.04	0.04	0.18	0.55	0.99	1.00	0.02
SAPLMA	0.13	0.15	0.88	0.92	0.99	0.11	0.15	0.15	0.88	0.93	0.97	0.06
P(IK)	0.08	0.16	0.82	0.97	1.00	0.00	0.08	0.13	0.86	0.98	1.00	0.00
Verbalized Conf.	0.59	0.56	0.65	0.98	1.00	0.08	0.71	0.68	0.68	0.99	1.00	0.03
Calib1	0.08	0.18	0.77	0.98	1.00	0.08	0.11	0.14	0.82	0.99	0.99	0.06

Table 21: PopQA metric values averaged across instruction and reasoning models, respectively. Darker color indicates better performance.

C Reproducibility Details

C.1 Use of AI Assistants

AI-based assistants were used to support code debugging, language polishing, grammar correction, and improving the clarity of presentation. The authors developed and verified all scientific contributions, experimental design, analyses, and conclusions. AI assistants were not used to generate experimental results or annotations without human oversight and verification.

C.2 Technical Details

We use the vLLM (vLLM Contributors, 2023) library for LLM inference and serving. All our experiments are conducted on NVIDIA HGX H100, which requires approximately 600 GPU hours to replicate.

C.3 Detailed Answer-elicite Prompts

In the main paper, all the CE methods are evaluated under the prompt variations in Table 22 for robustness.

C.4 LLM-as-a-Judge Prompts

The prompts used for GPT-4o as a judge model are shown in Table 23 and Table 24.

C.5 Evaluated LLMs

We evaluate 11 LLMs spanning five major families, which consist of instruction and reasoning models.

- From **Mistral**, we use Mistral-123B (Mistral-Large-Instruct-2411) (Mistral AI, 2024) and Ministral 8B and 14B reasoning models (Ministral-3-Reasoning-2512) (Mistral AI, 2025).

- From **Llama**, we test Llama-70B (Llama-3.3-70B-Instruct; Meta AI, 2024).
- From **Qwen**, we cover three model sizes: Qwen2.5-72B, Qwen2.5-32B, and Qwen2.5-14B (Qwen2.5-Instruct, Yang et al., 2024b).
- From **OLMo**, we assess a broad range of models (Team OLMo et al., 2024), including OLMo-2-0325-32B-Instruct (OLMo-32B), OLMo-2-1124-13B (OLMo-13B), and OLMo-2-1124-7B (OLMo-7B).
- From **GPT**, we include GPT-4o (gpt-4o-2024-11-20) (OpenAI, 2024), as it is a commercial model, confidence estimation methods (such as Boosted Prob. and internal states-based methods) that require beyond the top 20 output logits are not applicable.

C.6 Implementing Detail

Confidence Estimation Methods. Below are implementation details of the evaluated CE methods.

- **Seq. Likelihood.** We use the normalized probability of the generated response as a likelihood-based baseline. Specifically, we sum token log-probabilities over generated tokens only, excluding prompt tokens, and divide by the response length.
- **Boosted Prob.** Boosted probability is computed from generated-token probabilities following prior probability-based uncertainty estimators (Dinh and Niehues, 2025). We extract logits only for generated tokens and com-

Concept	Prompt Name	Prompt
Answer	A1	Answer the question, give ONLY the answer, no other words or explanation:
	A2	Answer the question, give ONLY the answer without explanation:
	A3	Answer the question, give ONLY the answer:
	A4	Answer the question as short as possible:
	A5	Answer the question with minimal words:
Provide	P1	Provide an answer for the question, give ONLY the answer, no other words or explanation:
	P2	Provide an answer for the question without explanation:
	P3	Provide an answer for the question, give ONLY the answer:
	P4	Provide as short an answer as possible to the question:
	P5	Provide an answer with minimal words for the question:

Table 22: Detail of prompts for eliciting answers from LLMs. These are the prompts used for evaluating robustness in the main paper.

pute the boosted-probability statistic over the model’s output distribution.

- **Platt Scaling (Platt, 1999).** Platt scaling trains two scalars over sequence likelihood and correctness labels. The input feature is sequence-likelihood, and the scalars learn a one-dimensional affine transformation followed by a sigmoid.
- **P(True) (Kadavath et al., 2022).** P(True) is implemented as a self-evaluation query to the same language model. We construct a binary verification prompt containing the question and proposed answer, ending with a request to choose whether the proposed answer is “True” or “False”. We then compute the normalized probability of the True option from the model’s next-token log probabilities.
- **Calib1 (Xia et al., 2025a).** Calib1 is implemented as a supervised text-pair calibrator. We encode the question and answer with a BERT-style encoder and train a classifier to predict answer correctness. The calibrator operates only on the textual question-answer pair and does not require access to internal states of the target LLM. We train the classifier with focal loss (Mukhoti et al., 2020) and use the predicted positive-class probability as the confidence score.
- **Verbalized Confidence (Tian et al., 2023).** Verbalized confidence is implemented as a second-stage prompting method. After the model produces an answer, we ask it to explicitly report a confidence value.

- **Attention Score (Sriramanan et al., 2024).** The attention score is implemented as a white-box statistic over transformer attention. During local inference, we extract attention from a mid-to-late layer, approximately 70% through the model depth. We immediately reduce the attention tensor to a scalar using an eigenspectrum-based statistic. The scalar attention feature is then mapped to a confidence score with a logistic regression calibrator trained on correctness labels.
- **Hidden Score (Sriramanan et al., 2024).** Similar to the attention score, the hidden score is implemented from the same mid-to-late layer hidden representations. We extract the relevant hidden-state vector and compute a singular-value-based scalar statistic. We then train a logistic regression calibrator over this scalar feature to produce a normalized confidence score.
- **SAPLMA (Azaria and Mitchell, 2023).** SAPLMA is implemented as a learned probe over the language model’s final-layer hidden state. For each generated answer, we use the hidden state associated with the final generated-token position. We train an MLP classifier on these fixed representations to predict correctness. The LLM parameters are frozen throughout training; only the MLP probe is optimized. The confidence score is the classifier’s positive-class probability.
- **P(IK) (Kadavath et al., 2022).** P(IK) is implemented as a pre-generation linear probe. Instead of using a representation from the gen-

Semantic Equivalence Judgment Prompt for Correctness

You are evaluating whether an answer is semantically equivalent to the reference answer. Judge meaning rather than exact wording. Two answers should be considered equivalent if they express the same meaning, even when using different phrasing, synonyms, or slightly different names referring to the same entity.

Also, treat answers as equivalent even if they include expressions of confidence or uncertainty (e.g., 'I'm pretty sure it's X' vs. 'X'). Focus on the core meaning of the answers and ignore differences in:

- wording or phrasing
- length
- fillers
- strengtheners or weakeners

Examples:

Answer: "Mount Everest" | Reference: "Everest" → YES;

Answer: "Albert Einstein" | Reference: "Einstein" → YES;

Answer: "Paris" | Reference: "The capital of France" → YES;

Answer: "Pacific Ocean" | Reference: "Atlantic Ocean" → NO;

Answer: "Jupiter" | Reference: "I think it's Jupiter" → YES;

Answer: "Mars" | Reference: "Maybe Venus" → NO.

Answer only with:

1 if semantically equivalent;

0 otherwise.

Question: <question>

Answer: <answer>

Reference answer: <target>

Equivalent:

Table 23: Prompt for judging correctness.

Semantic Equivalence Judgment Prompt for Semantic Grouping

You are grouping 10 answers by their semantic equivalence. Your goal is to judge meaning, not exact wording. Two answers should belong to one group if they are semantically equivalent, even if they use different phrasing, synonyms, or slightly different names that refer to the same entity.

Also, treat answers as equivalent even if they include expressions of confidence or uncertainty (e.g., 'I'm pretty sure it's X' vs. 'X'). Focus on the core meaning of the answers and ignore differences in:

- wording or phrasing
- length
- fillers
- strengtheners or weakeners

Examples:

- Answer: 'Mount Everest' | Reference: 'Everest' → YES

- Answer: 'Albert Einstein' | Reference: 'Einstein' → YES

- Answer: 'Paris' | Reference: 'The capital of France' → YES

- Answer: 'Pacific Ocean' | Reference: 'Atlantic Ocean' → NO

- Answer: 'Jupiter' | Reference: 'I think it's Jupiter' → YES

- Answer: 'Mars' | Reference: 'Maybe Venus' → NO

Answer in a dict format only, answers in the same group fall into one dict, for example:

{1: answer1, 2: ...}

{4: answer4, 5: ...}

Question: <question>

Answer: <sampled answers >

Table 24: Prompt for semantic grouping.

erated answer, we use the final-layer hidden state corresponding to the question's last token. A single linear layer with sigmoid output is trained to predict whether the model will answer correctly. The language model remains frozen, and only the linear head is trained.