Uncertainty in Language Models: Assessment through Rank-Calibration

Anonymous ACL submission

Abstract

Language Models (LMs) have shown promising performance in natural language generation. However, as LMs often generate incorrect or hallucinated responses, it is crucial to correctly quantify their uncertainty in responding to given inputs. In addition to verbalized confidence elicited via prompting, many uncertainty measures (*e.g.*, semantic entropy and affinity-graph-based measures) have been proposed. However, these measures can differ greatly, and it is unclear how to compare them, partly because they take values over different ranges (*e.g.*, $[0, \infty)$ or $[0, 1]$). In this work, we address this issue by developing a novel and practical framework, termed *Rank-Calibration*, to assess uncertainty and confidence measures for LMs. Our key tenet is that higher uncertainty (or lower confidence) should imply lower generation quality, on average. Rank-calibration quantifies deviations from this ideal relationship in a principled manner, without requiring ad hoc binary thresholding of the correctness score (*e.g.*, ROUGE or METEOR). The broad applicability and the granular interpretability of our methods are demonstrated empirically.

1 Introduction

Language Models (LMs), especially Large Language Models (LLMs), have shown promising performance in Natural Language Generation (NLG). These models, fitted on huge text corpora, can produce responses resembling those of humans (Touvron et al., 2023b; OpenAI, 2023). However, since LMs often generate wrong or hallucinated responses (Weidinger et al., 2021; Xiao and Wang, 2021), it is crucial to correctly quantify their level of uncertainty in responding to particular inputs.

Uncertainty quantification is well-explored in supervised learning, specifically in classification (*e.g.*, Lichtenstein et al., 1977; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017, etc). In classification, a *confidence measure* is an estimate of the probability that the predicted class $\widehat{Y}$ matches the true class label Y (Lichtenstein et al., 1977; Lee et al., 2023). A confidence measure C is considered *calibrated* if it reflects the probability of correct prediction, *i.e.*, $\mathbb{P}(\widehat{Y} = Y \mid C) = C$, for all values in C's range. The Expected Calibration Error (ECE) measures the miscalibration of a confidence measure (Harrell, 2015; Naeini et al., 2015):

$$\mathbb{E}_C\left[\left|\mathbb{P}(\widehat{Y}=Y\,|\,C)-C\right|\right]. \qquad \text{(ECE)}$$

In classification, confidence measures are predominantly built on model logits (Guo et al., 2017; Kull et al., 2019). However, these methods are less suitable for NLG tasks. First, the label space is often too large to assess correctness via $\widehat{Y} = Y$, since LMs produce potentially long textual responses $\widehat{Y}$ for any given input. Second, for LMs, logits encode the likelihood of selecting the next token and do not necessarily capture linguistic sense (Mielke et al., 2022). Third, even hand-crafted prompts intended to make LMs express confidence explicitly

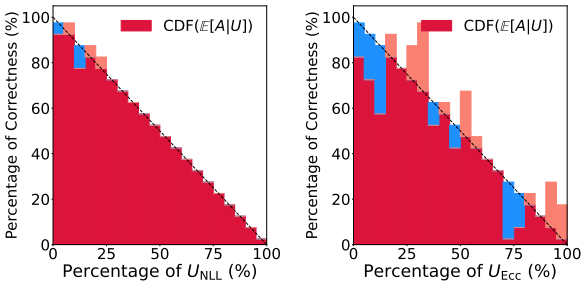


Figure 1: *Indication diagrams* comparing two uncertainty measures, U_{NLL} (negative log-likelihood) and U_{Ecc} (eccentricity), for the GPT-3.5-turbo model on the TriviaQA benchmark. The red bars indicate the average correctness of different outputs, as a function of the corresponding relative uncertainty levels. The blue and shallow red areas—deviating from the anti-diagonal line—indicate where the uncertainty measures are over-optimistic and pessimistic, respectively. Their sum is our *rank-miscalibration* metric (*i.e.*, RCE), which here is lower for U_{NLL} *than* U_{Ecc}. See Sec. 4.3 for details.

may not lead to reliable confidence values because elicitation is heavily tied to prompt formats (Zhao et al., 2021; Xiong et al., 2024).

Recent works have studied *uncertainty measures* as an alternative to confidence measures. These capture the "dispersion" of an LMs' potential outputs for a fixed input. Kuhn et al. (2023) introduce *semantic entropy*, which incorporates linguistic invariances arising from the shared meaning of generated responses. Lin et al. (2023) extend semantic entropy by leveraging the affinity matrices induced by entailment scores of generated outputs. Further, Chen et al. (2024) characterize differential entropy in the embedding space with EigenScore, via the covariance of embeddings of potential responses.

Uncertainty measures are more general and arguably more principled than confidence measures for LMs, but they lack a universal assessment metric such as ECE. A key issue is that uncertainty measures are not necessarily commensurate. For instance, the semantic entropy (Kuhn et al., 2023) can take arbitrarily large positive values, whereas the EigV measure of Lin et al. (2023) depends on the number of responses generated. This makes it difficult to understand, evaluate, and compare uncertainty measures via a unified lens.

This paper develops a principled framework to assess the quality of uncertainty and confidence measures for LMs. We provide a novel and practical framework, termed *Rank-Calibration*. Specifically, our contributions are as follows.

- We mathematically formalize the assessment of uncertainty/confidence measures for LMs in NLG tasks, going beyond binary correctness.
- We demonstrate empirically that existing assessment metrics (*e.g.*, AUROC, ECE, etc) have several limitations, including a heavy dependence on the LM's performance, instability caused by ad hoc binarization of correctness scores, and incompatibility with diverse uncertainty ranges.
- We address these limitations by starting from a basic principle: lower uncertainty/higher confidence should indicate higher-quality generation. We thus propose assessing uncertainty measures in terms of rank-calibration and introduce a suitable metric, the Rank-Calibration Error (RCE).
- To make rank-calibration practical, we introduce the Empirical RCE—an estimate of RCE based on a finite dataset. Moreover, we introduce novel indication diagrams, previewed in Fig. 1, that intuitively visualize the deviation of any uncertainty/confidence measure from the monotonicity required for rank-calibration.
- We experimentally demonstrate the broader applicability and granular interpretability of our proposed methods. Comprehensive ablation studies are conducted to examine its robustness.

2 Correctness and Uncertainty for LMs

Let $\mathcal{V}$ be the token vocabulary of an LM and $\mathcal{V}^\star := \cup_{\ell\geq 1}\mathcal{V}^\ell$ the space of sequences of arbitrary length. Given a query $\boldsymbol{x} \in \mathcal{V}^\star$, an LM $\mathcal{M}$ can generate output $\widehat{\boldsymbol{y}} \triangleq (\widehat{y}_\ell)_{\ell\geq 1} \in \mathcal{V}^\star$ by sequentially sampling from the distribution $\mathbb{P}(\widehat{\boldsymbol{y}}\,|\,\boldsymbol{x}) := \prod_{\ell\geq 1}\mathbb{P}(\widehat{y}_\ell\,|\,\boldsymbol{x}, \widehat{y}_{<\ell})$. Here, $\widehat{y}_\ell \in \mathcal{V}$ is the ℓ-th generated token and $\mathbb{P} \triangleq \mathbb{P}^{\mathcal{M}}$ is the generative distribution of $\mathcal{M}$.

We work with a deterministic *correctness function* $A\colon \mathcal{V}^\star \times \mathcal{V}^\star \to \mathbb{R}$ mapping each pair $(\boldsymbol{x}; \widehat{\boldsymbol{y}})$ to a correctness value $A(\boldsymbol{x}; \widehat{\boldsymbol{y}})$. In practice, correctness is often not a binary variable in NLG tasks and can be assessed in at least two different ways.

- **Reference matching.** Given certain reference answers $\{\boldsymbol{y}^{(m)}\}_{m=1}^M$ associated with $\boldsymbol{x}$, a similarity score between the output $\widehat{\boldsymbol{y}}$ and $\{\boldsymbol{y}^{(m)}\}_{m=1}^M$ can be interpreted as a correctness value. Similarity scores commonly utilized for this purpose include the *Rouge* score, *BLEU* score, and outputs of other discriminative LMs.
- **Human evaluation.** Correctness or quality may be evaluated by human experts, possibly integrating multiple opinions (*e.g.*, averaging). This approach does not require reference answers and is as "trustworthy" as the humans involved.

Notation	Description
$\mathcal{V}$	Token vocabulary
$\mathcal{V}^\star$	Space of token sequences
$\boldsymbol{x}$	Input context, $\boldsymbol{x} \in \mathcal{V}^\star$
$\widehat{\boldsymbol{y}}$	Gen. output $\widehat{\boldsymbol{y}} = (\widehat{y}_\ell)_{\ell\geq 1} \in \mathcal{V}^\star$
$\mathbb{P} \triangleq \mathbb{P}^{\mathcal{M}}$	Generative dist. of LM $\mathcal{M}$
$A(\cdot\,;\,\cdot)$	A deterministic correctness function
$\{\boldsymbol{y}^{(m)}\}_{m=1}^M$	Reference answers for input $\boldsymbol{x}$
$U^{\mathcal{M}}(\boldsymbol{x}; \widehat{\boldsymbol{y}})$	Uncertainty measure for LM $\mathcal{M}$
$C^{\mathcal{M}}(\boldsymbol{x}; \widehat{\boldsymbol{y}})$	Confidence measure for LM $\mathcal{M}$
$\mathrm{reg}(u)$	Regression fn. $\mathbb{E}_{\boldsymbol{x},\widehat{\boldsymbol{y}}}[A \mid U = u]$

Table 1: Summary of notations.

An *uncertainty measure* is a (possibly random) function $U^{\mathcal{M}}\colon \mathcal{V}^\star \times \mathcal{V}^\star \to \mathbb{R}$, $(\boldsymbol{x}; \widehat{\boldsymbol{y}}) \mapsto U^{\mathcal{M}}(\boldsymbol{x}; \widehat{\boldsymbol{y}})$ associated with the LM that maps any pair $(\boldsymbol{x}; \widehat{\boldsymbol{y}})$ to

an uncertainty value.[1] We will omit $\mathcal{M}$ and write $U(\boldsymbol{x}; \widehat{\boldsymbol{y}})$, $\mathbb{P}(\cdot \mid \boldsymbol{x})$ when the choice of the LM is clear. Some examples are reviewed below, while additional examples and details are in Appendix B.

- **NLL.** In classification, the softmax of the last-layer logits determines a model's prediction (Guo et al., 2017). In NLG tasks, one can view the Negative Log-Likelihood (NLL),

$$U_{\mathrm{NLL}}(\boldsymbol{x}, \widehat{\boldsymbol{y}}) := -\ln(\mathbb{P}(\widehat{\boldsymbol{y}} \mid \boldsymbol{x})),$$

  as an indicator of uncertainty where $\widehat{\boldsymbol{y}} = (\widehat{y}_\ell)_{\ell \geq 1}$ is a generated response. A natural extension accounting for the length of responses applies length normalization; this is also known as the *Perplexity* measure (Jelinek et al., 1977).
- **Entropy.** The predictive entropy of the distribution $\mathbb{P}(\cdot \mid \boldsymbol{x})$ is large when the same input may lead to diverse outputs, and it is defined as

$$U_{\mathrm{E}}(\boldsymbol{x}) := -\mathbb{E}_{\widehat{\boldsymbol{y}} \sim \mathbb{P}(\cdot|\boldsymbol{x})}[\ln(\mathbb{P}(\widehat{\boldsymbol{y}} \mid \boldsymbol{x}))].$$

  Malinin and Gales (2021) propose a variant of this, $U_{\text{E-LN}}(\boldsymbol{x})$, utilizing the length-normalized log-likelihood $\ln(\mathbb{P}(\widehat{\boldsymbol{y}}|\boldsymbol{x}))/\mathrm{len}(\widehat{\boldsymbol{y}})$. Kuhn et al. (2023) argue that different responses with the same meaning should be viewed as equals in this context, regardless of token-level differences. They propose the semantic entropy,

$$U_{\mathrm{SE}}(\boldsymbol{x}) := -\mathbb{E}_{\widehat{\boldsymbol{y}} \sim \mathbb{P}(\cdot|\boldsymbol{x})}[\ln(\mathbb{P}(c(\widehat{\boldsymbol{y}}) \mid \boldsymbol{x}))],$$

  where $c(\widehat{\boldsymbol{y}})$ is the semantic concept of $\widehat{\boldsymbol{y}}$, provided by another language modeling method.
- **Affinity graph.** Lin et al. (2023) calculate uncertainty using a weighted adjacency graph built upon semantic affinities. Consider an *affinity* model e, mapping pairs of responses $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{y}}'$ to values in $[0, 1]$. Given K independent samples $\{\widehat{\boldsymbol{y}}^{(k)}\}_{k=1}^K$ from $\mathbb{P}(\cdot \mid \boldsymbol{x})$, the model e induces a symmetric adjacency matrix $W = [w_{i,j}]_{i,j=1}^K$, with $w_{i,j} = (e(\widehat{\boldsymbol{y}}^{(i)}; \widehat{\boldsymbol{y}}^{(j)}) + e(\widehat{\boldsymbol{y}}^{(j)}; \widehat{\boldsymbol{y}}^{(i)}))/2$ for all i, j. Let $D = [\mathbb{1}[j = i] \sum_{k=1}^K w_{k,j}]_{i,j=1}^K$ be the corresponding degree matrix and $\{\lambda_k\}_{k=1}^K$ be the eigenvalues of the *Laplacian* $L = I - D^{-1/2} W D^{-1/2}$. Then, the uncertainty measures proposed in Lin et al. (2023) include

$$U_{\mathrm{EigV}}(\boldsymbol{x}) := \sum_{k=1}^{K} \max\{0, 1 - \lambda_k\},$$

$$U_{\mathrm{Deg}}(\boldsymbol{x}) := 1 - \mathrm{trace}(D)/K^2,$$

$$U_{\mathrm{Ecc}}(\boldsymbol{x}) := \|[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K]\|_2,$$

  where $\{\boldsymbol{v}_k\}_{k=1}^K$ are suitable vectors associated with L, see Lin et al. (2023). Intuitively, $U_{\mathrm{EigV}}(\boldsymbol{x})$ approximately counts the connected components in the graph represented by W, while $U_{\mathrm{Deg}}(\boldsymbol{x})$ and $U_{\mathrm{Ecc}}(\boldsymbol{x})$ reflect the diversity of outputs.

The diverse uncertainty measures reviewed above produce outputs with different ranges. For instance, U_{NLL}, U_{SE}, and U_{EigV} can yield any number in $[0, \infty)$, whereas U_{Deg} and U_{Ecc} are bounded in $[0, 1]$; see Fig. 3 [bottom] for a visual illustration. This mismatch in output ranges motivates the need for a novel unified assessment framework.

As we shall see, our assessment framework can handle not only any uncertainty measure but also the closely related concept of *confidence measures* (Zhao et al., 2021; Mielke et al., 2022; Xiong et al., 2024). A confidence measure can be cast as a (possibly random) function $C^{\mathcal{M}} : \mathcal{V}^\star \times \mathcal{V}^\star \to [0, 1]$, $(\boldsymbol{x}; \widehat{\boldsymbol{y}}) \mapsto C^{\mathcal{M}}(\boldsymbol{x}; \widehat{\boldsymbol{y}})$ with output taking values in $[0, 1]$. Intuitively, confidence and uncertainty measures serve similar purposes, although in a complementary way—high confidence should correlate with low uncertainty, and vice versa.

With this notation in place, we are now ready to state our goals and give a more detailed preview of our proposed framework. Given a benchmark dataset $\{(\boldsymbol{x}_i, \{\boldsymbol{y}_i^{(m)}\}_{m=1}^{M_i})\}_{i=1}^n$, where each $M_i \geq 0$ denotes the number of reference answers for $\boldsymbol{x}_i$, we aim to quantify the performance of an uncertainty measure U (or a confidence measure C) as follows. First, we obtain the paired values of uncertainty and correctness $\{(U(\boldsymbol{x}_i, \widehat{\boldsymbol{y}}_i), A(\boldsymbol{x}_i; \widehat{\boldsymbol{y}}_i))\}_{i=1}^n$ by independently sampling $\widehat{\boldsymbol{y}}_i \sim \mathbb{P}(\cdot \,|\, \boldsymbol{x}_i)$ for each $1 \leq i \leq n$. Then, we evaluate $\mathcal{E}(\{(U(\boldsymbol{x}_i, \widehat{\boldsymbol{y}}_i), A(\boldsymbol{x}_i; \widehat{\boldsymbol{y}}_i))\}_{i=1}^n)$ for each $1 \leq i \leq n$, using a suitable metric $\mathcal{E}$.[2] To account for the randomness in sampling $\widehat{\boldsymbol{y}}_i$, we may draw multiple independent responses $\{\widehat{\boldsymbol{y}}_i^{(k)}\}_{k=1}^K \overset{iid}{\sim} \mathbb{P}(\cdot \mid \boldsymbol{x}_i)$ and take the average as the final result $\sum_{k=1}^K \mathcal{E}(\{(U(\boldsymbol{x}_i, \widehat{\boldsymbol{y}}_i^{(k)}), A(\boldsymbol{x}_i, \widehat{\boldsymbol{y}}_i^{(k)}))\}_{i=1}^n)/K$.

3 Limitations of Existing Assessments

This section illustrates some limitations of existing assessments for LM uncertainty measures via a case study applying the *GPT-3.5-turbo* (Ouyang et al., 2022) model on the *TriviaQA* benchmark (Joshi et al., 2017). We use the validation

[1] In special cases, the uncertainty measure may only depend on the input $\boldsymbol{x}$ and the LM $\mathcal{M}$, not the output $\widehat{\boldsymbol{y}}$.

[2] A common practice is to map the correctness values to $\{0, 1\}$ by thresholding at an ad hoc value before feeding them into the evaluation metric; see Sec. 3 for a discussion of the limitations of this approach.

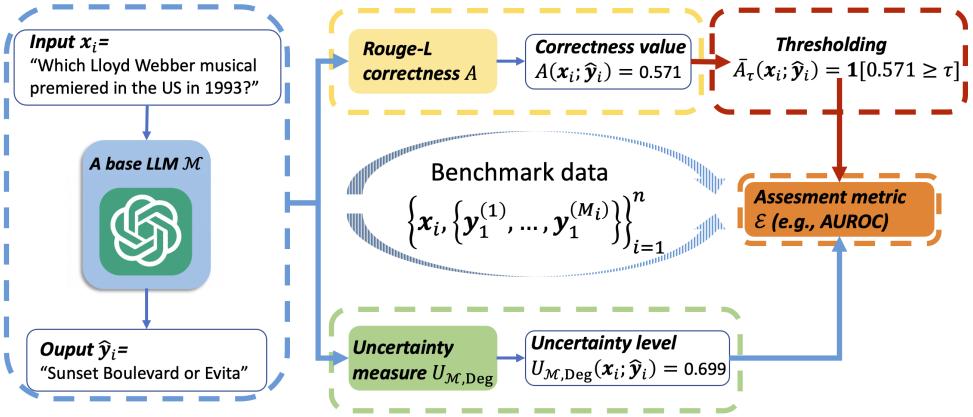


Figure 2: Common workflow for assessing the quality of an LM uncertainty/confidence measure. The key ingredients are: a base LM $\mathcal{M}$ (*e.g.*, Llama-2-7b-chat), a correctness function A (*e.g.*, the Rouge-L score), a benchmark dataset $\{\boldsymbol{x}_i, \{\boldsymbol{y}_i^{(m)}\}_{m=1}^{M_i}\}_{i=1}^n$ (*e.g.*, TriviaQA), an assessment metric $\mathcal{E}$ (*e.g.*, AUROC), and the uncertainty measure U (*e.g.*, U_{Deg}). The workflow proceeds in five stages: generation, correctness calculation, correctness discretization, uncertainty quantification, and evaluation. Notably, the threshold τ in correctness discretization is usually chosen heuristically (Kuhn et al., 2023; Xiong et al., 2024; Lin et al., 2023, etc), which can be problematic, as demonstrated in Sec. 3. Our proposed RCE-based assessment *removes* this stage by using the correctness values directly.

set of TriviaQA, which contains 11,322 question-answer pairs (after deduplication). We use the same prompt template as that in Lin et al. (2023). The template is shown in Appendix E.2.

The uncertainty measures examined here include the negative log-likelihood U_{NLL}, the semantic entropy U_{SE} (Kuhn et al., 2023), the affinity-graph-based measures U_{EigV}, U_{Ecc}, and U_{Deg} (Lin et al., 2023), with affinity determined by the NLI model (He et al., 2021), and the verbalized confidence C_{Verb} (Xiong et al., 2024); see definitions in Appendix B. These include both white box and grey box measures,[3] as well as a diversity of prompt strategies. We use the *Rouge-L* score as the correctness function A. We follow a common assessment pipeline (Kuhn et al., 2023; Lin et al., 2023; Xiong et al., 2024), as depicted in Fig. 2. The assessment metrics are detailed in Appendix C.

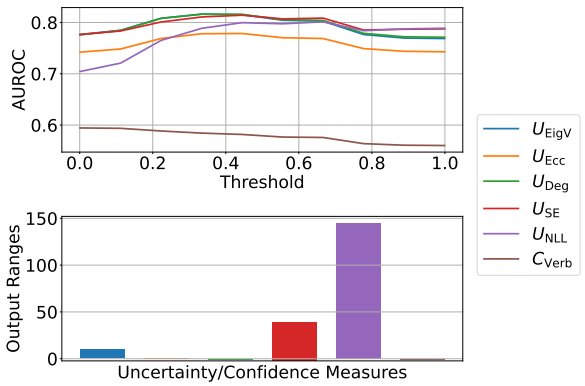


Figure 3: Top: AUROCs of uncertainty/confidence measures with various thresholds. Bottom: Output ranges of uncertainty/confidence measures. Both results are for *GPT-3.5-turbo* on the *TriviaQA* benchmark.

[3]The grey-box oracle refers to the access to model logits, which is partly feasible for commercial LMs, while the black-box oracle only relies on generated outputs.

Ad hoc correctness thresholding. Most existing assesment metrics (*e.g.*, AUROC, AUPRC, ECE, etc) are rooted in classification and require binary labels (*i.e.*, $A \in \{\textit{True} \text{ or } \textit{False}\}$). Consequently, an ad hoc threshold $\tau \in \mathbb{R}$ is often introduced to map continuous correctness values to binary labels, *i.e.*, $\bar{A}_\tau(\boldsymbol{x}; \widehat{\boldsymbol{y}}) := \mathbb{1}[A(\boldsymbol{x}; \widehat{\boldsymbol{y}}) \geq \tau]$ (Lin et al., 2023; Kuhn et al., 2023). Thus, the response is viewed as *correct* if the correctness value $A(\boldsymbol{x}; \widehat{\boldsymbol{y}})$ is at least τ, and *incorrect* otherwise.

However, thresholding can lead to inconsistencies. Taking AUROC as an example, we plot the assessed results of uncertainty/confidence measures under varying thresholds in Fig. 3 [top]. The relative AUROC results of distinct measures vary drastically with the choice of τ. For example, U_{NLL} appears inferior to other methods if $\tau < 0.2$, but it becomes the best measure if $\tau > 0.8$. This is especially concerning given that there seems to be no principled way to set this threshold. The same limitation also affects other metrics (*e.g.*, AUPRC, AUARC) and configurations; see Appendix E.4.

Diverse output ranges. The second limitation of existing assessments is rooted in the diverse output ranges of the uncertainty or confidence measures. As shown in Fig. 3 [bottom], the output ranges of different uncertainty measures vary significantly. For example, the values of U_{SE} can be higher than 100 while the values of U_{Ecc} and U_{Deg} are small by definition. This diversity of output ranges prevents the direct use of calibration-based metrics such as the ECE, which takes variables with inputs in $[0, 1]$.

Strong dependence on LM performance. While the quality of uncertainty/confidence measures should be disentangled from the generation performance of the LM, there is often a strong relation

between the two concepts. We argue that many existing metrics (*e.g.*, AUROC, AUPRC, AUARC) can be misleading due to this entanglement. Taking AUARC as an example, if the base LM is powerful and all correctness values of its responses are high (*e.g.*, within $[0.9, 1.0]$), then the evaluated AUARC will be high for any uncertainty/confidence measure, regardless of its quality. This is undesirable because our goal is to provide an overall assessment of the uncertainty measure, which may in the future need to be applied to different LMs. While the ECE metric provides a limited "disentangling" effect, in the sense that it can reflect that highly accurate models may be poorly calibrated (*i.e.*, with high ECE values) (Guo et al., 2017), it is not applicable to uncertainty measures in general.

Desiderata of evaluation. The aforementioned challenges suggest that the evaluation of LM uncertainty measures should take into account the following key desiderata: (1) avoidance of ad hoc correctness thresholding, (2) applicability to diverse output ranges of uncertainty measures, and (3) decoupling from the generative performance of the LM. Moreover, the evaluation framework should be practical. We view these criteria as important, but *not necessarily exhaustive* for an ideal assessment. Future research may identify other requisites and further improve our framework accordingly.

4 Rank-Calibration

In this section, we introduce a novel assessment framework satisfying the criteria outlined in Sec. 3.

4.1 Rank-Calibration & RCE

Define the regression function $\mathrm{reg}(\cdot): \mathbb{R} \to \mathbb{R}, u \mapsto \mathbb{E}_{\boldsymbol{x}, \widehat{\boldsymbol{y}}}[A(\boldsymbol{x}; \widehat{\boldsymbol{y}}) \mid U(\boldsymbol{x}; \widehat{\boldsymbol{y}}) = u]$, representing the *expected correctness level A conditional on an uncertainty level $U = u$*. Here, $\boldsymbol{x}$ is a random query sampled from the distribution associated with a specific benchmark dataset, while $\widehat{\boldsymbol{y}} \mid \boldsymbol{x} \sim \mathbb{P}(\cdot \mid \boldsymbol{x})$ is a random output sampled from the generative distribution of the LM. We start from the observation that, ideally, a lower uncertainty level should correspond to higher generation accuracy. This is equivalent to saying that the regression function should ideally be *monotone decreasing*.

Since U is a random variable depending on $(\boldsymbol{x}; \widehat{\boldsymbol{y}})$, $\mathrm{reg}(U)$ is also random. If $\mathrm{reg}(\cdot)$ is monotone decreasing, then $U \leq u$ implies $\mathrm{reg}(U) \geq \mathrm{reg}(u)$. Thus, for any value u in the range of U,

$$\mathbb{P}(U \leq u) = \mathbb{P}(\mathrm{reg}(U) \geq \mathrm{reg}(u)). \tag{1}$$

Equation (1) suggests a direct relation between an uncertainty level u and its corresponding expected correctness level $\mathrm{reg}(u)$. For example, for a value of u in the in bottom 10% of the distribution of U, the expected correctness level $\mathrm{reg}(u) = \mathbb{E}[A \mid U = u]$ is in the top 10% in the distribution of $\mathrm{reg}(U) = \mathbb{E}[A \mid U]$. We call this desired property of uncertainty measures *Rank-Calibration*.

Definition 1 (Rank-Calibration)**.** *We say that an uncertainty measure U is rank-calibrated if* (1) *holds for any u in U's range: on average, lower uncertainty implies higher generative quality.*

Rank-calibration is related to, yet distinct from, the usual notion of calibration in the classification context (Lichtenstein et al., 1977; Guo et al., 2017). We defer the detailed discussion to Sec. 4.2.

To quantify the distance of a given uncertainty measure from the ideal rank-calibration, we propose the following Rank-Calibration Error (RCE), inspired by ECE for calibration.

Definition 2 (Rank-Calibration Error)**.** *The RCE of an uncertainty measure U is defined as*

$$\mathbb{E}_U\left[\left|\mathbb{P}_{U'}(\mathrm{reg}(U') \geq \mathrm{reg}(U)) - \mathbb{P}_{U'}(U' \leq U)\right|\right], \tag{RCE}$$

where U' is an independent copy of U.

Extension to confidence measures. While primarily motivated by uncertainty measures with incommensurate ranges, rank-calibration also applies to confidence measures. Ideally, *higher values of a confidence measure should imply higher generation accuracy*. Thus, defining $\overline{\mathrm{reg}}(c) := \mathbb{E}[A \mid C = c]$ for all c in the range of C, we can adapt RCE to

$$\mathbb{E}_C\left[\left|\mathbb{P}_{C'}(\overline{\mathrm{reg}}(C') \geq \overline{\mathrm{reg}}(C)) - \mathbb{P}_{C'}(C' \geq C)\right|\right], \tag{2}$$

where C' is an independent copy of C. This gauges deviations from the equivalence between $C \geq c$ and $\overline{\mathrm{reg}}(C) \geq \overline{\mathrm{reg}}(c)$. Since rank-calibration provides a different perspective from calibration—see Sec. 4.2—(2) serves as a supplement to ECE in assessing confidence measures.

4.2 Comparison with Classical Calibration

For a binary correctness value function A taking values in $\{0, 1\}$, rank-calibration relaxes classical calibration by absorbing all *strictly decreasing* transformations.

Theorem 1. *Suppose the correctness function A takes values in $\{0, 1\}$. If an uncertainty measure U is rank-calibrated,* i.e.*, its RCE is zero, then*

there exists a unique strictly decreasing transformation $g^\star\colon \mathbb{R} \to [0,1]$ *such that* $C_{g^\star} := g^\star(U)$ *is calibrated,* i.e.*, its ECE is zero. If a confidence measure* C *is calibrated, then for any strictly decreasing transformation* $h\colon \mathbb{R} \to \mathbb{R}$*, the induced uncertainty measure* $U_h := h(C)$ *is rank-calibrated.*

Proof. If U is rank-calibrated, the regression function $u \mapsto \mathrm{reg}(u) = \mathbb{E}[A \mid U = u] \in [0,1]$ is strictly decreasing over all values in U's range with positive density (or mass). Moreover, $\mathbb{P}(A = 1 \mid \mathrm{reg}(U) = \mathrm{reg}(u)) = \mathbb{E}[A \mid U = u] = \mathrm{reg}(u)$. Therefore, $\mathrm{reg}(U)$ is a calibrated confidence measure, and reg is strictly decreasing. The uniqueness follows as $\mathbb{P}(A = 1 \mid g(U)) = \mathbb{E}[A \mid U] = \mathrm{reg}(U)$ for any strictly monotone function.

On the other hand, if C is calibrated, then $C = \mathbb{P}(A = 1 \mid C) = \mathbb{E}[A \mid C]$ almost surely. For any strictly decreasing h, we have $\mathbb{E}[A \mid U_h] = \mathbb{E}[A \mid C] = C$ almost surely because h is a one-to-one map. Therefore, for any given c and uncertainty value $u_h = h(c)$, it holds almost surely that

$$\begin{aligned} & U_h = h(C) \le u_h = h(c) \iff C \ge c \\ \iff & \mathbb{E}[A \mid C] \ge \mathbb{E}[A \mid C = c] \\ \iff & \mathbb{E}[A \mid U_h] \ge \mathbb{E}[A \mid U_h = u_h], \end{aligned}$$

which implies U_h is rank-calibrated. □

Theorem 1 implies that, for a binary correctness function, one can construct a calibrated confidence measure from an uncertainty measure with monotone transformations if and only if the uncertainty measure is rank-calibrated. However, RCE and ECE gauge different quantities: ECE captures the absolute difference between predicted and true probabilities, while RCE reflects the deviation from a monotonic correspondence between uncertainty and the expected correctness. These two notions are generally not directly comparable.

For example, consider the special case where a continuous-valued confidence measure C is completely uninformative and the regressed correctness $\overline{\mathrm{reg}}\colon c \mapsto \mathbb{E}[A \mid C = c]$ is a constant for all confidence level c. Then, the RCE defined in (2) reports a large value of $1/2$, reflecting its poor indicativeness. However, the ECE can be large or small depending on the averaged distance between C's output and $\overline{\mathrm{reg}}$. More generally, we find no relation in the results of ECE and RCE through the following result, proved in Appendix D.

Proposition 1. *Let the correctness function* $A \in \{0,1\}$ *be binary. For any* $\alpha, \beta \in (0, 1/2]$*, there is a confidence measure* C *such that its RCE is* α *while the ECE is* β*.*

4.3 Empirical RCE & Indication Diagram

Now, as in Sec. 2, consider a dataset $\{(u_i, a_i)\}_{i=1}^n$ of uncertainty and correctness values computed over a benchmark dataset where each $u_i = U(\boldsymbol{x}; \widehat{\boldsymbol{y}}_i)$, $a_i = A(\boldsymbol{x}_i; \widehat{\boldsymbol{y}}_i)$, and $\widehat{\boldsymbol{y}}_i$ is a response generated by the LM. The true value of RCE is unknown, as it refers to an average over the distribution from which the data are drawn.

Empirical RCE. The RCE involves the unknown probabilities $\mathbb{P}(U \le u)$ and $\mathbb{P}(\mathrm{reg}(U) \ge \mathrm{reg}(u))$, which generally need to be estimated. Estimating the latter is challenging as the regression function is also unknown and needs to be estimated.

To address this, we adopt a piecewise constant regression or binning strategy, as in non-parametric statistics (Tsybakov, 2009). First, we group the uncertainty values $\{u_i\}_{i=1}^n$ into B equal-mass intervals, each containing $\lceil n/B \rceil$—or, when needed, $\lfloor n/B \rfloor$—elements. The boundaries of the b-th ($1 \le b \le B$) bin are the $(b-1)/B$-th and b/B-th quantiles of $(u_i)_{i=1}^n$. Let $\mathcal{I}_b \subseteq \{1, \ldots, n\}$ be the set of indices of the datapoints whose uncertainty values fall into the b-th bin. Then, the expected correctness level over the b-th bin can estimated as

$$\mathrm{crc}_b := \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} a_i,$$

when $|\mathcal{I}_b| > 0$. From now on, we will interpret $0/0 := 0$; and we extend to $|\mathcal{I}_b| = 0$ in this way. Clearly, crc_b is an unbiased estimator of $\mathbb{E}[A \mid U \in \text{ the } i\text{-th bin}]$, which approximates $\mathrm{reg}(U)$ accurately given a narrow bin and abundant data. We similarly estimate the average uncertainty within the b-th bin as

$$\mathrm{uct}_b = \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} u_i.$$

As crc_b and uct_b estimate the per-bin averages of $\mathrm{reg}(U)$ and U, for each b. We estimate $\mathbb{P}(U \le u_i)$ and $\mathbb{P}(\mathrm{reg}(U) \ge \mathrm{reg}(u_i))$ for $i \in \mathcal{I}_b$ as follows:

$$\widehat{\mathbb{P}}(\mathrm{reg}(U) \ge \mathrm{reg}(u_i)) := \frac{1}{B-1} \sum_{b' \ne b} \mathbb{1}[\mathrm{crc}_b \ge \mathrm{crc}_{b'}],$$

$$\widehat{\mathbb{P}}(U \le u_i) := \frac{1}{B-1} \sum_{b' \ne b} \mathbb{1}[\mathrm{uct}_b \le \mathrm{uct}_{b'}].$$

A rank-calibrated measure has $\widehat{\mathbb{P}}(U \le u_i) \approx \widehat{\mathbb{P}}(\mathrm{reg}(U) \ge \mathrm{reg}(u_i))$ for all $1 \le i \le n$. We thus compute the empirical Rank-Calibration Error estimator (Empirical RCE) by taking an average of the

per-bin rank differences of correctness and uncertainty values. More precisely,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{\mathbb{P}}(\mathrm{reg}(U)\geq \mathrm{reg}(u_i))-\widehat{\mathbb{P}}(U\leq u_i)\right|. \quad \text{(Empirical RCE)}$$

The difference between the estimated probabilities for a given bin represent the ranking gap (*i.e.*, blue and shallow red areas in Fig. 1). We use the Empirical RCE as the main metric to assess uncertainty and confidence measures in the paper.

Indication diagram. Similar to reliability diagrams representing miscalibration (Lichtenstein et al., 1977; Niculescu-Mizil and Caruana, 2005), we can also visualize rank-miscalibration in diagrams (*e.g.*, Fig. 1). In particular, we plot the relative percentile (between 0% and 100%) of the expected correctness level (*i.e.*, $\mathrm{reg}(U)$) as a function of the relative percentile of uncertainty (*i.e.*, U). We term these plots *indication diagrams*. If a measure is rank-calibrated—*i.e.*, if (1) holds—then the indication diagram should lie on the anti-diagonal line $\mathrm{percent}(\mathrm{reg}(u)) = 1-\mathrm{percent}(u)$. Deviations from this line represent rank-miscalibration.

Advantages of rank-calibration. We summarize the advantages of the rank-calibration framework by revising the desiderata from Sec. 3. First, the empirical RCE does not require any thresholding of the correctness values. Second, rank-calibration assesses the monotonicity of uncertainty values by leveraging relative ranks, which makes it independent of the output range. Third, similar to ECE, the RCE is not directly tied to the generation performance of the LM. Finally, our assessment is practical for any uncertainty/confidence measures.

5 Experiments

We provide more comprehensive experiments and justify the advantages of our assessment.

5.1 Experiment Setup

We consider both open-source and commercial LMs, including *Llama-2-7b*, *Llama-2-7b-chat (Touvron et al., 2023b)* (an instruction fine-tuned version of *Llama-2-7b*), and *GPT-3.5-turbo* (Ouyang et al., 2022). See Appendix E.1 for more details.

We conduct assessments on the validation sets of four datasets: TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), SQuAD-1 (Rajpurkar et al., 2016), and Meadow (Wang et al., 2020). For assessment over the open-ended and challenging Meadow, we only use the more advanced model GPT-3.5-turbo. To account for randomness in the evaluation, we repeat experiments bootstrapping each dataset 20 times. See more details of datasets in Appendix E.2.

We use multiple correctness functions, including the *Rouge-L* score, *BERT similarity*, and *ChatGPT evaluation*, all widely applied before (Kuhn et al., 2023; Xiong et al., 2024). ChatGPT correctness is only used for GPT-3.5-turbo with temperature 1.0. See Appendix E.3 for more details.

The uncertainty/confidence measures to be assessed are the same as in Sec. 3, (*i.e.*, U_{NLL}, U_{SE}, U_{Ecc}, U_{Deg}, U_{EigV}, and C_{Verb}). We first illustrate that our proposed assessment has broad applicability and granular interpretability. Furthermore, we qualitatively show that uncertainty measures with lower RCE values reliably indicate correctness. Finally, we study robustness by empirically checking the impact of temperature and correctness functions on RCE (Demšar, 2006). More results for different configurations are in Table 2.

5.2 Broader Applicability

Previous assessments have some limitations in open-ended tasks. First, as shown in Fig. 4 [top], the correctness distribution in open-ended tasks (*e.g.*, the Meadow dataset) is less concentrated around zero and one compared to the TriviaQA correctness distribution. Consequently, if correctness were binarized with thresholding, the assessed results would be highly impacted by the threshold choice, as illustrated in Fig. 4 [bottom]. As such, using continuous-valued correctness scores is common in open-ended tasks (Cohan et al., 2018; Uppalapati et al., 2023). Since RCE does not require thresholding, our rank-calibration assessment does not suffer from the above issue.

5.3 Granular Interpretability

Beyond the rank-calibration error, the indication diagrams can be instrumental in understanding the performance of uncertainty measures. We show the indication diagrams of U_{NLL} and U_{SE} for GPT-3.5-turbo on TriviaQA in Fig. 1. More indication diagrams can be found in the Appendix.

First, indication diagrams consistently reflect the effect of rank-miscalibration. The indication diagram of U_{NLL} (Fig. 1 [left]) has more overlap between the red and blue bars, compared to that of U_{Ecc} (Fig. 1 [right]), reflecting a lower RCE level (0.038 with U_{NLL} v.s. 0.151 with U_{Ecc}). The high overlap suggests that the relative ranks of uncer-

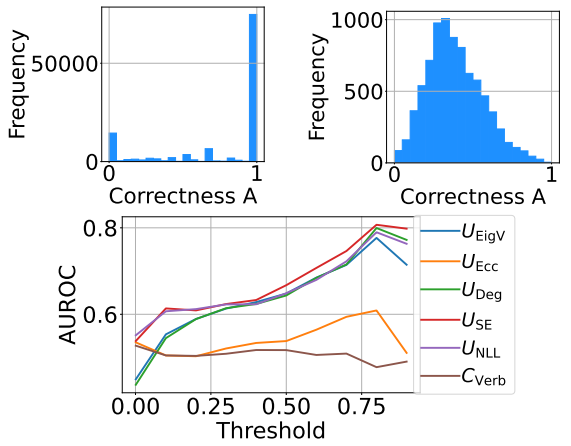


Figure 4: Top: Rouge-L correctness distributions of GPT-3.5-turbo on the *TriviaQA (left)* and *Meadow (right)* benchmarks. Bottom: AUROCs of assessed measures for GPT-3.5-turbo on *Meadow*, with Rouge-L correctness and various thresholds.

tainty values are more aligned with those of correctness levels, leading to better rank-calibration.

Second, indication diagrams can shed light onto which uncertainty levels may be problematic. For example, in Fig. 1 [right], we observe that for an uncertainty in the top 75th percentile, U_{Ecc} tends to be overpessimistic: U_{Ecc} assigns high uncertainty values to high-quality generations.

5.4 Qualitative Illustration

$\boldsymbol{x}$: On September 28th, NASA announced that what had been detected on Mars?
$\boldsymbol{y}$: flowing water
$\widehat{\boldsymbol{y}}$: Possible signs of life
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.813
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.930

$\boldsymbol{x}$: "Feel Like Making Love" and "The First Time Ever I Saw Your Face" were hit singles for which female artist?
$\boldsymbol{y}$: roberta flack
$\widehat{\boldsymbol{y}}$: Roberta Flack
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.864
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.046

To illustrate the effectiveness of the RCE as an evaluation metric for uncertainty measures, we present two TriviaQA instances and contrast U_{NLL} (having RCE 0.037) with U_{SE} (having RCE 0.051) for GPT-3.5. Here, $\boldsymbol{x}$ is the question input, $\boldsymbol{y}$ is the answer in the dataset, $\widehat{\boldsymbol{y}}$ is the LM response, and $\mathbb{P}(U \leq u)$ signifies the relative magnitudes of LM's uncertainty level according to U_{NLL} and U_{SE}.

In the first instance, the generation is factually incorrect and U_{NLL} assigns a high uncertainty value to the response, *i.e.* $\mathbb{P}(U_{\text{NLL}} \leq u) \approx 1$. In the second scenario, where the generation is correct, U_{NLL} succeeds in providing a lower uncertainty level, *i.e.* $\mathbb{P}(U_{\text{NLL}} \leq u) \approx 0$. Yet, U_{SE} assigns *a lower uncertainty to a poorer generation and a higher uncertainty to a better generation*! These instances showcase that U_{NLL} is more reliable than U_{SE} here, which is consistent with the RCE-assessed results. Additional qualitative results are given in Table 3.

5.5 Post-hoc Recalibration

Recalibrating uncertainty/confidence measures with poor rank-calibration can be of interest; for ECE, this is sometimes known as Mincer-Zamowitz regression (Mincer and Zarnowitz, 1969). As discussed in Sec. 4.2, an ECE-calibrated measure is also RCE-calibrated. However, RCE is invariant to monotone transformations, which means that approaches like Platt scaling (Platt, 1999) and isotonic regression (Zadrozny and Elkan, 2002) will not improve rank-calibration. Therefore, we suggest using histogram binning (or, piecewise constant regression), which includes non-monotone transforms (Zadrozny and Elkan, 2001). See the related experimental results in Appendix F.2.

5.6 Robustness Analysis

We conduct ablation studies to analyze the robustness of our assessment to key hyperparameters, including temperatures and correctness scores. We further propose a method to make robust comparisons between uncertainty measures via the *Critical Difference (CD) Diagram* (Demšar, 2006). Detailed information and results are in Appendix F.4.

6 Conclusion

This paper investigates the limitations of common assessments for LM uncertainty/confidence measures. We develop an alternate framework, termed rank-calibration, to assess their quality. Our approach does not require binarizing correctness at ad hoc thresholds and is compatible with uncertainty measures taking values in any output range. We experimentally show the broad applicability and the granular interpretability of our method, and provide a comprehensive robustness analysis. Future directions include developing uncertainty measures with guaranteed rank-calibration and enhancing generative pipelines of LMs (*e.g.*, the retrieval-augmented generation) with rank-calibrated measures.

Limitation & Broader Impact

The empirical RCE estimate has not been subjected to a thorough statistical analysis. The performance of assessed uncertainty and confidence measures (*e.g.*, the vanilla verbalized confidence C_{Verb}) have not been optimized, since the paper focuses on a new assessment approach rather than benchmarking. Human correctness evaluation is not performed, due to our limited budget.

This work is designed to unveil the issues in the existing approaches for evaluating LM uncertainty/confidence measures, and to introduce an alternate, principled assessment to the LM community. We believe there are no ethical concerns associated with our research.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

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

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2021. Calibration of neural networks using splines. In *International Conference on Learning Representations*.

Frank E Harrell. 2015. Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Siddhartha Jain, Ge Liu, Jonas Mueller, and David Gifford. 2020. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4264–4271.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

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

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

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

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

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. 2023. T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. 1977. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer.

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

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

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

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

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Meta. 2023. Llama access request form - meta ai. https://ai.meta.com/resources/models-and-libraries/llama-downloads/. (Accessed on 12/13/2023).

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Jacob A Mincer and Victor Zarnowitz. 1969. The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pages 625–632.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Georgios Papadopoulos, Peter J Edwards, and Alan F Murray. 2001. Confidence estimation methods for neural networks: A practical comparison. *IEEE transactions on neural networks*, 12(6):1278–1287.

Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*.

Leonard J Savage. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. *arXiv preprint arXiv:2205.12507*.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying uncertainty in natural language explanations of large language models. *arXiv preprint arXiv:2311.03533*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Alexandre B Tsybakov. 2009. *Introduction to Nonparametric Estimation*. Springer.

Padma Jyothi Uppalapati, Madhavi Dabbiru, · K. Venkata Rao, Omer F. Rana, Rajiv Misra, Alexander Pfeiffer, Luigi Troiano, Nishtha Kesswani, and K. Venkata Rao. 2023. A comprehensive survey on summarization techniques. *SN Computer Science*, 4:1–9.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Robert L Winkler, Javier Munoz, José L Cervera, José M Bernardo, Gail Blattenberger, Joseph B Kadane, Dennis V Lindley, Allan H Murphy, Robert M Oliver, and David Ríos-Insua. 1996. Scoring rules and the evaluation of probabilities. *Test*, 5:1–60.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

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

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Intertional*

Conference on Machine Learning, volume 1, pages 609–616.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Additional Related Work

Uncertainty measures in supervised learning. The quantification of uncertainties in model outputs in supervised learning has a long history (*e.g.*, Lichtenstein et al., 1977, etc). Overparametrized models such as neural networks pose unique challenges to estimating uncertainty and enhancing model calibration(Guo et al., 2017; Papadopoulos et al., 2001; Riquelme et al., 2018). Various approaches have been introduced to mimic Bayesian inference (Gal and Ghahramani, 2016), to utilize simple deep ensembles (Lakshminarayanan et al., 2017; Jain et al., 2020), and to identify training samples that are out-of-distribution (Liang et al., 2018; Papernot and McDaniel, 2018). Nonetheless, it is not clear how to adapt these strategies to language modeling, where the outputs can be text with complex structure.

Uncertainty measures in language modeling. To gauge the uncertainty level associated with the outputs of LMs, Kuhn et al. (2023) introduce the concept of semantic entropy, which integrates linguistic consistencies stemming from shared meanings. In a similar vein, Kadavath et al. (2022); Lin et al. (2022); Xiong et al. (2024) encourage LMs to analyze their own responses and come up with a "probability" that a response is correct. In related work, Manakul et al. (2023) uses sampling to identify instances of fabricated information. Recently, Tian et al. (2023) explore methods for deriving confidence measures for reinforcement-learning-trained LMs. Lin et al. (2023) draw a distinction between estimating uncertainty and confidence for LMs. Similarly, Chen and Mueller (2023) introduce a method for detecting bad and speculative responses from a pre-trained LM with a confidence score. Tanneru et al. (2023) propose two novel measures to quantify the uncertainty of LM-generated explanations. Although considerable research focuses on developing uncertainty and confidence measures for LMs, the evaluation of their effectiveness is less studied.

Assessments of uncertainty measures. Early assessment of confidence measures in classification scenarios leveraged proper scoring rules (Savage, 1971; DeGroot and Fienberg, 1983; Gneiting and Raftery, 2007), such as the Brier score (Brier, 1950) and the KL divergence (Winkler et al., 1996). Other assessments include plotting calibration curves, also known as reliability diagrams (estimated probabilities against predicted ones) (Harrell, 2015). More recently, the ECE metric—or mean absolute calibration error—has gained popularity in machine learning (Harrell, 2015; Naeini et al., 2015), along with many variants (Kumar et al., 2019; Nixon et al., 2019; Gupta et al., 2021; Lee et al., 2023; Si et al., 2022).

In the realm of uncertainty quantification for LMs, the assessment based on ECE remains viable. However, it necessitates the introduction of ad hoc threshold to derive binary labels. Moreover, the applicability of ECE is limited, as it does not directly apply to LM uncertainty measures that fall outside the interval $[0, 1]$. Our work introduces an assessment centered around rank-calibration, a critical property that ideal uncertainty measures should satisfy. This assessment is applicable to both confidence and uncertainty measures and eliminates the need for thresholding the correctness values.

B Common Uncertainty/Confidence Measures for LMs

In this section, we introduce common measures of uncertainty and confidence in detail.

- **NLL & Perplexity.** Let $\widehat{\boldsymbol{y}} = (\widehat{y}_\ell)_{\ell \geq 1}$ be the generated response. Then the Negative Log-Likelihood (NLL) is

$$U_{\text{NLL}}(\boldsymbol{x}, \widehat{\boldsymbol{y}}) := -\ln(\mathbb{P}(\widehat{\boldsymbol{y}} \mid \boldsymbol{x})) = -\sum_{\ell \geq 1} \ln(\mathbb{P}(\widehat{y}_\ell \mid \boldsymbol{x}, \widehat{y}_{<\ell})).$$

  A natural extension accounts for the variable length of responses by applying length normalization. Supposing the number of tokens of the response $\widehat{\boldsymbol{y}}$ is $\text{len}(\widehat{\boldsymbol{y}})$, the length-normalized NLL is defined as

$$U_{\text{NLL-LN}}(\boldsymbol{x}, \widehat{\boldsymbol{y}}) := -\frac{1}{\text{len}(\widehat{\boldsymbol{y}})} \sum_{\ell=1}^{\text{len}(\widehat{\boldsymbol{y}})} \ln(\mathbb{P}(\widehat{y}_\ell \mid \boldsymbol{x}, \widehat{y}_{<\ell})).$$

  Roughly speaking, this can be viewed as the average nats per token in the generated text; if using $\log_2$ instead of $\ln$, it would be the average bits per token. The exponential of the length-normalized NLL is

known as the Perplexity: $U_{\text{Perp}}(\boldsymbol{x}; \widehat{\boldsymbol{y}}) := \exp(U_{\text{NLL-LN}}(\boldsymbol{x}, \widehat{\boldsymbol{y}}))$ (Jelinek et al., 1977). The perplexity can also be viewed as the inverse of the geometric mean of the token-wise probabilities.

- **Entropy.** Entropy is a well-known type of uncertainty measure. The predictive entropy of the distribution $\mathbb{P}(\cdot \mid \boldsymbol{x})$ is defined as

$$U_{\text{E}}(\boldsymbol{x}) := -\mathbb{E}_{\widehat{\boldsymbol{y}} \sim \mathbb{P}(\cdot|\boldsymbol{x})}[\ln(\mathbb{P}(\widehat{\boldsymbol{y}} \mid \boldsymbol{x}))].$$

  Entropy gauges the information one has about the potential output given the input, and has high values when outputs are diverse. Malinin and Gales (2021) propose a variant $U_{\text{E-LN}}(\boldsymbol{x})$ using the length-normalized log-likelihood $\ln(\mathbb{P}(\widehat{\boldsymbol{y}} \mid \boldsymbol{x}))/\text{Length}(\widehat{\boldsymbol{y}})$. Kuhn et al. (2023) argues that responses with an identical meaning should be viewed as equal; even if they differ at the token level. They thus propose the semantic entropy

$$U_{\text{SE}}(\boldsymbol{x}) := -\mathbb{E}_{\widehat{\boldsymbol{y}} \sim \mathbb{P}(\cdot|\boldsymbol{x})}[\ln(\mathbb{P}(c(\widehat{\boldsymbol{y}}) \mid \boldsymbol{x}))],$$

  where $c(\widehat{\boldsymbol{y}})$ is a semantic concept of output $\widehat{\boldsymbol{y}}$, as determined by another machine learning method. We can similarly define the length-normalized semantic entropy as

$$U_{\text{SE-LN}}(\boldsymbol{x}) := \mathbb{E}_{\widehat{\boldsymbol{y}} \sim \mathbb{P}(\cdot|\boldsymbol{x})}[\ln(\mathbb{P}(c(\widehat{\boldsymbol{y}}) \mid \boldsymbol{x}))/\text{len}(\widehat{\boldsymbol{y}})].$$

- **Affinity graph.** Recently, Lin et al. (2023) use a weighted adjacency graph built upon semantic affinities between outputs to reflect uncertainty. Given an *entailment-contradiction* affinity model e that maps pairs $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{y}}'$ of responses to values in $[0, 1]$, e induces a symmetric adjacency matrix $W = [w_{i,j}]_{i,j=1}^K$ with responses $\{\widehat{\boldsymbol{y}}^{(k)}\}_{k=1}^K$ sampled from $\mathbb{P}(\cdot \mid \boldsymbol{x})$, where for all i, j, $w_{i,j} = (e(\widehat{\boldsymbol{y}}^{(i)}; \widehat{\boldsymbol{y}}^{(j)}) + e(\widehat{\boldsymbol{y}}^{(j)}; \widehat{\boldsymbol{y}}^{(i)}))/2$. Let $D = [\mathbb{1}[j = i] \sum_{k=1}^K w_{k,j}]_{i,j=1}^K$ be the matrix of degrees and $\{\lambda_k\}_{k=1}^K$ be the eigenvalues of the *Laplacian* $L = I - D^{-1/2} W D^{-1/2}$. Measures proposed in Lin et al. (2023) include

$$U_{\text{EigV}}(\boldsymbol{x}) := \sum_{k=1}^{K} \max\{0, 1 - \lambda_k\},$$

$$U_{\text{Deg}}(\boldsymbol{x}) := 1 - \text{trace}(D)/K^2, \qquad C_{\text{Deg}}(\boldsymbol{x}; \widehat{\boldsymbol{y}}^{(i)}) := D_{i,i}/K,$$

$$U_{\text{Ecc}}(\boldsymbol{x}) := \|[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K]\|_2.$$

  where $\{\boldsymbol{v}_k\}_{k=1}^K$ are certain centralized vectors associated with the spectral decomposition of L. Here, $U_{\text{EigV}}(\boldsymbol{x})$ is approximates the number of connected components in the graph represented by W, while $U_{\text{Deg}}(\boldsymbol{x})$ and $U_{\text{Ecc}}(\boldsymbol{x})$ reflect the diversity of outputs.

- **Verbalized confidence.** Verbalized confidence generally refers to the textual confidence output by an LM. For example, if an LM is highly uncertain about its answer, it may inform the user by saying, *e.g.*, "I am only 20% confident in this answer." This is often implemented by feeding handcrafted prompts to advanced LMs such as GPT-4 (OpenAI, 2023). Many prompting strategies have been used in the literature to enhance this procedure (Zhao et al., 2021; Kadavath et al., 2022; Lin et al., 2022; Xiong et al., 2024). Since optimizing the prompting strategy is not our focus and we do not want confidence elicitation to interfere with the generation of responses, we adopt a simple post-hoc strategy here by feeding a query-response pair to an LM and asking it how confident it believes the response correctly addresses the query. This post-hoc strategy is similar to the one used by Kadavath et al. (2022). We use the following specific prompt format:

```
Read the question and answer below.
{question} {generation}
Provide a numeric confidence that indicates your certainty about
this answer.
For instance, if your confidence level is 80%, it means you are 80%
```

```
certain that this answer is correct and there is a 20% chance that
it is incorrect.
Use the following format to provide your confidence: Confidence:
[Your confidence, a numerical number in the range of 0-100]%."
```

C Common Evaluation Metrics

In this section, we review evaluation metrics that have been commonly used to assess LM uncertainty/confidence measures. These metrics usually require binary correctness values.

- **AUROC.** AUROC refers to the area under the Receiver-Operating Curve (ROC). The ROC plots the true positive rate (a.k.a. recall) against the false positive rate (a.k.a. $1-$ specificity) at various thresholds of uncertainty levels. The true positive rate is on the y-axis, and the false positive rate is on the x-axis. An AUROC value of 1 may represent a perfect uncertainty measure; a value of 0.5 suggests no discriminative ability (equivalent to random uncertainty levels). The AUROC can be more useful for evaluation in imbalanced scenarios where correct responses are much more (or less) frequent than incorrect responses.

- **AUPRC.** AUPRC refers to the area under the Precision-Recall Curve (PRC), which plots the positive predictive value (a.k.a. precision) against the true positive rate (a.k.a. recall) at various threshold settings. Precision is on the y-axis, and recall is on the x-axis. Similar to AUROC, it is valuable in imbalanced dataset scenarios but focuses more on the performance of the positive (minority) class (*i.e.*, correct responses). Variants of AURPC include AUPRC-Positive and AUPRC-Negative, which focus on gauging the ability of uncertainty measures to identify correct responses and incorrect responses, respectively.

- **AUARC.** AUARC refers to the area under the Accuracy-Rejection Curve (ARC) that plots the accuracy of generation against a rejection rate (the proportion of generated responses for which the model abstains from making a prediction). The curve shows how the accuracy of generation improves as it is allowed to reject uncertain responses. A higher AUARC value means that an LM can generate more correct responses as it increasingly avoids uncertain (based on the level of specific uncertainty measures) cases. This metric is useful for evaluating uncertainty measures in scenarios where LMs can defer responses for which they are not confident.

- **ECE.** ECE stands for the expected calibration error, a metric used to evaluate the calibration of confidence measures, particularly in classification tasks. Calibration refers to how well the confidence levels align with the actual proportion of correct generation. For an ideally calibrated confidence measure, if the confidence level is 70%, then approximately 70% of generated responses should be correct. ECE quantifies the difference between the confidence levels and the realized correct proportion. A lower ECE indicates better calibration, meaning the confidence measure is more reflective of the actual correct proportion. A confidence measure with an ECE close to zero is considered well-calibrated.

D Proof of Proposition 1

Case 1. $\alpha = 1/2$**.** Consider the continuous case $C \sim \mathrm{Unif}[1/2-\beta, 1/2+\beta]$ and $\mathrm{reg}(C) \equiv 1/2+\beta$ almost surely (*i.e.*, $A \sim \mathrm{Bernoulli}(1/2+\beta)$). Then $\mathbb{P}_{C'}(\overline{\mathrm{reg}}(C') \geq \overline{\mathrm{reg}}(C)) \equiv 1$ for almost surely. Since C is continuous-valued, $\mathbb{P}_{C'}$ follows the uniform distribution over $[0,1]$. We thus have

$$\mathrm{RCE} = \int_0^1 |1-p| \mathrm{d}p = \frac{1}{2}.$$

On the other hand,

$$\mathrm{ECE} = \int_{1/2-\beta}^{1/2+\beta} \frac{|1/2+\beta-c|}{2\beta} \mathrm{d}c = \beta.$$

Case 2. $\alpha \in (0, 1/2)$**.** Consider the case $\mathrm{reg}(C) \equiv 1/2 + \beta$ almost surely. We construct the marginal distribution of C as follows. Let $\mathbb{P}(C = c_k) = p_k$ for $1 \le k \le K$ with $K \ge (1 - 2\alpha)^{-1}$. Here $p_1 = \cdots = p_{K-1} = p$ while $p_K = 1 - (K-1)p$ where p is the non-negative root of $(K-1)p^2 + (1 - (K-1)p)^2 = 1 - 2\alpha$. Since $K \ge (1-2\alpha)^{-1}$, such $p \in (0, (K-1)^{-1}]$ exists. Moreover, we let $\{c_k\}_{k=1}^K$ satisfy $0 \le c_1 < \cdots < c_{K-1} \le 1/2 + \beta$, $c_k + c_{K-k} \equiv 1$ with $c_k \ne 1/2$ for all $1 \le k < K$, $c_K = 1/2$. Then, by definition, we can calculate

$$\begin{aligned}\mathrm{RCE} &= \sum_{k=1}^K p_k \left(1 - \sum_{\ell \ge k} p_\ell\right) = \sum_{1 \le \ell < k \le K} p_k p_\ell \\ &= \frac{\left(\sum_{k=1}^K p_k\right)^2 - \sum_{k=1}^K p_k^2}{2} = \frac{1 - \sum_{k=1}^K p_k^2}{2} = \alpha.\end{aligned}$$

On the other hand, we have

$$\mathrm{ECE} = \sum_{k=1}^K \left|\frac{1}{2} + \beta - c_k\right| p_k = \beta + \frac{1}{2} - \sum_{k=1}^K c_k p_k = \beta.$$

This finishes the proof.

E Additional Experiment Details

E.1 Model Setup

Following Lin et al. (2023), we set the temperature to 0.6 for the two Llama-2 models and 1.0 for the GPT model. We quantize the two Llama-2 models to 16 bits. To ablate the influence of temperature, we also use generated responses of Llama-2-7b-chat with temperature 1.0.

E.2 Datasets

Dataset Descriptions. TriviaQA is a challenging reading comprehension dataset, containing question-answer pairs whose answers can be found on Wikipedia and the web. Similar to previous works, we use TriviaQA as an open-domain QA benchmark. Natural Questions is a question-answering dataset containing questions issued to the Google search engine. We use Natural Questions as an open-domain QA benchmark. SQuAD-1 is a reading comprehension dataset containing questions posed by crowdworkers based on Wikipedia articles. We include SQuAD-1 as a reading comprehension benchmark, where the annotated contexts are provided in the prompt. Meadow is created by research groups working on COVID-19 problems. We use this dataset for open-ended generation, where the LM is expected to provide a title for a paper given the abstract of the paper. The correctness is justified by comparing the generated title to the true title.

Dataset Setup. TriviaQA contains 11,322 data points, Natural Questions contains 3,600 data points, SQuAD-1 contains 10,570 data points, and Meadow contains 1,000 data points. The prompt templates used are similar to those in Kuhn et al. (2023); Lin et al. (2023), and are as follows:
TriviaQA: following from Lin et al. (2023), we use the exact same prompt used in Touvron et al. (2023a):

```
Answer these questions:
In Scotland, a bothy/bothie is a?
A: House
{question}
A:
```

Natural Question: Similar to Lin et al. (2023), we use an in-context learning prompt with five demonstrations:

```
where are the fa cup semi finals played.  [SEP] A: the new Wembley
Stadium.[SEP]
who was alf married to in home and away [SEP] A: Ailsa Stewart.[SEP]
```

what is the name of the first book in the twilight series [SEP] A: Twilight.[SEP]
when is tornado season in the united states [SEP] A: March through June.[SEP]
where did the idea of a messiah come from [SEP] A: Judaism.[SEP]
question [SEP] A:

SQuAD-1: Each data point in SQuAD-1 is a (question, context, reference) triplet, where the context is annotated to provide useful information to answer the question. We prompt SQuAD-1 using zero-shot prompting:

Answer the following question based on the context.
{question}
Context: {context}
A:

Meadow: Each data point in Meadow is a (abstract, title) pair. We prompt Meadow using one-shot prompting:

Abstract: Coronavirus disease 2019 (COVID-19) threatens vulnerable patient populations, resulting in immense pressures at the local, regional, national, and international levels to contain the virus. Laboratory-based studies demonstrate that masks may offer benefits in reducing the spread of droplet-based illnesses, but few data are available to assess mask effects via executive order on a population basis. We assess the effects of a county-wide mask order on per-population mortality, intensive care unit (ICU) utilization, and ventilator utilization in Bexar County, Texas. METHODS: We used publicly reported county-level data to perform a mixed-methods before-and-after analysis along with other sources of public data for analyses of covariance. We used a least-squares regression analysis to adjust for confounders. A Texas state-level mask order was issued on July 3, 2020, followed by a Bexar County-level order on July 15, 2020. We defined the control period as June 2 to July 2 and the postmask order period as July 8, 2020-August 12, 2020, with a 5-day gap to account for the median incubation period for cases; longer periods of 7 and 10 days were used for hospitalization and ICU admission/death, respectively. Data are reported on a per-100,000 population basis using respective US Census Bureau-reported populations. RESULTS: From June 2, 2020 through August 12, 2020, there were 40,771 reported cases of COVID-19 within Bexar County, with 470 total deaths. The average number of new cases per day within the county was 565.4 (95% confidence interval [CI] 394.6-736.2). The average number of positive hospitalized patients was 754.1 (95% CI 657.2-851.0), in the ICU was 273.1 (95% CI 238.2-308.0), and on a ventilator was 170.5 (95% CI 146.4-194.6). The average deaths per day was 6.5 (95% CI 4.4-8.6). All of the measured outcomes were higher on average in the postmask period as were covariables included in the adjusted model. When adjusting for traffic activity, total statewide caseload, public health complaints, and mean temperature, the daily caseload, hospital bed occupancy, ICU bed occupancy, ventilator occupancy, and daily mortality remained higher in the postmask period. CONCLUSIONS: There was no reduction in per-population daily mortality, hospital bed, ICU bed, or ventilator occupancy of COVID-19-positive patients attributable to the implementation of a mask-wearing mandate. [SEP]
Title: Analysis of the Effects of COVID-19 Mask Mandates on Hospital

```
Resource Consumption and Mortality at the County Level [SEP]
Abstract: {abstract} [SEP]
Title:
```

E.3 Correctness Functions

Rouge score. Recall-Oriented Understudy for Gist Evaluation (Rouge) score has originally been designed to evaluate machine translation or text summarization tasks. The Rouge score counts the overlapping n-grams between generated reference texts. Widely used n-grams include unigrams (Rouge-1), bigrams (Rouge-2), and the longest common subsequence (Rouge-L). Specifically, it is computed through

$$\text{ROUGE} = \frac{|(\text{n-gram} \in \text{Generation}) \cap (\text{n-gram}) \in \text{Reference}|}{|\text{Reference}|}.$$

METEOR score. The Metric for Evaluation of Translation with Explicit Ordering (METEOR) score has also been originally designed to evaluate machine translation and text summarization. Different from the Rouge score, the METEOR score considers the accuracy and fluency of the generation, as well as word order. Calculation of METEOR score can be found in Banerjee and Lavie (2005).

BERT-similarity. The BERT-similarity is based on sentence-bert (Reimers and Gurevych, 2019). Specifically, in the first step, reference and generation texts are encoded as 768-dimensional feature vectors, respectively. Then, the correctness values are computed by calculating the cosine similarity between reference and generation vectors. In our implementation, we use sentence-Bert with *bert-nli-mean-tokens* pre-trained weights as the encoding model.

ChatGPT evaluation. ChatGPT evaluation is calculated by prompting GPT-3.5-turbo with the question, reference, and generation; and asking it to evaluate the correctness of the generation. The template used in calculating ChatGPT correctness follows that in Lin et al. (2023):

```
Rate the level of consistency between the answer to the question and
the reference answer, from 0 to 100.
Question: In Scotland a bothy/bothie is a?
Reference: House
Answer: House
Rating: 100.
Question: Where in England was Dame Judi Dench born?
Reference: York
Answer: London
Rating: 0.
Question: {question}
Reference: {reference}
Answer: {generated}
Rating:
```

E.4 Inconsistency due to Correctness Thresholding

We provide more evidence to show the inconsistency of AUARC and AUPRC metrics caused by ad hoc correctness thresholding. The plots are in Fig 5, 6, 7, 8, and 9.

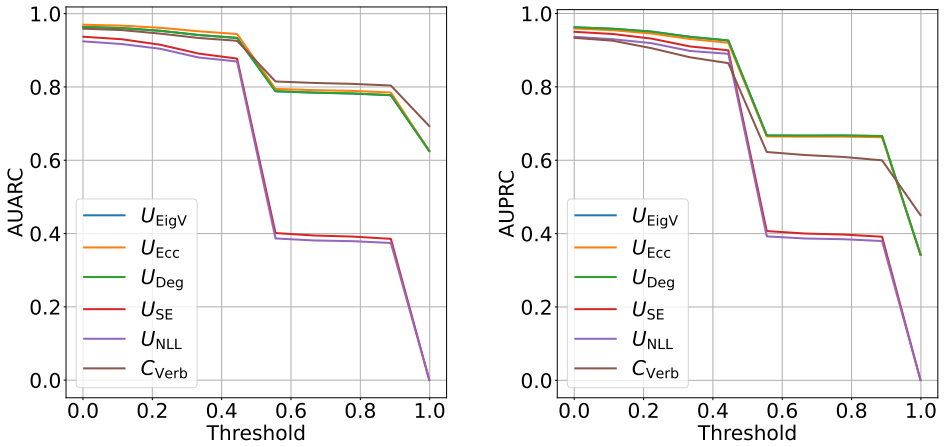
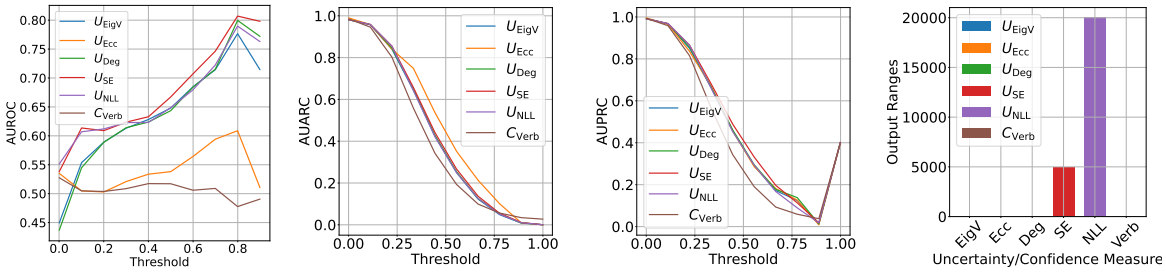
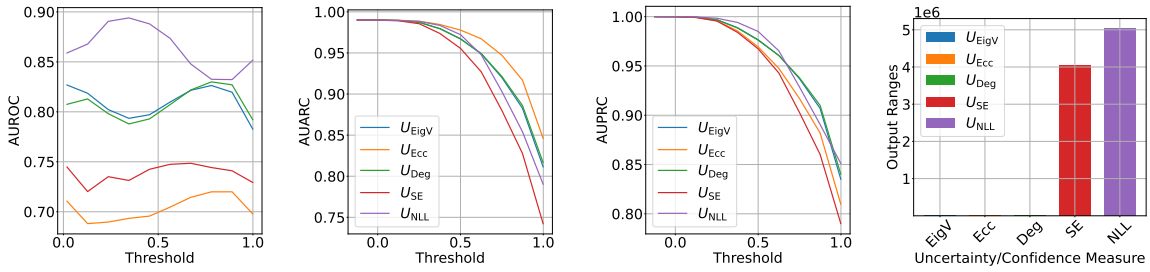


Figure 5: The assessed results for *AUARC (left)* and *AUPRC (right)* of uncertainty/confidence measures for GPT-3.5-turbo on the TriviaQA benchmark using the METEOR correctness score with varying thresholds.



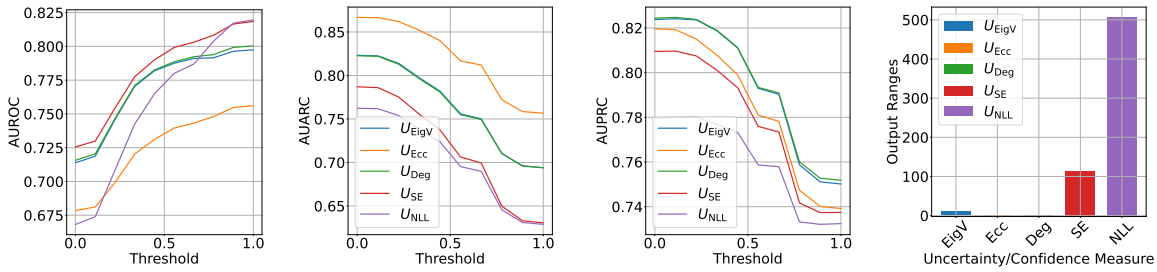
(a) AUROC (b) AUARC (c) AUPRC (d) Output ranges

Figure 6: Results for Meadow using GPT-3.5-turbo and the Rouge score.



(a) AUROC (b) AUARC (c) AUPRC (d) Output ranges

Figure 7: Results for TriviaQA using GPT-3.5-turbo with temperature 1.5 and the bert-similarity metric.

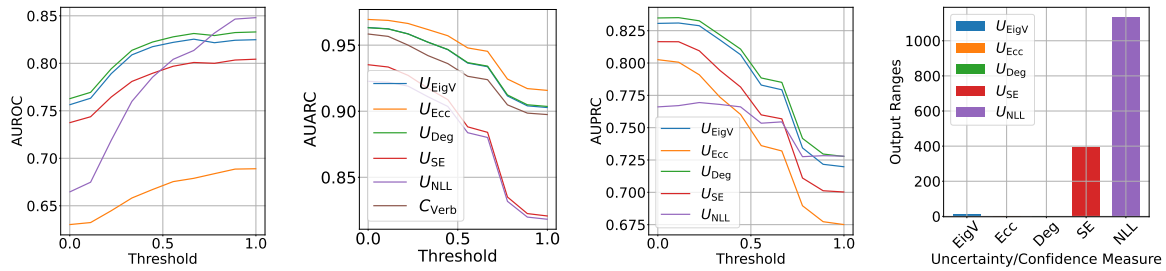


(a) AUROC (b) AUARC (c) AUPRC (d) Output ranges

Figure 8: Results for TriviaQA using Llama-2-7b-chat and the Rouge score.

model	dataset	correctness	temperature	U_{Ecc}	U_{Deg}	U_{EigV}	U_{NLL}	U_{SE}	C_{Verb}
Llama-2	nq-open	bert	0.6	$0.302_{\pm 0.044}$	$\mathbf{0.044}_{\pm 0.011}$	$0.046_{\pm 0.007}$	$0.121_{\pm 0.016}$	$0.122_{\pm 0.025}$	nan
		meteor	0.6	$0.293_{\pm 0.027}$	$\mathbf{0.072}_{\pm 0.010}$	$0.077_{\pm 0.015}$	$0.167_{\pm 0.021}$	$0.137_{\pm 0.024}$	nan
		rougeL	0.6	$0.297_{\pm 0.039}$	$0.058_{\pm 0.010}$	$\mathbf{0.051}_{\pm 0.010}$	$0.147_{\pm 0.021}$	$0.124_{\pm 0.019}$	nan
		rouge1	0.6	$0.297_{\pm 0.038}$	$0.057_{\pm 0.011}$	$\mathbf{0.051}_{\pm 0.010}$	$0.148_{\pm 0.021}$	$0.124_{\pm 0.020}$	nan
	squad	bert	0.6	$0.308_{\pm 0.041}$	$0.071_{\pm 0.013}$	$\mathbf{0.064}_{\pm 0.013}$	$0.072_{\pm 0.008}$	$0.181_{\pm 0.027}$	nan
		meteor	0.6	$0.299_{\pm 0.049}$	$0.252_{\pm 0.027}$	$\mathbf{0.247}_{\pm 0.029}$	$0.419_{\pm 0.018}$	$0.407_{\pm 0.024}$	nan
		rougeL	0.6	$0.359_{\pm 0.045}$	$\mathbf{0.139}_{\pm 0.033}$	$0.150_{\pm 0.027}$	$0.187_{\pm 0.028}$	$0.332_{\pm 0.036}$	nan
		rouge1	0.6	$0.360_{\pm 0.044}$	$\mathbf{0.141}_{\pm 0.034}$	$0.150_{\pm 0.027}$	$0.195_{\pm 0.032}$	$0.337_{\pm 0.035}$	nan
	triviaqa	bert	0.6	$0.312_{\pm 0.052}$	$\mathbf{0.020}_{\pm 0.005}$	$0.028_{\pm 0.007}$	$0.244_{\pm 0.012}$	$0.061_{\pm 0.008}$	nan
		meteor	0.6	$0.305_{\pm 0.048}$	$\mathbf{0.041}_{\pm 0.007}$	$0.049_{\pm 0.010}$	$0.271_{\pm 0.020}$	$0.052_{\pm 0.007}$	nan
		rougeL	0.6	$0.305_{\pm 0.050}$	$\mathbf{0.026}_{\pm 0.005}$	$0.033_{\pm 0.006}$	$0.206_{\pm 0.020}$	$0.051_{\pm 0.007}$	nan
		rouge1	0.6	$0.307_{\pm 0.049}$	$\mathbf{0.026}_{\pm 0.005}$	$0.034_{\pm 0.006}$	$0.209_{\pm 0.019}$	$0.052_{\pm 0.007}$	nan
Llama-2-chat	nq-open	bert	0.6	$0.199_{\pm 0.040}$	$\mathbf{0.046}_{\pm 0.008}$	$0.052_{\pm 0.010}$	$0.101_{\pm 0.015}$	$0.062_{\pm 0.010}$	nan
			1.0	$0.236_{\pm 0.033}$	$\mathbf{0.035}_{\pm 0.008}$	$0.038_{\pm 0.007}$	$0.097_{\pm 0.017}$	$0.055_{\pm 0.012}$	nan
		meteor	0.6	$0.190_{\pm 0.039}$	$\mathbf{0.062}_{\pm 0.008}$	$0.067_{\pm 0.010}$	$0.176_{\pm 0.018}$	$0.072_{\pm 0.009}$	nan
			1.0	$0.224_{\pm 0.034}$	$\mathbf{0.044}_{\pm 0.006}$	$0.046_{\pm 0.007}$	$0.209_{\pm 0.023}$	$0.074_{\pm 0.015}$	nan
		rougeL	0.6	$0.198_{\pm 0.039}$	$\mathbf{0.053}_{\pm 0.011}$	$0.057_{\pm 0.010}$	$0.167_{\pm 0.013}$	$0.060_{\pm 0.012}$	nan
			1.0	$0.227_{\pm 0.035}$	$0.035_{\pm 0.007}$	$\mathbf{0.033}_{\pm 0.006}$	$0.211_{\pm 0.021}$	$0.069_{\pm 0.016}$	nan
		rouge1	0.6	$0.199_{\pm 0.039}$	$\mathbf{0.054}_{\pm 0.010}$	$0.057_{\pm 0.010}$	$0.167_{\pm 0.014}$	$0.061_{\pm 0.013}$	nan
			1.0	$0.227_{\pm 0.035}$	$0.034_{\pm 0.007}$	$\mathbf{0.033}_{\pm 0.006}$	$0.212_{\pm 0.021}$	$0.069_{\pm 0.015}$	nan
	squad	bert	0.6	$0.208_{\pm 0.033}$	$0.065_{\pm 0.014}$	$0.075_{\pm 0.017}$	$\mathbf{0.048}_{\pm 0.007}$	$0.063_{\pm 0.012}$	nan
			1.0	$0.276_{\pm 0.039}$	$0.067_{\pm 0.011}$	$0.063_{\pm 0.010}$	$\mathbf{0.038}_{\pm 0.006}$	$0.098_{\pm 0.012}$	nan
		meteor	0.6	$0.216_{\pm 0.038}$	$0.303_{\pm 0.026}$	$0.265_{\pm 0.022}$	$\mathbf{0.063}_{\pm 0.013}$	$0.182_{\pm 0.029}$	nan
			1.0	$0.300_{\pm 0.046}$	$0.292_{\pm 0.035}$	$0.250_{\pm 0.027}$	$\mathbf{0.064}_{\pm 0.011}$	$0.274_{\pm 0.021}$	nan
		rougeL	0.6	$0.239_{\pm 0.036}$	$0.177_{\pm 0.026}$	$0.143_{\pm 0.020}$	$\mathbf{0.052}_{\pm 0.011}$	$0.127_{\pm 0.020}$	nan
			1.0	$0.304_{\pm 0.036}$	$0.179_{\pm 0.033}$	$0.137_{\pm 0.024}$	$\mathbf{0.053}_{\pm 0.012}$	$0.210_{\pm 0.027}$	nan
		rouge1	0.6	$0.238_{\pm 0.037}$	$0.183_{\pm 0.027}$	$0.148_{\pm 0.022}$	$\mathbf{0.053}_{\pm 0.010}$	$0.129_{\pm 0.021}$	nan
			1.0	$0.303_{\pm 0.035}$	$0.185_{\pm 0.033}$	$0.143_{\pm 0.025}$	$\mathbf{0.053}_{\pm 0.012}$	$0.213_{\pm 0.026}$	nan
	triviaqa	bert	0.6	$0.140_{\pm 0.024}$	$0.062_{\pm 0.016}$	$0.061_{\pm 0.015}$	$\mathbf{0.020}_{\pm 0.004}$	$0.027_{\pm 0.007}$	nan
			1.0	$0.213_{\pm 0.030}$	$0.025_{\pm 0.006}$	$0.034_{\pm 0.006}$	$\mathbf{0.014}_{\pm 0.002}$	$0.036_{\pm 0.006}$	nan
		meteor	0.6	$0.145_{\pm 0.027}$	$0.067_{\pm 0.017}$	$0.064_{\pm 0.015}$	$\mathbf{0.034}_{\pm 0.009}$	$0.075_{\pm 0.016}$	nan
			1.0	$0.206_{\pm 0.032}$	$\mathbf{0.035}_{\pm 0.007}$	$0.046_{\pm 0.005}$	$0.049_{\pm 0.008}$	$0.084_{\pm 0.007}$	nan
		rougeL	0.6	$0.141_{\pm 0.021}$	$0.062_{\pm 0.014}$	$0.061_{\pm 0.014}$	$\mathbf{0.024}_{\pm 0.005}$	$0.034_{\pm 0.005}$	nan
			1.0	$0.204_{\pm 0.035}$	$0.027_{\pm 0.006}$	$0.040_{\pm 0.004}$	$\mathbf{0.022}_{\pm 0.002}$	$0.051_{\pm 0.007}$	nan
		rouge1	0.6	$0.141_{\pm 0.021}$	$0.062_{\pm 0.014}$	$0.062_{\pm 0.013}$	$\mathbf{0.024}_{\pm 0.005}$	$0.034_{\pm 0.006}$	nan
			1.0	$0.203_{\pm 0.035}$	$0.027_{\pm 0.006}$	$0.040_{\pm 0.004}$	$\mathbf{0.022}_{\pm 0.002}$	$0.051_{\pm 0.007}$	nan
GPT-3.5	meadow	bert	1.0	$0.284_{\pm 0.035}$	$0.178_{\pm 0.030}$	$0.174_{\pm 0.025}$	$\mathbf{0.112}_{\pm 0.022}$	$0.177_{\pm 0.027}$	$0.288_{\pm 0.033}$
		meteor	1.0	$0.292_{\pm 0.045}$	$0.134_{\pm 0.027}$	$0.137_{\pm 0.026}$	$\mathbf{0.074}_{\pm 0.012}$	$0.132_{\pm 0.018}$	$0.263_{\pm 0.050}$
		rougeL	1.0	$0.278_{\pm 0.045}$	$0.130_{\pm 0.022}$	$0.131_{\pm 0.025}$	$\mathbf{0.056}_{\pm 0.010}$	$0.113_{\pm 0.022}$	$0.289_{\pm 0.046}$
		rouge1	1.0	$0.290_{\pm 0.047}$	$0.126_{\pm 0.018}$	$0.135_{\pm 0.020}$	$\mathbf{0.059}_{\pm 0.013}$	$0.113_{\pm 0.018}$	$0.299_{\pm 0.047}$
	nq-open	bert	1.0	$0.151_{\pm 0.025}$	$0.050_{\pm 0.012}$	$0.065_{\pm 0.014}$	$\mathbf{0.039}_{\pm 0.008}$	$0.050_{\pm 0.007}$	$0.487_{\pm 0.005}$
		meteor	1.0	$0.154_{\pm 0.027}$	$0.050_{\pm 0.011}$	$0.063_{\pm 0.011}$	$\mathbf{0.046}_{\pm 0.011}$	$0.060_{\pm 0.009}$	$0.452_{\pm 0.018}$
		rougeL	1.0	$0.151_{\pm 0.022}$	$0.048_{\pm 0.011}$	$0.062_{\pm 0.012}$	$\mathbf{0.034}_{\pm 0.009}$	$0.052_{\pm 0.008}$	$0.487_{\pm 0.006}$
		rouge1	1.0	$0.153_{\pm 0.023}$	$0.048_{\pm 0.011}$	$0.063_{\pm 0.012}$	$\mathbf{0.034}_{\pm 0.009}$	$0.051_{\pm 0.008}$	$0.487_{\pm 0.006}$
	squad	bert	1.0	$0.204_{\pm 0.025}$	$0.237_{\pm 0.024}$	$0.240_{\pm 0.019}$	$\mathbf{0.065}_{\pm 0.012}$	$0.113_{\pm 0.013}$	$0.181_{\pm 0.029}$
		meteor	1.0	$0.181_{\pm 0.012}$	$0.151_{\pm 0.016}$	$0.193_{\pm 0.020}$	$\mathbf{0.054}_{\pm 0.017}$	$0.086_{\pm 0.014}$	$0.182_{\pm 0.032}$
		rougeL	1.0	$0.222_{\pm 0.025}$	$0.270_{\pm 0.023}$	$0.269_{\pm 0.016}$	$\mathbf{0.037}_{\pm 0.010}$	$0.100_{\pm 0.011}$	$0.168_{\pm 0.035}$
		rouge1	1.0	$0.226_{\pm 0.024}$	$0.276_{\pm 0.023}$	$0.270_{\pm 0.017}$	$\mathbf{0.039}_{\pm 0.010}$	$0.103_{\pm 0.011}$	$0.168_{\pm 0.035}$
	triviaqa	bert	0.5	$0.215_{\pm 0.042}$	$0.212_{\pm 0.040}$	$0.212_{\pm 0.041}$	$\mathbf{0.043}_{\pm 0.006}$	$0.052_{\pm 0.009}$	nan
			1.0	$0.152_{\pm 0.025}$	$0.129_{\pm 0.020}$	$0.133_{\pm 0.020}$	$\mathbf{0.039}_{\pm 0.007}$	$0.052_{\pm 0.012}$	$0.182_{\pm 0.025}$
			1.5	$0.142_{\pm 0.018}$	$0.053_{\pm 0.011}$	$0.074_{\pm 0.012}$	$\mathbf{0.031}_{\pm 0.007}$	$0.081_{\pm 0.009}$	nan
		meteor	0.5	$0.215_{\pm 0.049}$	$0.211_{\pm 0.045}$	$0.208_{\pm 0.047}$	$\mathbf{0.179}_{\pm 0.021}$	$0.234_{\pm 0.019}$	nan
			1.0	$0.156_{\pm 0.026}$	$0.131_{\pm 0.024}$	$\mathbf{0.131}_{\pm 0.022}$	$0.146_{\pm 0.011}$	$0.209_{\pm 0.012}$	$0.194_{\pm 0.036}$
			1.5	$0.137_{\pm 0.024}$	$\mathbf{0.059}_{\pm 0.011}$	$0.077_{\pm 0.012}$	$0.119_{\pm 0.010}$	$0.176_{\pm 0.015}$	nan
		rougeL	0.5	$0.214_{\pm 0.046}$	$0.210_{\pm 0.042}$	$0.207_{\pm 0.041}$	$\mathbf{0.041}_{\pm 0.007}$	$0.050_{\pm 0.008}$	nan
			1.0	$0.151_{\pm 0.024}$	$0.126_{\pm 0.019}$	$0.129_{\pm 0.019}$	$\mathbf{0.038}_{\pm 0.007}$	$0.059_{\pm 0.009}$	$0.181_{\pm 0.026}$
			1.5	$0.138_{\pm 0.025}$	$0.059_{\pm 0.012}$	$0.079_{\pm 0.011}$	$\mathbf{0.034}_{\pm 0.008}$	$0.104_{\pm 0.007}$	nan
		rouge1	0.5	$0.216_{\pm 0.046}$	$0.212_{\pm 0.043}$	$0.209_{\pm 0.042}$	$\mathbf{0.040}_{\pm 0.007}$	$0.050_{\pm 0.008}$	nan
			1.0	$0.152_{\pm 0.024}$	$0.126_{\pm 0.018}$	$0.130_{\pm 0.021}$	$\mathbf{0.039}_{\pm 0.007}$	$0.060_{\pm 0.009}$	$0.176_{\pm 0.027}$
			1.5	$0.137_{\pm 0.023}$	$0.060_{\pm 0.011}$	$0.078_{\pm 0.012}$	$\mathbf{0.034}_{\pm 0.008}$	$0.105_{\pm 0.008}$	nan

Table 2: RCE results for various experimental configurations.



(a) AUROC (b) AUARC (c) AUPRC (d) Output ranges

Figure 9: Results for TriviaQA using Llama-2-7b-chat using temperature 1.0 and the Rouge score.

F Additional Experimental Results

Prompt	Reference	Generation	$\mathbb{P}(U_{\text{Ecc}} \leq u)$	$\mathbb{P}(U_{\text{Deg}} \leq u)$	$\mathbb{P}(U_{\text{EigV}} \leq u)$	$\mathbb{P}(U_{\text{SE}} \leq u)$	$\mathbb{P}(U_{\text{NLL}} \leq u)$
Q: Who did Dr. Crippen murder?	his wife	His wife	0.999	0.881	0.822	0.649	0.247
Q: What are the only two musical notes which have no flats?	c and f	B and F	0.999	0.761	0.769	0.898	0.691
Q: Which Eastenders actor has played the policeman Nick Rowan on TV?	nick berry	Mark Jordon	0.999	0.972	0.978	0.954	0.918
Q: Which ‘B’ was the name of the mechanical shark used in the original ‘Jaws‘ film?	bruce	Bruce	0.999	0.761	0.769	0.337	0.183
Q: Which actor does the interviewing in ’Interview with a Vampire’?	christian slater	Brad Pitt	0.999	0.858	0.856	0.861	0.893
Q: What did my true love bring to me on the Sixth Day of Christmas?	six geese-a-laying	Six geese a-laying	0.999	0.761	0.769	0.736	0.688
Q: In January 1957, Russell Endean became the first batsman to be dismissed from a test cricket match for doing what?	handling the ball	Handling the ball	0.999	0.761	0.769	0.901	0.368
Q: What are the first names of the two dancing instructors in the UK television series ‘Hi De Hi’?	barry and yvonne	Barry and Yvonne	0.999	0.761	0.769	0.846	0.627
Q: Who became the host of the UK television game show Blankety Blank in 1984?	les dawson	Les Dawson	0.999	0.761	0.769	0.180	0.040
Q: How much, in pounds sterling, does the Best in Show Winner receive at the annual Crufts Dog Show?	100 pounds	£100	0.999	0.920	0.908	0.830	0.787
Q: In the Billy Bunter stories, what is the surname of Bunter’s form teacher?	quelch	Quelch	0.999	0.761	0.769	0.999	0.558
Q: Which play is featured in the film The Producers?	springtime for hitler	Springtime for Hitler	0.999	0.761	0.769	0.967	0.341
Q: What provoked the war between Honduras and El Salvador in 1969?	a football match	A soccer match	0.999	0.761	0.769	0.535	0.711
Q: Which character was played by Linda Thorson in The Avengers?	tara king	Tara King	0.999	0.824	0.885	0.919	0.399
Q: According to a traditional English proverb, what is better than none?	half a loaf	A bad excuse	0.999	0.972	0.978	0.931	0.908
Q: In which Welsh village is there only one gay, apparently?!	llanddewi breffi	Llanddewi Brefi	0.999	0.926	0.963	0.950	0.906
Q: On September 28th, NASA announced that what had been detected on Mars?	flowing water	Possible signs of life	0.999	0.965	0.963	0.813	0.930
Q: What are the first four words of the Bible, as recorded in Genesis?	in the beginning god	In the beginning, God	0.653	0.650	0.651	0.574	0.557
Q: Which national anthem was originally called the ’War Song for the Rhine Army’?	marsellaise	German national anthem	0.694	0.858	0.837	0.785	0.888
Q: Name the UK budget holiday company specialising in Turkey and Greece which went bust in July 2010?	goldtrail	Goldtrail	0.999	0.920	0.902	0.894	0.655
Q: Who has been President of France twice, but never been elected to the position?	alain poher	François Mitterrand	0.999	0.920	0.902	0.854	0.864
Q: What is the name of Madonna’s proposed chain of fitness clubs?	hard candy fitness	Hard Candy Fitness	0.999	0.761	0.769	0.996	0.183
Q: Elvis Presley sang a few lines in German on which US hit song?	wooden heart	Wooden Heart	0.999	0.761	0.769	0.998	0.270
Q: What was the name of the book that was a collection of Aubrey Beardsley’s work, published by Leonard Smithers in 1897?	a book of fifty drawings	The Yellow Book	0.999	0.761	0.769	0.950	0.775
Q: Dishes prepared with spinach can be referred to as what?	la florentine	Spinach dishes	0.999	0.920	0.902	0.943	0.899
Q: Which English civil engineer’s most famous project was the construction of Tower Bridge over the River Thames in London?	sir john wolfe-barry	Sir John Wolfe Barry	0.999	0.761	0.769	0.830	0.633
Q: Where did the space probe New Horizons launched by NASA in 2006 aim to investigate?	pluto and the kuiper belt	Pluto and the Kuiper Belt	0.999	0.905	0.904	0.905	0.576
Q: Where woud you find a nave or an apse?	in a church	In a church	0.999	0.761	0.769	0.236	0.185
Q: What is the name of Jay-Z and Beyonce’s daughter?	blue ivy	Blue Ivy	0.999	0.976	0.965	0.975	0.354
Q: ’Feel Like Making Love’ and ’The First Time Ever I Saw Your Face’ were hit singles for which female artist?	roberta flack	Roberta Flack	0.999	0.761	0.769	0.864	0.046
Q: In the nursery rhyme, who pulled pussy out of the well?	little tommy stout	Tommy	0.999	0.976	0.987	0.962	0.882
Q: "In the film of the same name, what was the name of ""The Hustler""?"	"""fast eddie"" felson"	Fast Eddie Felson	0.999	0.761	0.769	0.708	0.692
Q: In Camberwick Green on Children’s TV who was the commander of Pippin Fort?	captain snort	Captain Snort	0.999	0.761	0.769	0.961	0.156
Q: In Chigley on Children’s TV who owned the steam railway and drove the steam engine ’Bessie’?	lord belborough	Lord Belborough	0.999	0.761	0.769	0.951	0.401
Q: Who won the gold medal in the women’s Skeleton Bob at the 2010 Vancouver Winter Olympics?	amy williams	Amy Williams	0.999	0.881	0.822	0.676	0.265
Q: What decoration, a Cross, was first awarded in 1995 to Corporal Wayne Mills for his actions in Bosnia?	conspicuous gallantry	George Cross	0.999	0.844	0.783	0.801	0.899
Q: What was the French sounding winner of the 2011 Epsom Derby?	pour moi	Pour Moi	0.999	0.761	0.769	0.321	0.101
Q: Who originally provided the voice for TV’s ’Basil Brush’?	ivan owen	Ivan Owen	0.999	0.761	0.769	0.987	0.454
Q: "Which actress played ’Valeria"" in the film Carry On Screaming?"	fenella fielding	Fenella Fielding	0.999	0.761	0.769	0.862	0.206
Q: Which of the ’Spice Girls’ advertised ’Milky Way’ ob t.v.?	emma bunton (baby spice)	Victoria Beckham (Posh Spice)	0.999	0.949	0.963	0.985	0.847
Q: Give any year in the life of the Portuguese prince known as Henry the Navigator.	1394-1460	1394-1460	0.999	0.761	0.769	0.680	0.671
Q: On which horse did Sir Gordon Richards ride his only Epsom Derby winner?	pinza	Pinza	0.999	0.824	0.885	0.987	0.229
Q: What was the name of the aeroplane in which Wiley Post became the first pilot to fly solo around the world?	’winnie mae’	Winnie Mae	0.999	0.761	0.769	0.849	0.654
Q: Who was the husband of Rebekah Brooks from 2002 to 2009?	ross kemp	Ross Kemp	0.999	0.761	0.769	0.826	0.746
Q: Whole Again and Eternal Flame were Number Ones for which girl group in 2001?	atomic kitten	Atomic Kitten	0.999	0.761	0.769	0.180	0.026
Q: During a penalty shoot out in soccer where should the non participating players be	in the centre circle	Outside of the penalty area	0.999	0.985	0.987	0.987	0.960
Q: On which game show was Bobby Charlton once a contestant and winner	double your money	A Question of Sport	0.999	0.961	0.963	0.987	0.952
Q: From ’On Her Majesty’s Secret Service’ (1969), as Bond passes a janitor in Draco’s headquarters, the man can be heard whistling what?	the goldfinger (1964) theme	"Goldfinger"	0.999	0.944	0.940	0.984	0.886
Q: A Paris grocer was jailed for two years in 1978 stabbing wife what?	a wedge of hard cheese	Knife	0.999	0.976	0.987	0.974	0.849

Table 3: Examples of correctness and the according uncertainty levels.

F.1 Qualitative Illustration

$\boldsymbol{x}$: In 1840 the world's first postage stamps printed were the Penny Black and which other?
$\boldsymbol{y}$: twopenny blue
$\widehat{\boldsymbol{y}}$: The Penny Red
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.825
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.864

$\boldsymbol{x}$: Championship dragon boat racing calls for a specialised long boat, a team of paddlers (typically 20), a sweeper to steer and which other of these?
$\boldsymbol{y}$: a drummer and drum
$\widehat{\boldsymbol{y}}$: A drummer
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.946
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.704

$\boldsymbol{x}$: Who has the highest suicide rate in the UK?
$\boldsymbol{y}$: men - by a ratio of roughly 4 to 1
$\widehat{\boldsymbol{y}}$: Middle-aged men
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.745
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.894

$\boldsymbol{x}$: Which East Midlands club holds the Football League record for most games played?
$\boldsymbol{y}$: nots county
$\widehat{\boldsymbol{y}}$: Notts County
$\mathbb{P}(U_{\text{SE}} \leq u)$: 0.842
$\mathbb{P}(U_{\text{NLL}} \leq u)$: 0.793

We provide more instances to show the qualitative effect of our RCE-based assessment in Table 3.

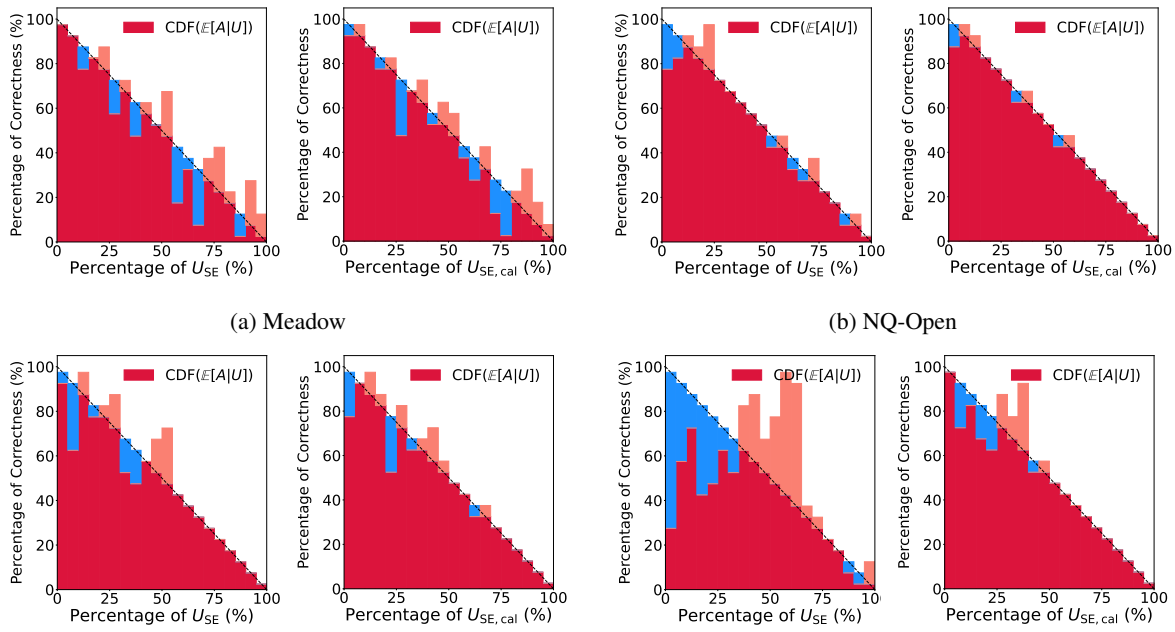
F.2 Recalibration with Histogram Binning

We use equal-mass histogram binning to recalibrate, in a post-hoc manner, the performance of an uncertainty (or confidence) measure on a specific benchmark. Specifically, given a dataset $\{(u_i, a_i)\}_{i=1}^{n}$ of uncertainty and correctness values computed over a benchmark, where each $u_i = U(\boldsymbol{x}; \widehat{\boldsymbol{y}}_i)$, $a_i = A(\boldsymbol{x}_i; \widehat{\boldsymbol{y}}_i)$, and $\widehat{\boldsymbol{y}}_i$ is a response generated by the LM. Then, we first randomly split it into the calibration set $\{(u_i, a_i)\}_{i=1}^{n_{\text{cal}}}$ and the test set $\{(u_i, a_i)\}_{i=n_{\text{cal}}+1}^{n}$. Similar to the operations in Sec. 4.3, we partition the range of U into B bins $\{\text{bin}_b\}_{b=1}^{B}$ whose boundaries are quantiles of $\{(u_i, a_i)\}_{i=n_{\text{cal}}+1}^{n}$. Then, we estimate the expected correctness level over the bin_b as

$$\text{crc}_{b,\text{cal}} := \frac{1}{|\mathcal{I}_{b,\text{cal}}|} \sum_{i \in \mathcal{I}_{b,\text{cal}}} a_i$$

where $\mathcal{I}_{b,\text{cal}} \triangleq \{i : 1 \leq i \leq n_{\text{cal}}, u_i \in \text{bin}_b\}$. We re-calibrate the measure U, defining U_{cal} via $U_{\text{cal}}(\boldsymbol{x}; \widehat{\boldsymbol{y}}) = \text{crc}_{b,\text{cal}}$ for any $U(\boldsymbol{x}; \widehat{\boldsymbol{y}}) \in \text{bin}_b$. We evaluate the performance of the calibrated measure on the test set. Table 4 and Fig. 10 and 11 list the RCE results of U_{SE} for GPT-3.5-turbo before and after calibration. We observe the calibrated measure is significantly better rank-calibrated, showing the effectiveness of this strategy. Here, we split the benchmark data equally into calibration and test sets.

While effective, one should note that such a post-hoc recalibration strategy concerns a specific benchmark and is not a focus of our work. We leave devising benchmark-agnostic calibrated uncertainty/confidence measures for future work.



(a) Meadow

(b) NQ-Open

(c) Squad

(d) TrviaQA

Figure 10: Indication diagrams of U_{SE} and $U_{\text{SE,cal}}$ (post-calibrated) for GPT-3.5-turbo (temperature 1.0) on various benchmarks with the Meteor correctness.

F.3 Critical Difference Diagrams

Here, we propose to combine the RCE metric with the *critical difference* (CD) diagram (Demšar, 2006). Critical Difference diagrams are built on the Wilcoxon signed rank test and the Friedman test, giving a non-parametric comparison of multiple approaches aggregated over several trials.

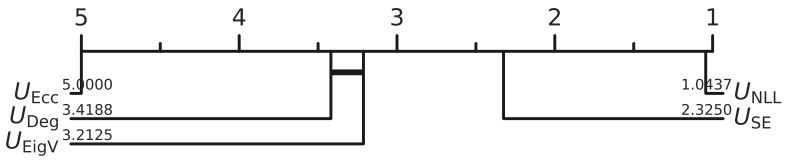
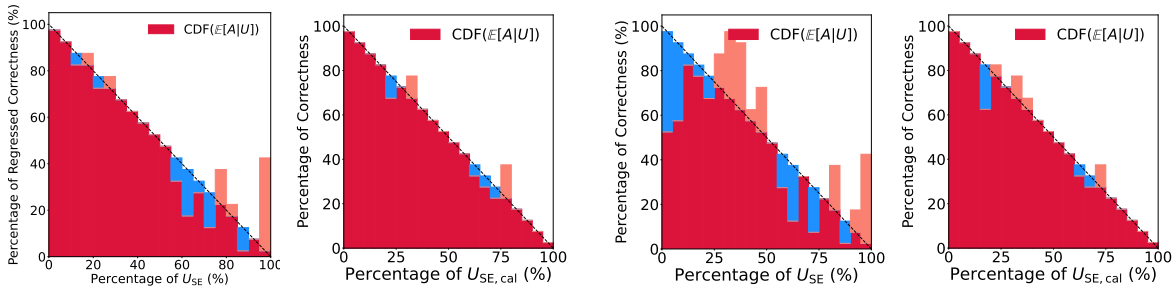


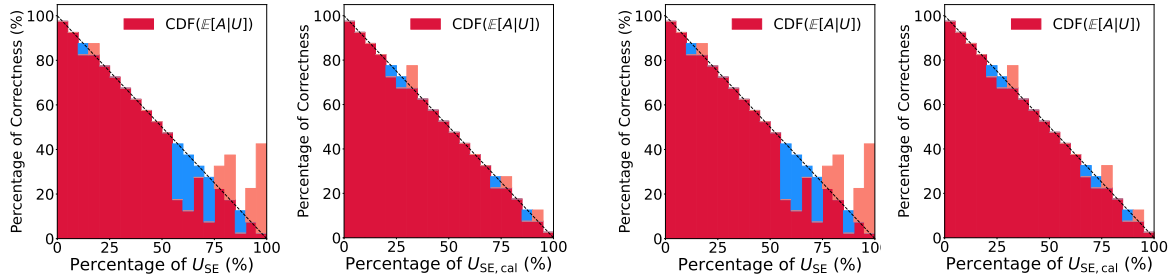
Figure 12: CD diagram of Llama-2-chat on TriviaQA.

As a demonstration, the CD diagram of assessed measures for Llama-2-chat on TriviaQA is shown in Fig. 12. The positions of various methods represent their averaged ranks over various experimental configurations (*e.g.*, temperature, LM, bootstrap, etc), where a lower averaged rank indicates that the corresponding measure (*e.g.*, 1.04 for U_{NLL}) performs better than others in an averaged sense. Here, a thick horizontal segment connects measures (*e.g.*, U_{Deg} and U_{EigV}) if the difference between their averaged ranks is within the critical length determined by related hypothesis testing procedures. Measures that are disconnected (*e.g.*, U_{Ecc}, U_{Deg}, and U_{NLL}) have statistically significant differences in performance.



(a) Bert Similarity

(b) Meteor Score



(c) Rouge Score

(d) Rouge1 Score

Figure 11: Indication diagrams of U_{SE} and $U_{\text{SE,cal}}$ (post-calibrated) for GPT-3.5-turbo (temperature 1.5) on TriviaQA with various correctness scores.

dataset	correctness	temperature	U_{SE}	$U_{\text{SE,cal}}$
meadow	bert	1.0	$0.177_{\pm 0.027}$	$\mathbf{0.083}_{\pm 0.016}$
	meteor	1.0	$0.132_{\pm 0.018}$	$\mathbf{0.066}_{\pm 0.015}$
	rougeL	1.0	$0.113_{\pm 0.022}$	$\mathbf{0.063}_{\pm 0.014}$
	rouge1	1.0	$0.113_{\pm 0.018}$	$\mathbf{0.061}_{\pm 0.012}$
nq-open	bert	1.0	$0.050_{\pm 0.007}$	$\mathbf{0.026}_{\pm 0.007}$
	meteor	1.0	$0.060_{\pm 0.009}$	$\mathbf{0.033}_{\pm 0.011}$
	rougeL	1.0	$0.052_{\pm 0.008}$	$\mathbf{0.030}_{\pm 0.010}$
	rouge1	1.0	$0.051_{\pm 0.008}$	$\mathbf{0.029}_{\pm 0.010}$
squad	bert	1.0	$0.113_{\pm 0.013}$	$\mathbf{0.050}_{\pm 0.013}$
	meteor	1.0	$0.086_{\pm 0.014}$	$\mathbf{0.046}_{\pm 0.010}$
	rougeL	1.0	$0.100_{\pm 0.011}$	$\mathbf{0.037}_{\pm 0.008}$
	rouge1	1.0	$0.103_{\pm 0.011}$	$\mathbf{0.039}_{\pm 0.007}$
triviaqa	bert	0.5	$0.052_{\pm 0.009}$	$\mathbf{0.030}_{\pm 0.010}$
		1.0	$0.052_{\pm 0.012}$	$\mathbf{0.027}_{\pm 0.008}$
		1.5	$0.081_{\pm 0.009}$	$\mathbf{0.029}_{\pm 0.007}$
	meteor	0.5	$0.234_{\pm 0.019}$	$\mathbf{0.058}_{\pm 0.015}$
		1.0	$0.209_{\pm 0.012}$	$\mathbf{0.047}_{\pm 0.014}$
		1.5	$0.176_{\pm 0.015}$	$\mathbf{0.036}_{\pm 0.012}$
	rougeL	0.5	$0.050_{\pm 0.008}$	$\mathbf{0.028}_{\pm 0.007}$
		1.0	$0.059_{\pm 0.009}$	$\mathbf{0.026}_{\pm 0.007}$
		1.5	$0.104_{\pm 0.007}$	$\mathbf{0.028}_{\pm 0.006}$
	rouge1	0.5	$0.050_{\pm 0.008}$	$\mathbf{0.028}_{\pm 0.006}$
		1.0	$0.060_{\pm 0.009}$	$\mathbf{0.027}_{\pm 0.006}$
		1.5	$0.105_{\pm 0.008}$	$\mathbf{0.028}_{\pm 0.008}$

Table 4: RCE results of U_{SE} and $U_{\text{SE,cal}}$ after rank-calibration for GPT-3.5-turbo with various experimental configurations.

F.4 Robustness Analysis

The RCE of uncertainty measures in practice may be affected by several factors. Therefore, we conduct ablation studies to analyze whether RCE is robust to two crucial key factors: correctness scores and model temperatures.

Correctness functions. We show RCEs for various models and correctness scores on TriviaQA and SQuAD in Fig 13. Each result is obtained using bootstrapping with 20 fixed seeds. We observe that the ranking of uncertainty measures is robust to correctness scores. For instance, we show the critical diagrams using GPT-3.5 on TriviaQA with varying correctness scores in Fig 14. In this setting, U_{NLL}, U_{SE} and C_{Verb} rank consistently higher across different correctness scores. Second, as shown in Table 2, RCE values using different correctness scores are relatively stable. For instance, when using GPT-3.5 on TriviaQA, the RCE values of NLL are 0.065, 0.054, 0.037, and 0.039 with bert_similarity, meteor, rouge-L, and rouge-1 scores, which are close.

Temperature setting. We show the RCEs for various models and temperatures on TriviaQA and SQuAD in Fig. 15. As above, each result is obtained using bootstrapping with 20 fixed seeds. The findings are similar to those regarding correctness scores. First, as shown in Fig. 16, while RCE values are not constant, U_{NLL} ranks consistently highest across different temperatures. When only the best uncertainty measure is considered, the RCE rankings at different temperatures give consistent results. Second, the RCE values are stable across different temperatures. For instance, when using GPT-3.5 with the Rouge-L score, the RCE values are 0.041, 0.038, 0.034 with temperatures 0.5, 1.0, and 1.5.

F.5 Conclusive Comparison

While the RCE values and rankings are often stable when correctness score and temperature vary, there are exceptional situations where uncertainty measures rankings might fluctuate. This poses a challenge when aiming for conclusive comparisons for uncertainty measures across varying hyperparameter situations. To make conclusive comparisons aiming to identify a best method, we can use CD diagrams by taking multiple hyperparameter choices into account. For example, to draw conclusions agnostic to model temperature, we plot CD diagrams that show RCE rankings averaged from data collected at different temperatures, as shown in Fig. 17. Based on these results, comparisons agnostic to the temperature can be made: U_{NLL} overall outperforms other methods with GPT-3.5 and Llama-2-chat on TriviaQA; U_{EigV} and U_{Deg} overall show statistically similar performance with Llama-2-chat on TriviaQA.

F.6 Library Information

The details of the main libraries used in our experiments are as in Table 5.

Package	Version	Package	Version
transformer (Wolf et al., 2020)	4.32.1	nltk (Bird et al., 2009)	3.8.1
spacy (Honnibal and Montani, 2017)	3.6.1	torch (Paszke et al., 2019)	2.0.1
rouge-score (Lin, 2004)	0.1.2		

Table 5: Information on main libraries used.

F.7 Artifact License and Terms

We use four datasets, namely, Natural Questions, TriviaQA, SQuAD-1, and Meadow. Natural Questions is under the **CC BY-SA 3.0 license**, TriviaQA and Meadow are under the **Apache License 2.0**, and SQuAD-1 is under the **CC BY-SA 4.0 license**. We used two LLMs, namely *ChatGPT-3.5* and *Llama-2*. ChatGPT-3.5-turbo usage is subject to OpenAI's *Sharing & Publication Policy* and *Usage Policies*. Llama-2 is under the Llama-2 Community License (Meta, 2023). Our implementation and the data collected are under the **MIT License**.

Our use of the existing artifacts is consistent with their original intended use. Our created artifacts intend to verify our proposed method in our submission, which is consistent with the original access conditions.

G AI Assistant Usage

We used *Copilot* to assist with coding.

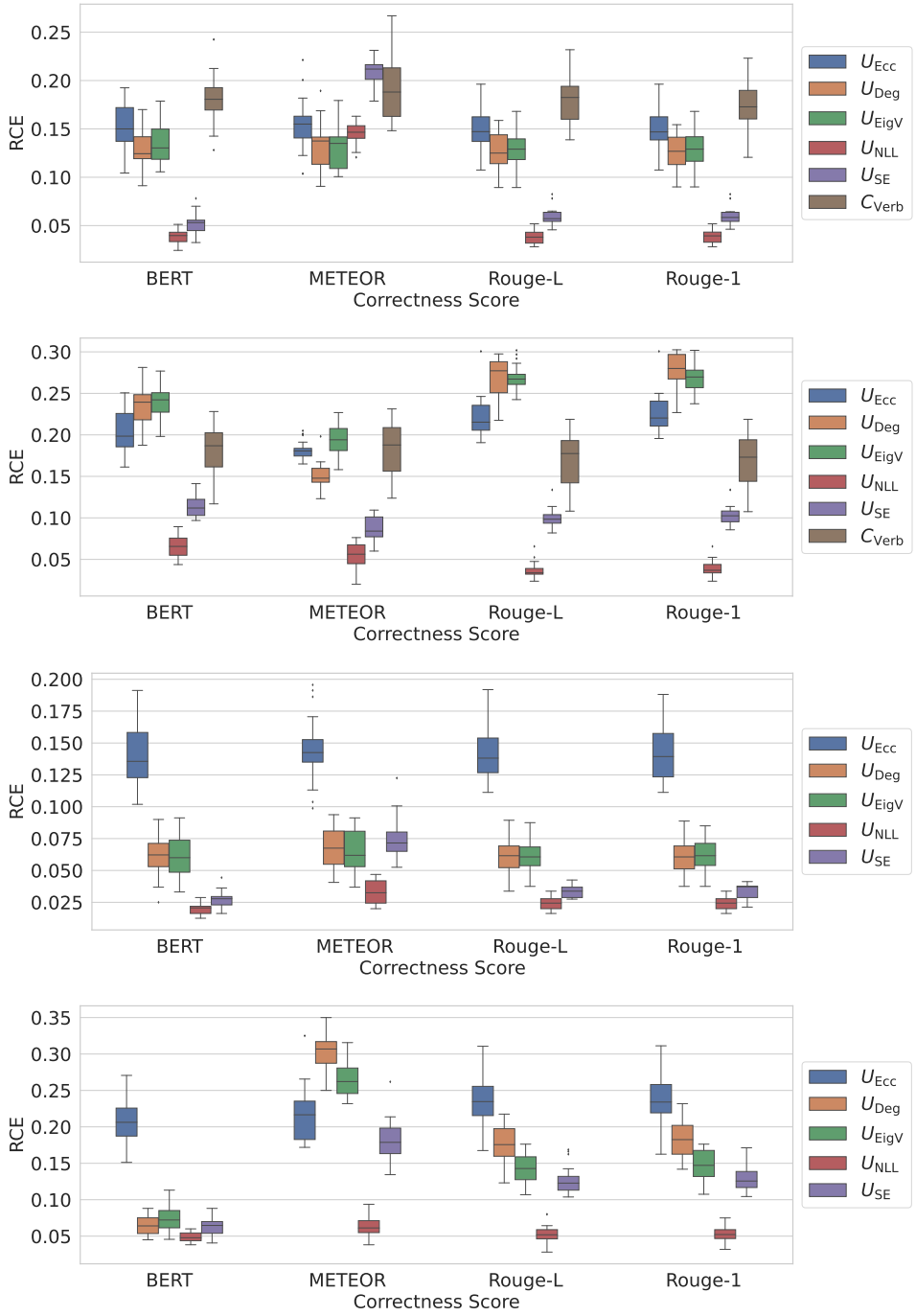


Figure 13: Box plots with various correctness functions under various configurations. The first row is for GPT-3.5-turbo on TriviaQA; the second row is for GPT-3.5-turbo on SQuAD; the third is for Llama-2-7b-chat on TriviaQA; and the fourth row is for Llama-2-7b-chat on SQuAD.

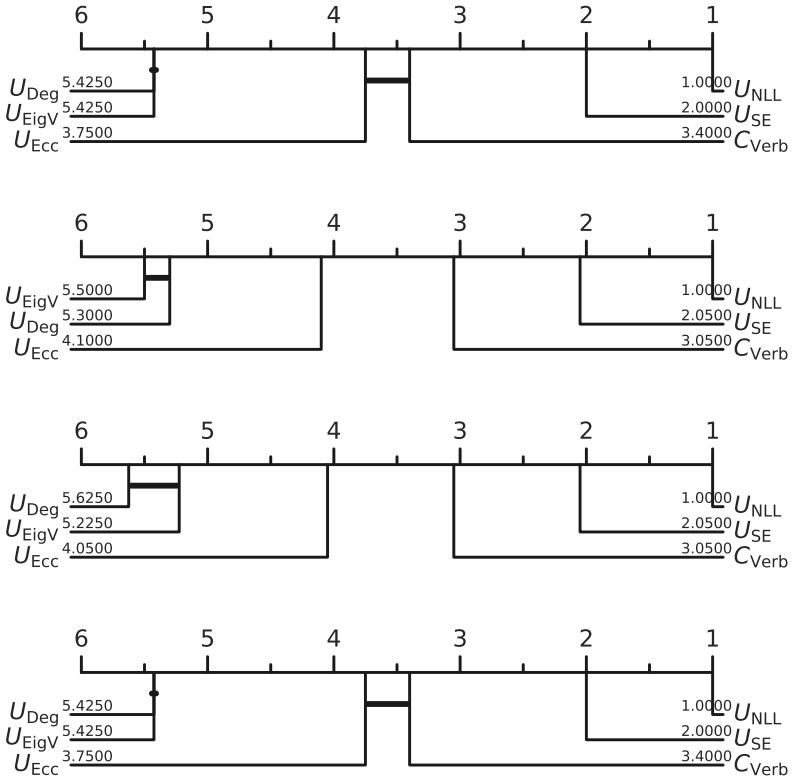


Figure 14: CD diagrams using GPT-3.5 on TriviaQA with different correctness scores.

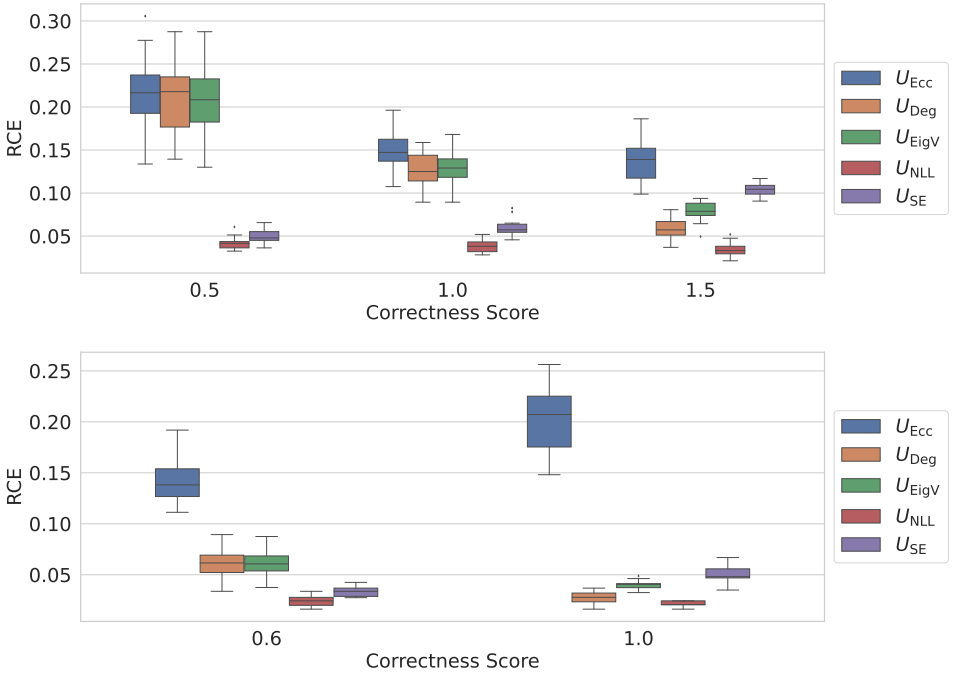


Figure 15: Box plots based on the generations of GPT-3.5-turbo and Llama-2-7b-chat with varying temperatures. The first row represents GPT-3.5-turbo with temperatures 0.5, 1.0, and 1.5, while the second row represents Llama-2-7b-chat with temperatures 0.6 and 1.0. Both results are evaluated on TriviaQA dataset.

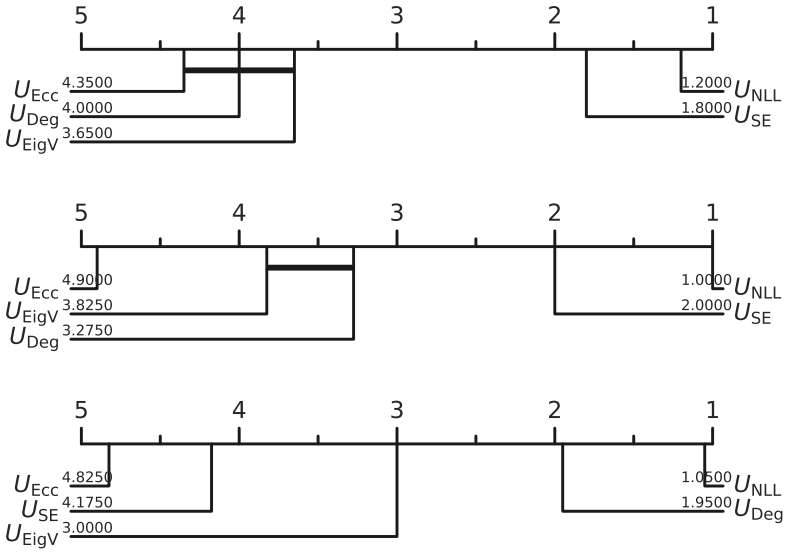


Figure 16: CD diagrams on using GPT-3.5 TriviaQA with temperature 0.5, 1.0, and 1.5.

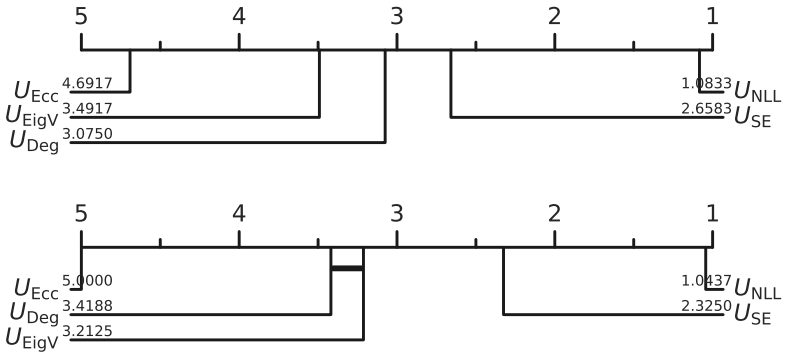


Figure 17: Conclusive comparison via critical difference diagrams. The first plot is with GPT-3.5-turbo on TriviaQA with temperatures 0.5, 1.0, and 1.5; the second is with Llama-2-chat on TriviaQA with temperatures 0.6 and 1.0.