Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph

Roman Vashurin[1♢] **Ekaterina Fadeeva**[2♢] **Artem Vazhentsev**[3♢]
Lyudmila Rvanova[6,7] **Daniil Vasilev**[4] **Akim Tsvigun**[5] **Sergey Petrakov**[3]
Rui Xing[1,8] **Abdelrahman Sadallah**[1] **Kirill Grishchenkov**[9] **Alexander Panchenko**[3]
Timothy Baldwin[1,8] **Preslav Nakov**[1] **Maxim Panov**[1] **Artem Shelmanov**[1]

[1]MBZUAI, UAE [2]ETH Zurich, Switzerland [3]Center for Artificial Intelligence Technology, Russia
[4]HSE University, Russia [5]Nebius [6]Laboratory for Analysis and Controllable Text Generation Technologies RAS, Russia [7]Weakly-Supervised NLP Group, Russia [8]The University of Melbourne, Australia [9]Independent Researcher

{roman.vashurin, Rui.Xing, Abdelrahman.Sadallah, Timothy.Baldwin, Preslav.Nakov, maxim.panov, artem.shelmanov}@mbzuai.ac.ae
ekaterina.fadeeva@inf.ethz.ch {artiomvazh99, Milarv99, aktsvigun, sergeypetrakof, kirillgrish, panchenkoalexander}@gmail.com
davasilev_4@edu.hse.ru {vazhentsev, panchenko}@airi.net

Abstract

The rapid proliferation of large language models (LLMs) has stimulated researchers to seek effective and efficient approaches to deal with LLM hallucinations and low-quality outputs. Uncertainty quantification (UQ) is a key element of machine learning applications in dealing with such challenges. However, research to date on UQ for LLMs has been fragmented in terms of techniques and evaluation methodologies. In this work, we address this issue by introducing a novel benchmark that implements a collection of state-of-the-art UQ baselines and offers an environment for controllable and consistent evaluation of novel UQ techniques over various text generation tasks. Our benchmark also supports the assessment of confidence normalization methods in terms of their ability to provide interpretable scores. Using our benchmark, we conduct a large-scale empirical investigation of UQ and normalization techniques across eleven tasks, identifying the most effective approaches.

1 Introduction

Uncertainty quantification (UQ) is increasingly being recognized as a critical safety component in AI applications. It enables systems to abstain from uncertain model predictions, allowing the associated inputs to be handled through alternative means—for example, by escalating them to a human operator (El-Yaniv et al., 2010). This safety mechanism is crucial in areas where the cost of errors is high, such as healthcare. Besides that, uncertainty scores can be used for out-of-distribution detection (OOD) (Podolskiy et al., 2021; Vazhentsev et al., 2023b), annotation with active learning (Gal et al., 2017; Shelmanov et al., 2021a; Tsvigun et al., 2022; Rubashevskii et al., 2023), adversarial attack detection (Smith and Gal, 2018), reducing model response latency (Xin et al., 2020; Schwartz et al., 2020; Schuster et al., 2022; Leviathan et al., 2023; Chen et al., 2023), among many other applications.

A plethora of UQ methods has been developed for classification and regression models (Gal, 2016). There has also been a surge of research devoted to UQ specifically in the context of encoder-only language models (LMs) such as BERT (Zhang et al., 2019; He et al., 2020; Shelmanov et al., 2021b; Xin et al., 2021; Vazhentsev et al., 2022; Kotelevskii et al., 2022; Wang et al., 2022; Kuzmin et al., 2023). The rapid proliferation of large language models (LLMs) has stimulated researchers to seek efficient and effective approaches to UQ in text generation tasks, in an attempt to make LLMs safer to use in downstream applications. As with any machine language (ML) model, LLMs can make incorrect predictions, ‘‘hallucinate’’ by fabricating claims (Xiao and Wang, 2021; Dziri et al., 2022), or simply generate low-quality outputs. These problems stem from the peculiarities of the LLM training objective, the general nature of ML models in being susceptible to errors due to the limited amount

♢ Equal contribution.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 220–248, 2025. https://doi.org/10.1162/tacl_a_00737
Action Editor: Hinrich Schütze. Submission batch: 6/2024; Revision batch: 10/2024; Published 3/2025.

of training data, and the inherent ambiguity of tasks.

Several methods exist for censoring the outputs of LLMs: output filtering using stop-word lists, post-processing using classifiers (Xu et al., 2023), fact-checking with external tools (Wang et al., 2023), output rewriting (Logacheva et al., 2022), and model alignment via preference optimization (Rafailov et al., 2024). However, these techniques alone are insufficient to entirely eliminate incorrect/inappropriate outputs. For instance, fact-checkers target a very narrow sub-problem and usually require external knowledge sources such as knowledge bases, which are generally incomplete. Building an efficient external system to verify the LLM output for every possible task is infeasible.

UQ offers a more general solution to the problem by relying on the model's internal capabilities without requiring access to external knowledge, which also enables the potential for greater computational efficiency. Several recent studies have focused on developing UQ methods for LLMs in text generation tasks (Malinin and Gales, 2021; van der Poel et al., 2022; Kuhn et al., 2023; Ren et al., 2023; Vazhentsev et al., 2023b; Fadeeva et al., 2023; Lin et al., 2024; Fadeeva et al., 2024). However, the current UQ research landscape is quite fragmented, with many non-comparable and concurrent studies. Researchers have proposed highly divergent methods for benchmarking UQ techniques, making it challenging to consolidate research findings and draw general conclusions.

In this work, we strive to bridge these disparate research efforts and resolve some issues found in their evaluation protocols by developing a benchmark for UQ techniques in text generation tasks. The benchmark is based on the LM-Polygraph framework (Fadeeva et al., 2023), which implements state-of-the-art UQ baselines in a unified way, enabling a large-scale, consistent comparison of methods developed in recent work. It includes the tasks of selective question-answering (QA), selective generation (machine translation [MT] and text summarization [TS]), and claim-level fact-checking. For the latter, we developed an automatic fact-checking pipeline for four languages: English, Chinese, Arabic, and Russian. Besides common metrics related to UQ performance, we also introduce a metric related to the calibration of confidence scores. It enables the evaluation of confidence normalization methods according to their ability to produce interpretable scores. We propose a strong baseline for normalization and investigate its performance in comparison to simpler approaches. Using the developed benchmark, we conduct a large-scale empirical investigation of UQ and normalization methods across eleven datasets.

This work both lowers the barrier to entry into UQ research for individual researchers and developers, and enables more robust, reliable, and trustworthy LLM deployment for end users.

Our **contributions** are as follows:

- We propose a new comprehensive benchmark for the evaluation of UQ and uncertainty normalization methods for LLMs. The benchmark can assess the calibration of uncertainty scores and their effectiveness in selective QA/generation and claim-level fact-checking (hallucination detection).[1]
- As part of the benchmark, we develop a novel multilingual automatic evaluation pipeline for claim-level UQ methods, focusing on claim-level fact-checking of LLM outputs in multiple languages, including English, Mandarin Chinese, Arabic, and Russian.
- We develop methods for producing normalized and bounded confidence scores that preserve the performance of raw uncertainty scores while providing better calibration and improved interpretability for end users.
- Using the developed benchmark, we perform a large-scale empirical evaluation of state-of-the-art UQ techniques.

2 Uncertainty Quantification Methods

2.1 Background

Uncertainty is a fundamental concept in ML and statistics, indicating that model predictions have a degree of variability due to the lack of complete information. Estimating predictive uncertainty is crucial for various tasks, such as selective classification, where the model abstains from making a prediction if its confidence is insufficient.

Despite recent efforts to establish a common definition of predictive uncertainty (Kotelevskii and Panov, 2024; Hofman et al., 2024), multiple approaches to its quantification exist based on

[1] All code is published under the MIT license and available at `https://github.com/IINemo/lm-polygraph`.

Type	Uncertainty Quantification Method	Category	Compute	Memory	Needs Training Data	Level
White-box	Maximum Sequence Probability (MSP)	Information-based	Low	Low	No	Seq./claim
	Perplexity (Fomicheva et al., 2020)		Low	Low	No	Seq./claim
	Mean/Max Token Entropy (TE; Fomicheva et al., 2020)		Low	Low	No	Seq./claim
	Pointwise Mutual Information (PMI; Takayama and Arase, 2019)		Medium	Low	No	Seq./claim
	Conditional PMI (van der Poel et al., 2022)		Medium	Medium	No	Seq.
	Rényi Divergence (Darrin et al., 2023)		Low	Low	No	Seq.
	Fisher-Rao Distance (Darrin et al., 2023)		Low	Low	No	Seq.
	TokenSAR (Duan et al., 2024)		Low	Low	No	Seq.
	CCP (Fadeeva et al., 2024)		Low	Low	No	Seq./claim
	Monte Carlo Sequence Entropy (MC-SE; Kuhn et al., 2023)	Sample diversity	High	Low	No	Seq.
	Monte Carlo Norm. Seq. Entropy (MC-NSE; Malinin and Gales, 2021)		High	Low	No	Seq.
	Semantic Entropy (Kuhn et al., 2023)		High	Low	No	Seq.
	SentenceSAR (Duan et al., 2024)		High	Low	No	Seq.
	SAR (Duan et al., 2024)		High	Low	No	Seq.
	Mahalanobis Distance (MD; Lee et al., 2018)	Density-based	Low	Low	Yes	Seq.
	Robust Density Estimation (RDE; Yoo et al., 2022)		Low	Low	Yes	Seq.
	Relative Mahalanobis Distance (RMD; Ren et al., 2023)		Low	Low	Yes	Seq.
	Hybrid Uncertainty Quantification (HUQ; Vazhentsev et al., 2023a)		Low	Low	Yes	Seq.
	P(True) (Kadavath et al., 2022)	Reflexive	Medium	Low	No	Seq./claim
Black-box	Number of Semantic Sets (NumSet; Lin et al., 2024)	Sample diversity	High	Low	No	Seq.
	Sum of Eigenvalues of the Graph Laplacian (EigV; Lin et al., 2024)		High	Low	No	Seq.
	Degree Matrix (Deg; Lin et al., 2024)		High	Low	No	Seq.
	Eccentricity (Ecc; Lin et al., 2024)		High	Low	No	Seq.
	Lexical Similarity (LexSim; Fomicheva et al., 2020)		High	Low	No	Seq.
	BB Semantic Entropy		High	Low	No	Seq.
	LabelProb	Information-based	Low	Low	No	Seq.
	BB P(True)	Reflexive	Medium	Low	No	Seq./claim
	Verbalized 1S (Tian et al., 2023)		Low	Low	No	Seq.
	Verbalized 2S (Tian et al., 2023)		Medium	Low	No	Seq.

Table 1: UQ methods implemented in the benchmark.

probabilities, entropies, distances, risks, etc. From a practical perspective, any of these scores could serve as a measure of uncertainty as long as they accurately reflect the relevant properties and help to solve reliability tasks.

While there are principled ways of expressing and reasoning about uncertainty, e.g., in terms of information theory and Bayesian modeling (Blundell et al., 2015), they are often difficult to implement and may lead to worse model performance. Therefore, UQ practitioners usually rely on approximations or even heuristics. For example, one popular approach to UQ is ensembling (Ashukha et al., 2019). For classification tasks, it is considered a very strong baseline, but it introduces large computational overhead due to the need for repetitive inference and storing multiple versions of weights. One of the main research questions related to UQ that has been addressed in recent work is how to perform it efficiently while keeping the performance of the uncertainty scores reliably high (Shelmanov et al., 2021b).

UQ for text generation tasks represents a greater challenge than classification. In generation, a model makes multiple predictions: one for each token. Therefore, the uncertainty scores for each token should be somehow aggregated into a single value. At the same time, in many cases, we would like to have an uncertainty score not for the entire output but for text fragments such as individual claims. Another problem is that the raw probability distributions of LLMs reflect multiple types of uncertainty, some of which might be irrelevant to a given generation task. Usually, we should not take into account the uncertainty related to the choice of the surface forms of the answer, as long as they convey the same meaning (Kuhn et al., 2023). Similarly, uncertainty related to the type of conveyed information might be irrelevant, as long as it is correct, and we care only about its veracity (Fadeeva et al., 2024). Finally, LLM predictions are not conditionally independent (Zhang et al., 2023), and therefore incorrect claims generated by an LLM at the start of an output can cause flown-on hallucinations through the subsequent generation process.

2.2 Overview of Uncertainty Quantification Methods for LLMs

Here, we provide an overview of the UQ methods implemented in our benchmark, as outlined in Table 1. The methods are implemented using the LM-Polygraph framework (Fadeeva et al., 2023), which has been extended to incorporate several recently proposed approaches. A detailed

description of the methods can be found in Appendix A.

There are two major types of techniques: white-box and black-box. *White-box* methods require access to logits, internal layer outputs, or the LLM itself. *Black-box* methods only need access to the generated text, and can easily be integrated with third-party online services such as OpenAI's API. Methods also differ in their computational requirements: Some pose high computational or memory overheads, e.g., due to repeated inference, making them less suitable for practical usage. The application of some methods can also be hindered by the need to access the model's training data. Finally, different methods might be restricted only to the UQ of the whole text (sequence level), while others might also be applicable to text fragments, such as atomic claims (claim level).

2.2.1 White-Box Methods

Let us consider the input sequence $\mathbf{x}$ and the output sequence $\mathbf{y} \in \mathcal{Y}$ of length L, where $\mathcal{Y}$ is the set of all possible output sequences. Then the probability of an output sequence given an input sequence for autoregressive language models is given by

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{l=1}^{L} P(y_l \mid \mathbf{y}_{<l}, \mathbf{x}), \qquad (1)$$

where the distribution of each y_l is conditioned on all previous tokens in a sequence $\mathbf{y}_{<l} = \{y_1, \ldots, y_{l-1}\}$.

We begin the discussion with **information-based methods**, which focus on analyzing token probability distributions $P(y_l \mid \mathbf{y}_{<l}, \mathbf{x})$. The simplest UQ baseline in this group is *Maximum Sequence Probability* (MSP) score: $U_{\text{MSP}}(\mathbf{x}) = 1 - P(\mathbf{y} \mid \mathbf{x})$. MSP discards a lot of information from the LLM probability distribution, which in theory might affect the UQ performance. This issue is addressed in various *entropy*-based techniques (Fomicheva et al., 2020). *Claim-Conditioned Probability* (CCP; Fadeeva et al., 2024) is another method in this category that aims to eliminate the impact of irrelevant sources of uncertainty reflected in original $P(y_l \mid \mathbf{y}_{<l}, \mathbf{x})$. We will return to the discussion of CCP in detail in Section 2.2.3. Information-based methods offer the advantage of being simple to implement and cost-effective, while still providing performance that is often on par with more computationally demanding techniques.

Information-based methods can be improved by sampling multiple outputs from the LLM and aggregating their confidence scores or assessing their diversity. We refer to these techniques as **sample diversity** methods. Malinin and Gales (2021) suggest to sample several sequences $\mathbf{y}^{(k)}, k = 1, \ldots, K$ and compute entropy on the sequence level through approximate Monte Carlo estimation (*Monte Carlo sequence entropy*). This approach does not take into account that many sampled responses do not diverge in meaning and only vary in surface form. This problem is addressed by *Semantic Entropy* (SE; Kuhn et al., 2023), which clusters sampled responses by meaning and computes entropy of the distribution over obtained clusters. This approach is further extended in *Shifting Attention to Relevance* (SAR; Duan et al., 2024). Instead of clustering, SAR performs a soft aggregation of word or sentence probabilities using their semantic similarity. Additionally, SAR mitigates the influence of irrelevant tokens and sequence samples.

Density-based methods (Lee et al., 2018; Yoo et al., 2022; Kotelevskii et al., 2022; Ren et al., 2023) approximate the training data distribution using embeddings of training instances. Usually, this distribution is modeled by one or multiple Gaussians. Uncertainty is quantified by estimating the likelihood of the input under the approximated distribution. As such, they are good at spotting OOD instances (Vazhentsev et al., 2023b). They are computationally efficient at inference time, with no additional inference steps required. However, they require access to the model's training data to fit the approximated distribution. One can also combine information-based and density-based methods as suggested by Vazhentsev et al. (2023a) and Ren et al. (2023). For example, the *Hybrid Uncertainty Quantification* (HUQ) method (Vazhentsev et al., 2023a) performs a ranking-based aggregation and leverages the strengths of both information-based and density-based methods.

Directly asking the model to provide confidence for its responses is another approach to UQ (Kadavath et al., 2022; Tian et al., 2023). We refer to such techniques as **reflexive**. In one of the simplest techniques of this kind P(True) (Kadavath et al., 2022), the authors generate a response from an LLM and subsequently ask the same LLM to verify the answer. The uncertainty score is

calculated using the probability of the option ‘‘True’’ in the output distribution of the LLM for the second prompt. Kadavath et al. (2022) showed that this method achieves better performance than MSP at the cost of an additional inference step on a variety of tasks for large LLMs ($\geq$70b parameters).

2.2.2 Black-Box Methods

LLM providers often expose models in a black-box fashion, where users have access to generated text only. Among **sample diversity** black-box methods (also known as consistency-based methods in the black-box setting Zhang et al., 2024), we consider Lexical Similarity (Fomicheva et al., 2020), Number of Semantic Sets, Sum of Eigenvalues of the Graph Laplacian, Degree Matrix, and Eccentricity. These methods are grounded in a common methodological framework (Lin et al., 2024):

- Obtain K responses $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}$ for a particular input $\mathbf{x}$.
- Compute $K \times K$ similarity matrix S between responses, where $S_{ij} = s\big(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}\big)$ for some similarity score s (e.g., NLI or Jaccard score).
- Analyze the similarity matrix S and aggregate the information in this matrix to compute the resulting uncertainty score.

Number of Semantic Sets is the simplest method, which clusters semantically similar responses into non-overlapping groups and counts the resulting clusters. A larger number of clusters indicates greater uncertainty. Other methods do a more sophisticated analysis of the matrix S. For example, *Sum of Eigenvalues of the Graph Laplacian* computes the sum of eigenvalues of the normalized matrix S, providing a continuous relaxation of the Number of Semantic Sets score.

Reflexive techniques for the black-box setting are enabled by the fact that most popular LLMs deployed as a service are instruction-tuned and are able to follow a multi-turn conversation. This allows the user to either prompt a model to explicitly verbalize its confidence level as part of its response or request a confidence estimate in a follow-up conversation turn. Tian et al. (2023) propose several variations of such *Verbalized* UQ methods and conduct an extensive empirical evaluation.

2.2.3 Claim-Level Extensions

While the methods discussed above provide uncertainty scores for entire generated sequences, it is often desirable to quantify the uncertainty for short text fragments (claims) within the LLM output. Assuming that claims have been extracted from text sentences and there is a mapping between them and the tokens in the original text, we can obtain probability distributions for each token in each claim. Some aforementioned sequence-level methods can be modified to operate on the claim level (Fadeeva et al., 2024), but not all of them. For example, sampling-based methods cannot work on the claim level because sampled texts may diverge too much and miss some claims.

Let C denote a set of token indices corresponding to a claim. To adapt *MSP* to the claim level, we can compute the joint probability of tokens solely within the claim instead of the whole sequence: $\prod_{l \in C} P(y_l \mid \mathbf{y}_{<l})$. In a similar way, we can adapt *Mean/Max Token Entropy*, *Perplexity*, and *PMI*. *P(True)* could be adapted to the claim level by querying an LLM about the correctness of each claim in the generated response individually.

Claim-Conditioned Probability (CCP; Fadeeva et al., 2024) is designed specifically for the claim-level UQ (but can also be applied at the sequence level). It assesses the semantic similarity between the original claim and perturbed versions where individual tokens are replaced with alternative generations. This approach provides a more nuanced understanding of uncertainty by considering the potential impact of different word choices on the overall meaning of the claim.

3 Uncertainty Normalization Methods

Raw uncertainty scores are good for ranking outputs by their potential quality but can be confusing to the end user. To address this issue, we consider several methods for producing confidence scores bounded within the range $[0, 1]$.

All methods require fitting on a held-out calibration set $\mathcal{D}_{calib} = \{\mathbf{x}_i, \mathbf{y}_i^*\}_{i=1}^N$. We assume that for each input $\mathbf{x}_i$ in this set, output $\mathbf{y}_i$ is generated by LLM, and some uncertainty score $u_i = U(\mathbf{x}_i)$ as well as the generation quality score $q_i = Q(\mathbf{y}_i, \mathbf{y}_i^*)$ are computed.

Among simple normalization approaches, we consider *linear scaling* and *quantile scaling*, as they provide simple rules to normalize uncertainty

scores in $[0, 1]$ based solely on uncertainty values u_i; see Appendix B for more details.

To convey meaningful information about the model's confidence to the end user, the confidence score should not only be bounded within a fixed interval but also directly reflect the expected quality of the model's output. We term this *confidence calibration*, as opposed to confidence normalization. To achieve this, we introduce two methods referred to as *Performance-Calibrated Confidence* (PCC).

The first approach, *Binned PCC*, splits the calibration set into non-intersecting bins based on the values of uncertainty u_i and considers the confidence to be an estimate of the mean quality of the outputs in the bin, as measured by some quality measure of choice. The downside of this approach is that the ordering of the instances based on raw uncertainty and normalized confidence scores can be different, and thus the quality of UQ can vary substantially and unpredictably.

To address this problem, we propose a second approach: *Isotonic PCC*. It fits Centered Isotonic Regression (CIR; Oron and Flournoy, 2017) on pairs of uncertainty and quality values from the calibration set. CIR produces a monotonic piecewise linear function, which allows the use of the relationship between uncertainty and quality while keeping the order of the inputs intact.

Both approaches in the PCC family produce calibrated confidence scores as a local estimate of some quality measure in the neighborhood of the raw uncertainty estimate. This directly ties the confidence with the estimated quality of the output, thus making it more interpretable than raw uncertainty scores. We provide a more detailed discussion on the specifics of these methods in Appendix B.

4 Approaches to Evaluating Uncertainty Quantification Methods

In general, a valid UQ technique should produce scores that are well correlated with some measure of output quality. Thus, the most straightforward way of comparing different UQ methods is to measure the rank correlation between some generation quality metric (e.g., ROUGE-L) and uncertainty scores (Fomicheva et al., 2020; Ren et al., 2023). However, this way of evaluating UQ is not very informative of the performance gain that a particular UQ method achieves.

Another popular evaluation approach is based on designating outputs as either correct or incorrect based on a threshold over a quality metric, and measuring how well uncertainty scores can predict the output as being one or the other (Kuhn et al., 2023; Duan et al., 2024). This reduces the uncertainty score to being a predictor in a binary classification task, and thus one can use ROC-AUC or PR-AUC as a measure of how well a UQ method performs. The problem with this approach, which makes results across different works incomparable, is that it requires selecting the quality threshold, and its choice has been quite arbitrary in the literature.

A more comprehensive approach is called *rejection verification* (Malinin and Gales, 2021; Lin et al., 2024). It does not require thresholding the quality metric to formulate a binary classification task. Instead, it computes the average quality of the outputs for which the uncertainty is relatively low. By continuously lowering the uncertainty threshold above which data points are discarded, one obtains a set of average quality values of outputs with progressively lower maximum uncertainty. These pairs of uncertainty thresholds and associated average output quality give a prediction–rejection curve. The area under this curve quantifies the overall quality of an UQ method (Malinin et al., 2017).

One problem with the majority of evaluation approaches in previous work on UQ for LLMs is the usage of n-gram output quality metrics such as ROUGE-L, which often do not reflect the actual quality of the generated output. For example, this discrepancy occurs when the gold standard answer and the LLM output differ only by a negation token. In this case, n-gram metrics would rank such a model answer higher than it deserves, failing to capture the substantial semantic difference caused by the negation. In addition to n-gram metrics, we suggest using the recently proposed AlignScore (Zha et al., 2023), where the gold-standard answer and LLM output are compared by another LLM. This metric has a higher correlation with human annotators due to its ability to capture deeper semantic similarities and differences between texts.

Another issue with evaluation protocols in recent works is their tendency to overlook simple yet effective baselines. For instance, MSP often proves difficult to surpass in tasks with short outputs. Many studies neglect this baseline,

favoring comparisons with entropy-based metrics (He et al., 2020; Xiao and Wang, 2021; Kuhn et al., 2023), which often perform worse. By not considering this straightforward baseline, the evaluation protocols may give an incomplete picture of UQ performance.

We also note that research on UQ for text fragments, such as sentences and claims, has been limited. Furthermore, the evaluation protocols for this setting exhibit significant limitations. For example, Manakul et al. (2023) manually annotated texts generated by one LLM, then evaluated the UQ performance for a proxy model by inferring the probability distributions of the tokens for the fixed output. We argue that such an approach introduces a big discrepancy between the generated text and what a proxy LLM actually wants to generate, which results in biased UQ performance. In this work, we mitigate this problem by building an automatic evaluation pipeline for UQ in claim-level fact-checking for various languages. It allows unrestricted generation from LLMs and leverages LLM-as-a-judge (Zheng et al., 2023) instead of manual annotation to support experiments with various LLMs.

5 Evaluation Benchmark

Our benchmark features three sections: (1) evaluation of UQ performance in selective QA/generation; (2) evaluation of UQ performance in fact-checking; (3) evaluation of confidence calibration.

5.1 Selective QA / Generation

In this section of the benchmark, we evaluate how well the uncertainty scores of the considered methods detect low-quality LLM generations.

Datasets. For selective QA, we use four datasets: CoQA (Reddy et al., 2019) with free-form answers about conversations; TriviaQA (Joshi et al., 2017) with complex and compositional questions without context; MMLU (Hendrycks et al., 2021), a multitask dataset structured as multiple-choice QA; and GSM8k (Cobbe et al., 2021), consisting of grade school math word problems. The first two datasets have been used widely in UQ research, while the last two are popular datasets for evaluating the English-language generation quality of LLMs. For selective generation, we use two MT datasets: WMT-14 French to English (Bojar et al., 2014) and WMT-19 German to English (Barrault et al., 2019), and one text summarization dataset: XSum (Narayan et al., 2018).

Dataset	Type	Num. Instances train/test	Avg. Document Len.	Avg. Target	Language
CoQA	QA	7,199 / 500	405.9	4	English
TriviaQA	QA	138,384 / 17,210	18.8	4.3	English
MMLU	QA	99,842 / 14,042	64.9	3	English
GSM8k	QA	7,473 / 1,319	64.2	128.6	English
XSum	ATS	204,045 / 11,334	544.2	30.4	English
WMT'14	NMT	40.8M / 3,003	49.3	32.9	Fr.-to-Eng.
WMT'19	NMT	34.8M / 2,998	52.5	33.5	Ger.-to-Eng.

Table 2: The statistics of the benchmark datasets. The lengths in tokens are provided according to the Mistral 7B v0.2 tokenizer.

For each dataset, we limit the evaluation set to 2,000 instances, except for MMLU, where we restrict the number of questions to 100 per subject. We also reserve 1,000 instances from the training set for UQ techniques that require ''pre-training'', such as density-based methods. The detailed statistics about the datasets are given in Table 2.

We primarily use prompt formats from the lm-evaluation-harness framework (Gao et al., 2023). For TriviaQA, MMLU, and GSM8k, we use a 5-shot prompt. For XSum, WMT-14 Fr-En, and WMT-19 De-En, we use a zero-shot prompt. For CoQA, we use a few-shot prompt with all preceding questions in the conversation before the target question. The maximum length of the generated sequence in selective QA and generation tasks was set to the 99th percentile of the target sequence length in the respective training set.

The datasets of the benchmark could be found in the HuggingFace repository.[2]

Metrics. Following previous work on UQ in text generation (Malinin and Gales, 2021; Vazhentsev et al., 2022), we compare the methods using the Prediction Rejection Ratio (PRR) metric (Malinin et al., 2017). Consider a test dataset $\mathcal{D} = \{(\mathbf{x}_j, \mathbf{y}_j^*)\}$. Let $\mathbf{y}_j$ be the output generated by an LLM for an input $\mathbf{x}_j$ and $u_j = U(\mathbf{x}_j)$ be the uncertainty score of a prediction. The rejection curve indicates how the average quality $Q(\mathbf{y}_j, \mathbf{y}_j^*)$ of the instances with uncertainty $u_j < a$ depends on the value of the rejection parameter a. PRR computes the ratio of the area between the rejection curves for a considered uncertainty score and a random score and the area between the

[2] https://huggingface.co/LM-Polygraph.

oracle (the best possible uncertainty that sorts instances according to their text quality metric) and a random score:

$$PRR = \frac{\text{AUC}_{\text{unc}} - \text{AUC}_{\text{rnd}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{rnd}}}. \quad (2)$$

A higher PRR indicates a better uncertainty score.

The choice of the generation quality measure $Q(\mathbf{y}_j, \mathbf{y}_j^*)$ depends on the dataset. In contrast to previous work that employs n-gram-based measures, our benchmark primarily relies on LLM-based metrics, such as AlignScore (Zha et al., 2023) and COMET (Rei et al., 2020), while retaining n-gram measures for comparability with prior research. For MT datasets, we use COMET and AlignScore. For GSM8K and MMLU, we use accuracy. For CoQA and TriviaQA, we use AlignScore. For XSum, we use ROUGE-L and AlignScore.

When calculating AlignScore for QA outputs, we treat the model's generations as context and a ground truth response as the claim. This approach is designed to accurately score instances where the model mentions the correct entity in its response but includes an additional explanation or rephrases the correct answer differently from the ground truth, which would otherwise lead to failure in exact match scoring. This is especially important in evaluation on the CoQA dataset.

5.2 Claim-Level Fact-Checking

In this part of the benchmark, we evaluate claim-level UQ techniques and their ability to spot hallucinations on the task of generating biographies, as proposed by Fadeeva et al. (2024). The difference over previous work such as Manakul et al. (2023) is that in our benchmark, we perform unrestricted generation using the LLM, which is much closer to the practical use case. However, this raises the problem that each generation is unique and needs to be reannotated, posing a significant challenge for human annotation.

Evaluation Pipeline. We implement an automatic benchmarking pipeline to enable the evaluation of UQ methods on unrestricted LLM outputs. The pipeline supports English, Mandarin Chinese, Arabic, and Russian. It is intended to work with any modern LLM that has functionality over these languages. The main feature of our pipeline is its full automation without needing human labeling at any step. The automation is achieved via extensive use of GPT-4. Note that GPT-4 is used only as a tool for benchmarking, while UQ is performed solely based on the particular LLM in question.

Language	Acc.	F1 score	# claims	% of false claims	Fleiss Kappa
English, Mistral-7B-v0.1	0.98	0.93	97	16.5%	0.90
Arabic, Jais 13B	0.89	0.80	132	28.3%	0.86
Russian, Vikhr 7B-0.2	0.89	0.80	275	15.6%	0.85
Chinese, Yi 7B	0.89	0.89	100	35.0%	0.87

Table 3: Classification metrics of GPT-4 annotation against manual annotation in claim-level fact-checking (unsupported claims represent a positive class) and annotation agreement (Fleiss Kappa) using True/False labels from 3 annotators.

Using a LLM, e.g., Mistral 7b, we first generate responses to biography prompts such as *Give me a biography of <person name>*. The set of people was generated by asking GPT-4 to list the most famous people since 1900. The maximum output length is 256 tokens. The generated texts are then decomposed into atomic claims, with each claim mapped to a corresponding subset of tokens from the original LLM output. The decomposition and mapping are done using GPT-4 with a language-specific prompt. Usually, about 5% of claims cannot be mapped to tokens because GPT-4 abstains from responding or outputs words not present in the original text. The further evaluation considers only successfully matched claims.

The annotation of the extracted claims is also done automatically using GPT-4. We use language-specific prompts that facilitate chain-of-thought reasoning to ask whether the presented claims are supported, unsupported, or unknown. Usually, the percentage of claims classified as unknown is negligible; these claims are discarded from the evaluation. To assess the quality of automatic annotation, we manually annotated a random subset of claims for each language.[3] Table 3 summarizes the binary performance metrics of GPT-4 against human labels and presents annotation agreement scores. The results indicate that GPT-4 is a reliable evaluator, capable of serving as a ''ground truth'' for assessing UQ techniques.

[3]Human annotations are available at https://huggingface.co/datasets/LM-Polygraph/bio-claim-human-anno.

Metrics. The performance of the claim-level UQ methods is evaluated using ROC-AUC and PR-AUC with unsupported claims as a positive class.

5.3 Effect of Uncertainty Normalization

This section of the benchmark is designed to analyze how uncertainty normalization procedures (as presented in Section 3) affect the performance of the scores and their correlation with the quality of the generated text.

Datasets. The reserved training partitions of the datasets from the selective QA and generation section are used as calibration sets for normalization methods. Evaluation can be performed either on the concatenated test partitions or on each of them individually.

Metrics. The benchmark offers two metrics: PRR before and after normalization to ensure that performance does not degrade; and a metric similar to ECE that measures ‘‘calibration’’, i.e., the ability of normalized uncertainty scores to represent the expected quality of the output in a bounded range—Mean Squared Error (MSE) between a normalized quality metric and a confidence score. Lower MSE indicates better confidence ‘‘calibration’’.

5.4 Models

For selective QA / generation, we conducted experiments with white-box models that provide access to logits and their internal states. For selective QA, we also conducted experiments with black-box models that provide only the generated text. Black-box models represent the scenario where LLMs are deployed as services and are available only via an API (e.g., ChatGPT and models deployed on platforms like HuggingFace). For selective QA/generation in the white-box setting, we use Mistral 7B v0.2 (Jiang et al., 2023) and Stable LM 2 12B (Bellagente et al., 2024) base models without instruction tuning. For black-box evaluation on selective QA, we use the corresponding instruction-tuned versions of these models and also GPT-4o-mini. The experiments on claim-level fact-checking are conducted using the instruction-tuned versions of Mistral 7B-v0.1 (Jiang et al., 2023) (for English), Jais 13B (Sengupta et al., 2023) (for English and Arabic), Vikhr 7B-0.2 (Nikolich et al., 2024) (for Russian), and Yi 7B (for Chinese). The detailed generation hyperparameters can be found in the code base. The text generation quality of the models is presented in Table 14 in Appendix E.

For white-box models, we use continuation-style prompts. For black-box models with verbalized UQ techniques, we use prompts specified by Tian et al. (2023), and for other black-box UQ methods, we use general instruction-oriented prompts. For both model types, we compute the same metrics. However, for black-box models with verbalized UQ methods, we perform more extensive output post-processing to extract the model's predictions and disentangle them from reported confidence.

6 Experiments

Using our benchmark, we evaluated the implemented UQ and normalization methods.

6.1 Selective QA and Generation

Selective QA. Tables 6 and 7 in Appendix C present detailed results on the selective QA task for white-box models. Figures 2 and 1 present the aggregated results.

Despite being a simplistic baseline, MSP demonstrates strong performance. On MMLU, it is the best method for Stable LM and the second best for Mistral. It achieves the second-best result on CoQA for Mistral and outperforms entropy-based methods in most cases. For GSM8k, MSP substantially lags behind the best techniques, but still outperforms entropy-based methods for Mistral.

Among information-based methods, it is also worth noting CCP as it demonstrates the best performance on MMLU and GSM8k for Mistral and has the second best result on TriviaQA and MMLU for Stable LM.

The majority of density-based methods demonstrate poor results across all tasks. One exception is HUQ-MD, which is a hybrid method that leverages strengths from both information- and density-based approaches. HUQ-MD delivers strong performance across all tasks and LLMs, achieving the best results on GSM8k for Stable LM with a substantial improvement over the nearest competitor.

State-of-the-art methods based on sample diversity typically perform well on CoQA and TriviaQA. The black-box method based on the degree matrix, DegMat with NLI similarity metric, achieves the best results for both LLMs. However,

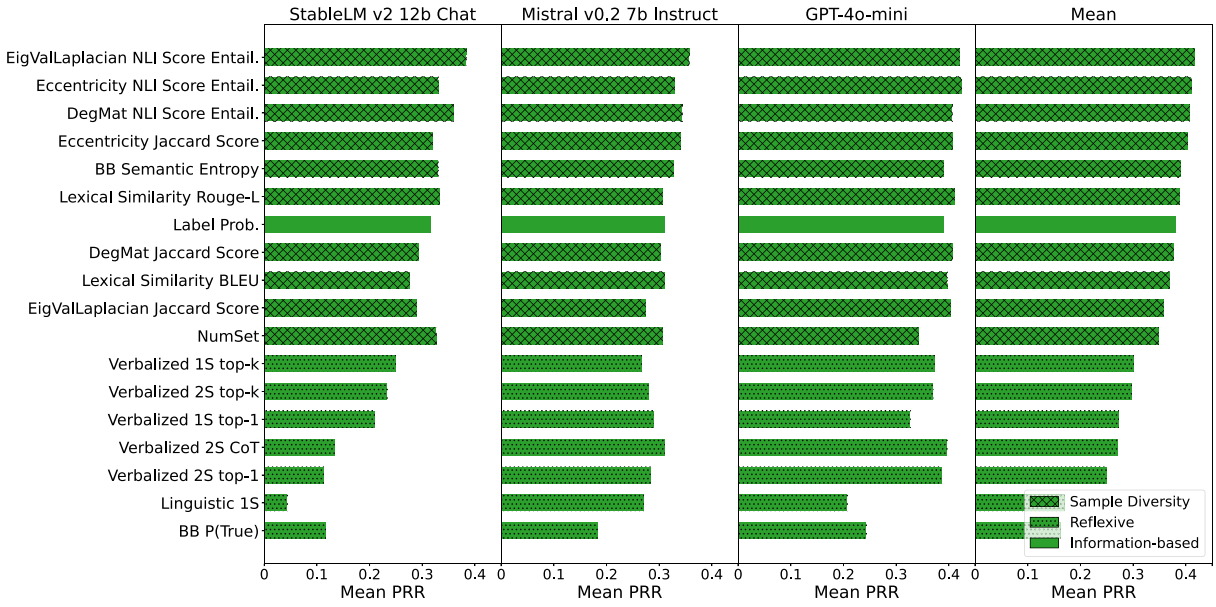

Figure 1: Mean PRR ↑ aggregated over all selective QA tasks for each black-box LLM separately (the lower the better). Column *Mean* corresponds to the mean PRR across all LLMs.

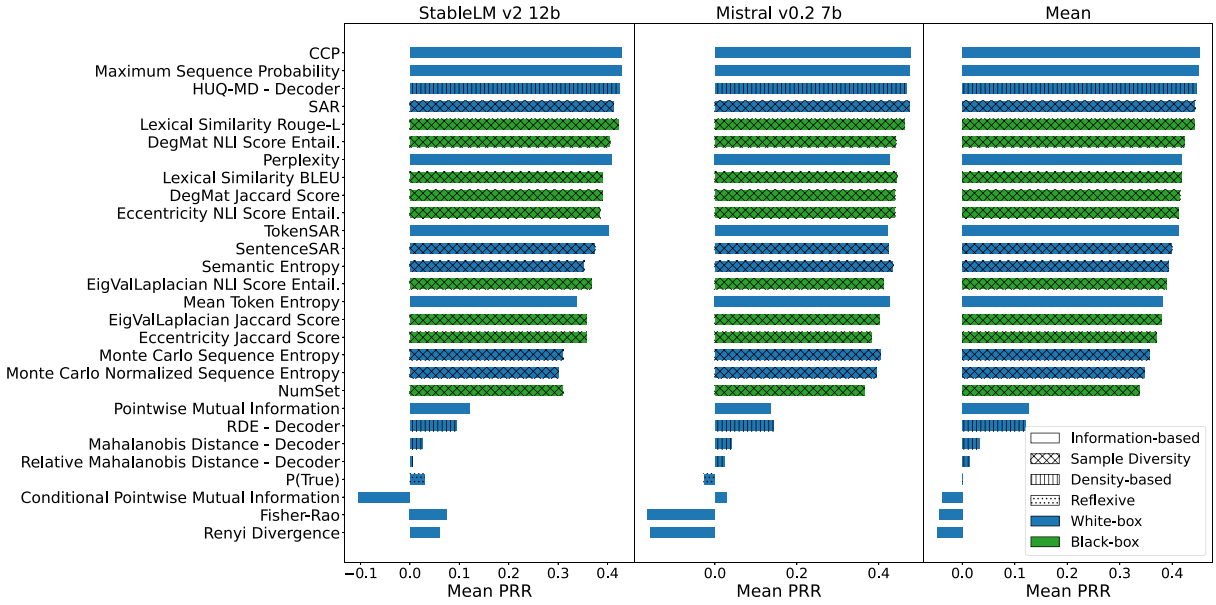

Figure 2: Mean PRR ↑ aggregated over all selective QA tasks for each white-box LLM separately (the lower the better). Column *Mean* corresponds to the mean PRR across all LLMs.

for MMLU, these methods fall significantly behind information-based techniques. This can be attributed to the nature of the multiple-choice QA task in MMLU, where LLMs are constrained to choose from a limited set of options and generate very short responses, limiting their capacity to exploit sample diversity. For GSM8k, the results are mixed: Methods based on sample diversity perform similarly to information-based approaches. In the sample diversity group, it is worth also highlighting SAR, which demonstrates strong overall performance, achieving the second-best result on GSM8k and the best result on TriviaQA for Mistral.

The reflexive method P(True) in most cases is not better than random. This could be due to the used LLMs being too small to develop an awareness of their own knowledge gaps, a capability observed in larger LLMs (Kadavath et al., 2022).

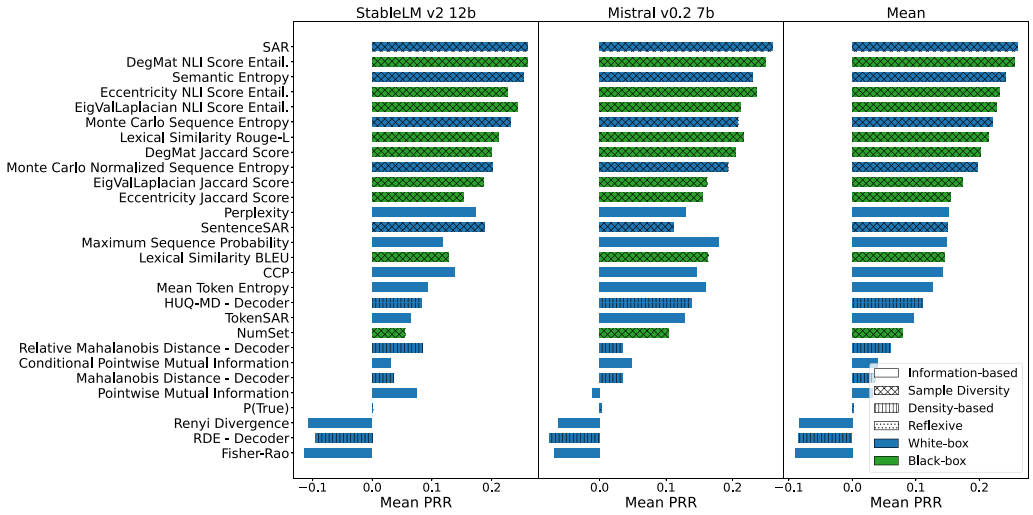

Figure 3: The mean PRR ↑ aggregated over all selective generation tasks for each LLM separately (the lower the better). Column *Mean* corresponds to the mean PRR across all LLMs.

Figure 1 presents the mean PRRs for each white-box model individually, as well as the mean PRR across both models. The best methods overall in selective QA for white-box models are CCP, MSP, HUQ-MD, SAR, and Lexical Similarity.

Tables 8 to 10 in Appendix C present evaluation results for the instruction-tuned LLMs treated as black-box models. For all LLMs, the pattern is similar. On the CoQA and TriviaQA datasets, empirical information-based and sample diversity methods confidently outperform reflexive techniques: P(True) and verbalized uncertainty. On MMLU, this pattern is reversed: Verbalized UQ methods notably outperform other techniques. For GPT-4o-mini, improvements are consistent across the majority of verbalized uncertainty methods. For Stable LM and Mistral, they are still the best, though many of them perform comparably to sample diversity techniques.

Averaging results across all datasets and models (see Figure 2) still places sample diversity-based methods at the top. Three best techniques in this experiment—EigValLaplacian, DegMat, and Eccentricity—use the NLI-based similarity measure for capturing the semantic diversity of responses.

UQ Method	Output length in symbols						
	1–2	3–6	7–24	25–85	86–138	139–210	211–1k
MSP	0.71	0.62	0.67	0.57	0.54	0.38	0.26
Perplexity	0.71	0.59	0.71	0.56	0.41	0.27	0.26
Mean Token Entropy	0.68	0.30	0.54	0.47	0.39	0.22	0.17
PMI	0.44	0.19	0.11	−0.24	−0.28	−0.17	0.0
Conditional PMI	0.26	0.04	−0.21	−0.12	−0.12	−0.11	−0.12
Rényi Divergence	0.05	0.05	−0.54	0.18	0.21	−0.06	−0.15
Fisher-Rao Distance	0.08	0.06	−0.31	−0.07	0.09	0.21	0.23
TokenSAR	0.70	0.59	0.70	0.56	0.40	0.26	0.26
CCP	0.59	0.63	0.62	0.54	0.56	0.50	0.36
MC−SE	0.62	0.39	0.57	0.54	0.68	0.60	0.33
MC−NSE	0.62	0.45	0.67	0.54	0.57	0.52	0.36
Semantic Entropy	0.64	0.45	0.66	0.63	0.72	0.68	0.44
SentenceSAR	0.68	0.57	0.71	0.53	0.28	0.22	0.23
SAR	0.67	0.55	0.73	0.64	0.68	0.67	0.52
MD − Decoder	0.12	−0.11	−0.48	−0.14	−0.23	−0.18	−0.14
RDE − Decoder	−0.06	0.14	0.16	−0.11	−0.22	0.05	0.09
RMD − Decoder	−0.10	0.02	−0.05	0.14	0.12	−0.03	−0.14
HUQ-MD − Decoder	0.63	0.51	0.40	0.02	−0.03	0.01	0.10
P(True)	−0.05	−0.22	−0.51	−0.23	−0.37	−0.16	−0.03
NumSet	0.60	0.49	0.62	0.41	0.48	0.47	0.17
EigValLaplacian NLI Entail.	0.63	0.54	0.71	0.71	0.77	0.77	0.62
EigValLaplacian NLI Contra.	0.55	0.52	0.65	0.58	0.58	0.49	0.22
EigValLaplacian Jaccard	0.60	0.49	0.68	0.61	0.69	0.66	0.51
DegMat NLI Entail.	0.64	0.56	0.74	0.71	0.77	0.78	0.63
DegMat NLI Contra.	0.48	0.54	0.62	0.58	0.58	0.47	0.20
DegMat Jaccard	0.64	0.53	0.7	0.62	0.7	0.68	0.55
Eccentricity NLI Entail.	0.60	0.50	0.71	0.7	0.77	0.77	0.64
Eccentricity NLI Contra.	0.55	0.52	0.67	0.58	0.57	0.47	0.21
Eccentricity Jaccard	0.60	0.46	0.66	0.59	0.68	0.64	0.49
Lexical Similarity Rouge-L	0.64	0.55	0.71	0.64	0.72	0.71	0.57
Lexical Similarity BLEU	0.64	0.52	0.65	0.59	0.69	0.68	0.56

Table 4: PRR↑ with AlignScore aggregated over all selective QA/generation tasks and both white-box LLMs. The results are grouped by output length, with each interval representing approximately the same number of instances.

Selective Generation. Tables 11 and 12 in Appendix C present detailed results on the selective generation task for white-box LLMs: Stable LM 2 12B and Mistral 7B v0.2, respectively. For both models, on the text summarization task, the best results are achieved by sample-diversity techniques in terms of both metrics. The majority of information-based techniques have negative or near-zero PRR in most of the cases, which indicates that they perform similarly or worse than random. One exception is the performance of Conditional PMI in terms of ROUGE-L-based PRR, which is drastically better than the performance of all other methods. This discrepancy might be

UQ Method	English, Mistral 7b		English, Jais 13b		Arabic, Jais 13b		Russian, Vikhr 7b		Chinese, Yi 6b	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
MSP	0.65 ± 0.03	0.33 ± 0.01	0.65 ± 0.03	0.41 ± 0.01	0.6 ± 0.02	0.2 ± 0.01	0.59 ± 0.02	0.74 ± 0.04	0.51 ± 0.01	0.17 ± 0.01
Perplexity	0.62 ± 0.02	0.29 ± 0.01	0.62 ± 0.02	0.35 ± 0.01	0.62 ± 0.02	0.23 ± 0.01	0.48 ± 0.01	0.63 ± 0.02	0.5 ± 0.01	0.16 ± 0.01
Max Token Entropy	0.64 ± 0.03	0.33 ± 0.01	0.61 ± 0.02	0.38 ± 0.01	0.58 ± 0.02	0.24 ± 0.01	0.49 ± 0.01	0.67 ± 0.03	0.52 ± 0.01	0.19 ± 0.01
PMI	0.44 ± 0.01	0.19 ± 0.01	0.45 ± 0.01	0.23 ± 0.01	0.47 ± 0.01	0.14 ± 0.01	0.39 ± 0.01	0.61 ± 0.02	0.5 ± 0.01	0.17 ± 0.01
CCP	**0.74 ± 0.04**	**0.46 ± 0.01**	**0.68 ± 0.03**	**0.47 ± 0.01**	**0.73 ± 0.04**	**0.3 ± 0.01**	**0.67 ± 0.03**	**0.8 ± 0.04**	**0.61 ± 0.02**	**0.25 ± 0.01**
P(True)	0.62 ± 0.02	0.3 ± 0.01	0.55 ± 0.02	0.3 ± 0.01	0.53 ± 0.01	0.16 ± 0.01	0.64 ± 0.03	0.75 ± 0.04	**0.61 ± 0.02**	**0.25 ± 0.01**

Table 5: ROC-AUC↑ and PR-AUC↑ (with unsupported claims as the the positive class) on the claim-level fact-checking benchmark. Warmer colors indicate better results.

due to limitations of the ROUGE-L metric for measuring the quality of text summarization.

On MT, in terms of PRR-COMET, there is no clear winner. For Stable LM, on both datasets SAR achieves the best results, while for Mistral, basic information-based techniques are the best. However, in terms of PRR-AlignScore, the black-box techniques based on the NLI similarity are clear winners. It is worth highlighting DegMat, as in terms of PRR-AlignScore, it outperforms all techniques for all considered datasets for Stable LM and achieves the second best results for Mistral.

Figure 3 presents the mean PRR values for each white-box model individually, as well as the mean PRR across both models. The best methods are sample diversity techniques: SAR, DegMat, and Semantic Entropy. The best white-box techniques, SAR and Semantic Entropy, are closely competing with black-box methods. The MSP baseline trails behind in this task but still achieves reasonable performance.

Output Length Impact. Table 4 presents the aggregated PRR scores for the two white-box LLMs across all datasets, categorized by the length of their outputs. For short generations (<7 symbols), information-based methods (Maximum Probability, Perplexity, Mean Token Entropy, CCP, and TokenSAR) perform the best. Conversely, for longer outputs, sample diversity methods, especially black-box techniques based on NLI similarity, achieve superior performance. SAR, despite being inferior to black-box techniques for long outputs and slightly inferior to information-based techniques for short outputs, demonstrates the most robust results across all output lengths.

6.2 Fact-Checking

In the fact-checking evaluation, we generated biographies for 100 individuals across five languages: English (using Mistral 7B-v0.1 and Jais 13B), Arabic (using Jais 13B), Russian (using Vikhr 7B-0.2), and Chinese (using Yi 6B). From biographies in English generated using Mistral 7B, we have extracted 2,499 claims; from biographies in English generated by Jais 13B – 1,100 claims; in Arabic – 1,031 claims; in Russian – 3,104 claims, and in Chinese – 2,703 claims. The percentage of claims annotated as unsupported by the automatic pipeline was 20.5% for Mistral 7B-v0.1 and 16.7% for Jais 13B in English, 17.4% in Arabic, 58.6% in Russian, and 20.0% in Chinese. No more than 8% of claims were classified as unknown across all models and languages and were discarded from the evaluation. Table 5 shows the performance of UQ techniques in fact-checking obtained using the automatic evaluation pipeline. For all tested models and languages, except Chinese, CCP consistently achieves the best performance, surpassing other methods. For Chinese, CCP and P(True) yield comparable results across both evaluated metrics. This may be attributed to peculiarities of the Yi model resulting from specific fine-tuning, which enhances the model's ability to assess its own confidence.

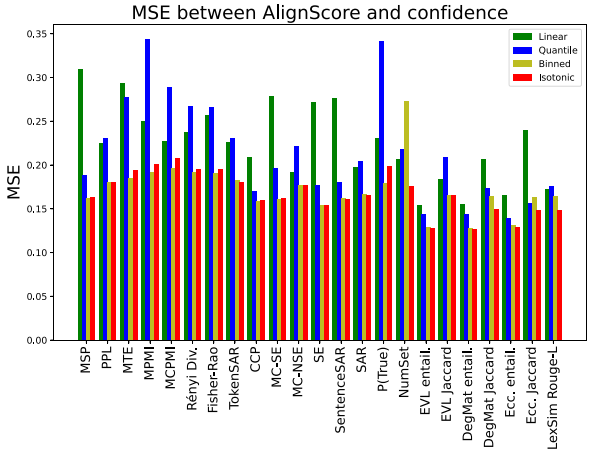

Figure 4: MSE↓ between AlignScore and confidence scores obtained by various normalization methods. Scorers are fitted on combined train partitions of all datasets, and MSE is averaged over their combined test partitions.

6.3 Effect of Uncertainty Normalization

The calibration (MSE) of confidence scores obtained by various normalization methods is presented in Figure 4. Binned and isotonic PCC confidently outperform the linear and quantile normalization for almost all of the considered UQ techniques. To verify that the quality of the normalized scores does not degrade after normalization, we also evaluate their performance in selective QA/generation tasks (see Table 13 in Appendix D). We compute PRR for raw and normalized uncertainty scores and analyze the difference. Our observations indicate that scores normalized using linear, quantile, and isotonic PCC methods perform similarly to the raw scores, which is anticipated due to their monotonic properties.

7 Conclusion

In this work, we proposed a comprehensive benchmark for evaluating UQ techniques in text generation tasks. The empirical investigation conducted on the developed benchmark provides several useful insights. Overall, methods based on sample diversity perform well across selective QA/generation tasks. However, for shorter answers, we recommend using information-based methods because they often perform on par with diversity-based techniques, while introducing much less computational overhead. Specifically, for multiple-choice QA, information-based methods MSP, Perplexity, and CCP would be substantially superior. For the tasks that assume longer outputs, methods based on sample diversity such as Semantic Entropy, DegMat, or Lexical Similarity are preferable. It is worth highlighting that SAR consistently stands out as one of the most effective methods for short and long outputs. Reflexive methods in general do not demonstrate good performance. Only for big LLMs, such as GPT-4o-mini, it might be reasonable to use verbalized techniques. For the fact-checking task, the best method is CCP: It demonstrates the best results and is computationally efficient. We should also note that MSP appears to be a very strong and robust baseline across all tasks and should not be discarded from the evaluation protocols. For generating human-interpretable confidence scores, we suggest normalization based on isotonic PCC as it improves confidence calibration and does not degrade performance in terms of PRR.

Acknowledgments

We thank anonymous reviewers for their insightful feedback towards improving this paper. This work is supported by a grant #848011 from the MBZUAI & WIS Collaborative Research Program.

References

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2019. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, Meng Lee, Emad Mostaque, Michael Pieler, Nikhil Pinnaparju, Paulo Rocha, Harry Saini, Hannah Teufel, Niccolo Zanichelli, and Carlos Riquelme. 2024. Stable LM 2 1.6b technical report. *arXiv preprint arXiv:2402.17834*.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ales Tamchyna.

2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3302

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023. RainProof: An umbrella to shield text generator from out-of-distribution data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5857, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.357

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.276

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.387

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

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.558

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461. https://doi.org/10.18653/v1/2023.emnlp-demo.41

Wade Fagen-Ulmschneider. 2023. Perception of probability words. https://waf.cs.illinois.edu/visualizations/Perception-of-Probability-Words/

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555. https://doi.org/10.1162/tacl_a_00330

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles

Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation. https://zenodo.org/records/10256836

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and Chang-Tien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, pages 8362–8372. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.671

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. 2024. Quantifying aleatoric and epistemic uncertainty with proper scoring rules. *arXiv preprint arXiv:2404.12215*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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. https://doi.org/10.18653/v1/P17-1147

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. In *Advances in Neural Information Processing Systems*, volume 35, pages 36308–36323. Curran Associates, Inc.

Nikita Kotelevskii and Maxim Panov. 2024. Predictive uncertainty quantification via risk decompositions for strictly proper scoring rules. *arXiv preprint arXiv:2402.10727*.

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, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*.

Gleb Kuzmin, Artem Vazhentsev, Artem Shelmanov, Xudong Han, Simon Suster, Maxim Panov, Alexander Panchenko, and Timothy Baldwin. 2023. Uncertainty estimation for debiased models: Does fairness hurt reliability? In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the*

Asia-Pacific Chapter of the Association for Computational Linguistics, pages 744–770, Nusa Dua, Bali. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.48

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, volume 31, pages 7167–7177.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

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

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.469

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–50, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2008

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 2018*, pages 1797–1807. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1206

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. In *Proceedings of the 4rd Workshop on Multilingual Representation Learning (MRL) @ EMNLP-2024*. https://doi.org/10.18653/v1/2024.mrl-1.15

Assaf P. Oron and Nancy Flournoy. 2017. Centered isotonic regression: Point and interval estimation for dose-response studies. *Statistics in Biopharmaceutical Research*, 9:258–267. https://doi.org/10.1080/19466315.2017.1286256

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682. https://doi.org/10.1609/aaai.v35i15.17612

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael

Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

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

Peter J. Rousseeuw. 1984. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880. https://doi.org/10.1080/01621459.1984.10477105

Aleksandr Rubashevskii, Daria Kotova, and Maxim Panov. 2023. Scalable batch acquisition for deep bayesian active learning. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 739–747. SIAM. https://doi.org/10.1137/1.9781611977653.ch83

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.593

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021a. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.145

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021b. How certain is your transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.157

Lewis Smith and Yarin Gal. 2018. Understanding measures of uncertainty for adversarial example detection. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6–10, 2018*, pages 560–569. AUAI Press.

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4115

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.330

Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gusev, Manvel Avetisian, and Leonid Zhukov. 2022. Towards computationally feasible deep active learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1198–1218, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.90

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.399

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.566

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023a. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.652

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023b. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.93

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696. https://doi.org/10.1162/tacl_a_00483

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2023. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.236

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.204

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.84

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564. https://doi.org/10.1162/tacl_a_00563

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.289

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.299

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932. https://doi.org/10.18653/v1/2023.emnlp-main.58

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1316

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Detailed Description of Uncertainty Quantification Methods

Here, we provide details of UQ methods implemented in LM-Polygraph that were omitted from the main part of the paper; see also Table 1. For ease of notation, we will write all the uncertainty measures as functions $U(\mathbf{x})$ of an input $\mathbf{x}$, though they might depend on various other instances like a generated output $\mathbf{y}$.

A.1 White-Box Methods

A.1.1 Information-Based Methods

Length-normalized log probability computes the average negative log probability of generated

tokens. If the score is exponentiated, it corresponds to *perplexity*:

$$U_{\text{Perp}}(\mathbf{x}) = \exp\left\{-\frac{1}{L}\log P(\mathbf{y} \mid \mathbf{x})\right\},$$

while it is convenient also to denote length-normalized sequence probability by $\bar{P}(\mathbf{y} \mid \mathbf{x}) = \exp\Big\{\frac{1}{L}\log P(\mathbf{y} \mid \mathbf{x})\Big\}$.

Mean token entropy simply averages entropy of each token in the generated sequence:

$$U_{\mathcal{H}_T}(\mathbf{x}) = \frac{1}{L}\sum\nolimits_{l=1}^{L} \mathcal{H}(y_l \mid \mathbf{y}_{<l}, \mathbf{x}),$$

where $\mathcal{H}(y_l \mid \mathbf{y}_{<l}, \mathbf{x})$ is an entropy of the token distribution $P(y_l \mid \mathbf{y}_{<l}, \mathbf{x})$.

Generalizing length-normalized log probability, *TokenSAR* (Duan et al., 2024) computes the weighted average of the negative log probability of generated tokens based on their relevance for the entire generated text. For a given sentence similarity function $g(\cdot, \cdot)$ and token relevance function $R_T(y_k, \mathbf{y}, \mathbf{x}) = 1 - g(\mathbf{x} \cup \mathbf{y}, \mathbf{x} \cup \mathbf{y} \setminus y_k)$, the resulting estimate is given by the following formula:

$$U_{\text{TokenSAR}}(\mathbf{x}) = \text{TokenSAR}(\mathbf{y}, \mathbf{x}) = -\sum\nolimits_{l=1}^{L} \tilde{\text{R}}_T(y_l, \mathbf{y}, \mathbf{x}) \log P(y_l \mid \mathbf{y}_{<l}, \mathbf{x}),$$

where $\tilde{\text{R}}_T(y_k, \mathbf{y}, \mathbf{x}) = \frac{\text{R}_T(y_k, \mathbf{y}, \mathbf{x})}{\sum_{l=1}^{L} \text{R}_T(y_l, \mathbf{y}, \mathbf{x})}$.

Takayama and Arase (2019) proposed an uncertainty score based on *Pointwise Mutual Information* (PMI) between generation conditioned on the prompt and unconditional generation:

$$U_{\text{PMI}}(\mathbf{x}) = \frac{1}{L}\sum\nolimits_{l=1}^{L} \log \frac{P(y_l \mid \mathbf{y}_{<l})}{P(y_l \mid \mathbf{y}_{<l}, \mathbf{x})}.$$

van der Poel et al. (2022) suggested a modification of this approach called *Conditional Pointwise Mutual Information (CPMI)* that considers only the probabilities of those tokens, for which the entropy of the conditional distribution is above a certain threshold τ:

$$U_{\text{CPMI}}(\mathbf{x}) = -\frac{1}{L}\sum_{l=1}^{L} \log P(y_l \mid \mathbf{y}_{<l}, \mathbf{x}) + \frac{\lambda}{L}\sum\nolimits_{l:\, \mathcal{H}(y_l \mid \mathbf{y}_{<l}, \mathbf{x}) \geq \tau} \log P(y_l \mid \mathbf{y}_{<l}),$$

where $\lambda > 0$ is another tunable parameter.

Rényi divergence (Darrin et al., 2023) computes the divergence between the probability distribution for each token and the uniform distribution:

$$U_{\text{RD}}(\mathbf{x}) = \frac{1}{L}\sum\nolimits_{l=1}^{L} \frac{1}{\alpha - 1} \log \sum\nolimits_{i=1}^{N} \frac{P(y_i \mid \mathbf{y}_{<l}, \mathbf{x})^{\alpha}}{\mathbf{q}_i^{\alpha-1}},$$

where $\alpha > 0$ is a tunable parameter, N is the number of tokens in the vocabulary, and $\mathbf{q} = \left[\frac{1}{N}, \ldots, \frac{1}{N}\right]$ is a probability vector with a uniform distribution over the vocabulary.

The other way to compute the distance between probability distributions is the *Fisher-Rao distance* (Darrin et al., 2023):

$$U_{\text{FR}}(\mathbf{x}) = \frac{1}{L}\sum\nolimits_{l=1}^{L} \frac{2}{\pi} \arccos \sum\nolimits_{i=1}^{N} \sqrt{P(y_i \mid \mathbf{y}_{<l}, \mathbf{x}) \cdot \mathbf{q}_i}.$$

A.1.2 Methods Based on Sample Diversity

We can compute the entropy on the sequence level $\mathbb{E}\big[-\log P(\mathbf{y} \mid \mathbf{x})\big]$, where the expectation is taken over the sequences $\mathbf{y}$ randomly generated from the distribution $P(\mathbf{y} \mid \mathbf{x})$. Unfortunately, while for token level, we have an exact way of computing the entropy, for the sequence level, we need to adhere to some approximations. In practice, we can use Monte-Carlo integration, i.e. generate several sequences $\mathbf{y}^{(k)},\ k = 1, \ldots, K$ via random sampling and compute *Monte Carlo Sequence Entropy*:

$$U_{\mathcal{H}_S}(\mathbf{x}) = -\frac{1}{K}\sum\nolimits_{k=1}^{K} \log P(\mathbf{y}^{(k)} \mid \mathbf{x}). \quad (3)$$

We can replace $P(\mathbf{y}^{(k)} \mid \mathbf{x})$ with its length-normalized version $\bar{P}(\mathbf{y}^{(k)} \mid \mathbf{x})$ leading to a more reliable uncertainty measure in some cases.

Semantic Entropy (Kuhn et al., 2023) aims to deal with the generated sequences that have similar meanings while having different probabilities according to the model, which can significantly affect the resulting entropy value (3). The idea is to cluster generated sequences $\mathbf{y}^{(k)}, k = 1, \ldots, K$ into several semantically homogeneous clusters $\mathcal{C}_m, m = 1, \ldots, M$ with $M \leq K$ with bi-directional entailment algorithm and average

the sequence probabilities within the clusters. The resulting estimate of entropy is given by:

$$U_{\text{SE}}(\mathbf{x}) = -\sum_{m=1}^{M} \frac{|\mathcal{C}_m|}{K} \log \hat{P}_m(\mathbf{x}),$$

where $\hat{P}_m(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{C}_m} P(\mathbf{y} \mid \mathbf{x})$.

SentenceSAR (Duan et al., 2024) enlarges the probability of those sentences that are more relevant and convincing than others. Given sentence relevance measure $g\big(\mathbf{y}^{(j)}, \mathbf{y}^{(k)}\big)$ of $\mathbf{y}^{(j)}$ concerning to $\mathbf{y}^{(k)}$, SentenceSAR is computed as:

$$\begin{aligned} &\text{R}_S(\mathbf{y}^{(j)}, \mathbf{x}) = \sum_{k \neq j} g\big(\mathbf{y}^{(j)}, \mathbf{y}^{(k)}\big) P\big(\mathbf{y}^{(k)} \mid \mathbf{x}\big). \\ &U_{\text{SentSAR}}(\mathbf{x}) = \\ &-\frac{1}{K} \sum_{k=1}^{K} \log\Big(P(\mathbf{y}^{(k)} \mid \mathbf{x}) + \frac{1}{t} \text{R}_S(\mathbf{y}^{(k)}, \mathbf{x})\Big), \end{aligned} \tag{4}$$

where t is a temperature parameter used to control the scale of shifting to relevance.

Combining SentenceSAR and TokenSAR results in a new method *SAR* (Duan et al., 2024). In particular, in equation (4), the generative probability $P(\mathbf{y} \mid \mathbf{x})$ is replaced with the token-shifted probability $P'(\mathbf{y} \mid \mathbf{x}) = \exp\{-\text{TokenSAR}(\mathbf{y}, \mathbf{x})\}$.

A.1.3 Density-Based Methods

Let $h(\mathbf{x})$ be a latent representation of an instance $\mathbf{x}$. The *Mahalanobis Distance* (MD; Lee et al., 2018) method fits a Gaussian centered at the training data centroid μ with an empirical covariance matrix Σ. The uncertainty score is the Mahalanobis distance between $h(\mathbf{x})$ and μ:

$$U_{\text{MD}}(\mathbf{x}) = \big(h(\mathbf{x}) - \mu\big)^T \Sigma^{-1} \big(h(\mathbf{x}) - \mu\big).$$

We suggest using the last hidden state of the encoder averaged over non-padding tokens or the last hidden state of the decoder averaged over all generated tokens as $h(\mathbf{x})$.

The *Robust Density Estimation* (RDE; Yoo et al., 2022) method improves over MD by reducing the dimensionality of $h(\mathbf{x})$ via the PCA decomposition. Additionally, the covariance matrix Σ for each class is computed using the Minimum Covariance Determinant estimation method (Rousseeuw, 1984). The uncertainty score is computed as the Mahalanobis distance but in the space of reduced dimensionality.

Ren et al. (2023) showed that it might be useful to adjust the Mahalanobis distance score by subtracting from it the other Mahalanobis distance $\text{MD}_0(\mathbf{x})$ computed for some large general-purpose dataset covering many domains (e.g., C4; Raffel et al., 2020). The resulting *Relative Mahalanobis Distance* score is

$$U_{\text{RMD}}(\mathbf{x}) = \text{MD}(\mathbf{x}) - \text{MD}_0(\mathbf{x}).$$

A.2 Black-Box Methods

A.2.1 Methods Based on Sample Diversity

Sample diversity methods sample multiple predictions from a LLM for the same prompt and analyze the diversity of the outputs across different samples. The idea is that if the LLM consistently outputs similar answers, it is confident, whereas varying outputs indicate high uncertainty. Since the LLM might output the same meaning in various surface forms by rephrasing its answers, the approaches from this category usually construct a matrix $S = (s_{ij})$ representing similarities between responses based on some semantic similarity measure and then cluster the responses into groups of answers with the same meanings.

Following Lin et al. (2024), we consider two similarity measures for responses. The first one is the Jaccard similarity $s(\mathbf{y}, \mathbf{y}') = |\mathbf{y} \cap \mathbf{y}'| / |\mathbf{y} \cup \mathbf{y}'|$, where the sequences $\mathbf{y}$ and $\mathbf{y}'$ are considered just as sets of words. Another similarity measure is based on Natural Language Inference (NLI). For each pair of input sequences, an NLI model provides two probabilities: $\hat{p}_{\text{entail}}(\mathbf{y}, \mathbf{y}')$—the degree of entailment between the sequences and $\hat{p}_{\text{contra}}(\mathbf{y}, \mathbf{y}')$—the degree the contradiction between them. The similarity between sequences $\mathbf{y}$ and $\mathbf{y}'$ is computed as $s_{\text{entail}}(\mathbf{y}, \mathbf{y}') = \hat{p}_{\text{entail}}(\mathbf{y}, \mathbf{y}')$ or $s_{\text{contra}}(\mathbf{y}, \mathbf{y}') = 1 - \hat{p}_{\text{contra}}(\mathbf{y}, \mathbf{y}')$. Following Kuhn et al. (2023), we use the DeBERTa-large NLI model (He et al., 2021).

One of the simplest techniques that leverages the idea of meaning diversity for UQ is *Number of Semantic Sets*. We adopt an iterative approach by sequentially examining responses from the first to the last while making pairwise comparisons between them (each pair has indexes j_1 and j_2, $j_2 > j_1$). The number of semantic sets initially equals the total number of generated answers K. If the condition $\hat{p}_{\text{entail}}(\mathbf{y}_{j_1}, \mathbf{y}_{j_2}) > \hat{p}_{\text{contra}}(\mathbf{y}_{j_1}, \mathbf{y}_{j_2})$ and $\hat{p}_{\text{entail}}(\mathbf{y}_{j_2}, \mathbf{y}_{j_1}) > \hat{p}_{\text{contra}}(\mathbf{y}_{j_2}, \mathbf{y}_{j_1})$ is fulfilled we put this two sentences into one cluster. The

computation is done for all the pairs of answers, and then the resulting number of distinct sets $U_{NumSemSets}$ is reported. This measure is simple yet it has many limitations. It can only take integer values and it assumes that the semantic equivalences derived from the NLI model are always transitive.

Sum of Eigenvalues of the Graph Laplacian (Lin et al., 2024) represents a more advanced approach from this category. Let's consider a similarity matrix $S_{j_1 j_2} = \big(s(\mathbf{y}_{j_1}, \mathbf{y}_{j_2}) + s(\mathbf{y}_{j_2}, \mathbf{y}_{j_1})\big)/2$. Averaging is done to obtain better consistency. The Laplacian for matrix S is given by the following formula $L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, where D is a diagonal matrix and $D_{ii} = \sum_{j=1}^{K} S_{ij}$. Consequently, the following formula is derived: $U_{EigV} = \sum_{k=1}^{K} \max(0, 1 - \lambda_k)$, where λ_k are the eigenvalues of matrix L. This value is a continuous analogue of $U_{NumSemSets}$.

U_{EigV} and $U_{NumSemSets}$ have a common disadvantage: they can not provide uncertainty for each answer. Lin et al. (2024) demonstrates that we can extract it from the diagonal *Degree Matrix* D computed above. The idea is that elements on the diagonal of D are sums of similarities between the given answer and all other answers, and the corrected trace of D provides an average pairwise distance between answers. The larger the pairwise distance is the higher is uncertainty: $U_{Deg} = 1 - trace(D)/K^2$.

A drawback of previously considered methods is the limited knowledge of the actual embedding space for the different answers since we only have measures of their similarities. Nevertheless, we can overcome this limitation by taking advantage of the inferential capabilities of the graph Laplacian, which makes it easier to obtain the coordinates of the answers. Let us introduce $\mathbf{u}_1, \ldots, \mathbf{u}_k \in R^K$ as the eigenvectors of L that correspond to k smallest eigenvalues. We can efficiently construct an informative embedding $\mathbf{v}_j = [\mathbf{u}_{1,j}, \ldots, \mathbf{u}_{k,j}]$ for an answer $\mathbf{y}_j$. Lin et al. (2024) suggest *Eccentricity* as uncertainty score—the average distance from the center in the space of constructed embeddings: $U_{Ecc} = \big\|[\tilde{\mathbf{v}}_1^T, \ldots, \tilde{\mathbf{v}}_K^T]\big\|_2$, where $\tilde{\mathbf{v}}_j = \mathbf{v}_j - \frac{1}{K}\sum_{\ell=1}^{K} \mathbf{v}_\ell$.

Lexical Similarity is a measure proposed by (Fomicheva et al., 2020) that computes how similar two words or phrases are in terms of their meaning. Since the original article is dedicated to machine translation, this measure calculates the average similarity score between all pairs of translation hypotheses in a set, using a similarity measure based on the overlap of their lexical items. Different metrics can be used, such as ROUGE-1, ROUGE-2, ROUGE-L, and BLEU. For our task, this measure iterates over all responses and calculates the average score with other answers.

A.2.2 Empirical Approximations of White-Box Methods

Following Tian et al. (2023), we introduce several empirical variations of white-box methods that can be computed in a black-box setting. *LabelProb* is a black-box approximation of MSP. Given K sampled outputs from the model, we can estimate the model-assigned probability for each of the outputs $\hat{\mathrm{P}}\big(\mathbf{y}^{(j)} \mid \mathbf{x}\big)$, based on its relative frequency among the samples: $\hat{\mathrm{P}}(\mathbf{y}^{(j)} \mid \mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}\big(\mathbf{y}^{(i)} = \mathbf{y}^{(j)}\big)$. LabelProb is computed by considering the probability of the most likely sample:

$$U_{\text{LabelProb}}(\mathbf{x}) = 1 - \max_j \hat{\mathrm{P}}\big(\mathbf{y}^{(j)} \mid \mathbf{x}\big).$$

Similarly, it is possible to estimate the black-box equivalent of *Semantic Entropy* by calculating semantic cluster probabilities based on the relative frequencies of samples within these clusters: $\hat{P}_m(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{C}_m} \hat{P}(\mathbf{y} \mid \mathbf{x})$.

We also implement a black-box approximation of P(True) (Kadavath et al., 2022), where the model is repeatedly queried for its confidence, and the relative frequency of the ‘‘True’’ output is taken as the confidence measure:

$$\hat{\mathrm{P}}(\text{‘‘True’’} \mid \mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}(\mathbf{y}^{(i)} = \text{‘‘True’’}),$$

$$U_{\text{PTrue}_{BB}}(\mathbf{x}) = 1 - \hat{\mathrm{P}}(\text{‘‘True’’} \mid \mathbf{x}).$$

A.2.3 Reflexive Methods

Instruction-tuned LLMs can also be directly prompted to output a level of confidence as a part of their output, as shown by Tian et al. (2023).

Linguistic 1S prompts the model to output its confidence along with the answer by selecting it from the list of predetermined linguistic expressions of confidence. The selected answer is mapped to a floating-point confidence level following (Fagen-Ulmschneider, 2023).

Verbalized 1S Top1/TopK prompts the model to generate both the answer and its confidence

in a single output, with the confidence expressed directly as a floating-point number. Top1/TopK approaches specify how many guesses (with corresponding confidences) the model is asked to output.

Verbalized 2S Top1/TopK differs from the previous approach by separating the answer and confidence estimation into two distinct turns of interaction with a LLM.

Verbalized 2S CoT asks the model to reason about the question and output its answer in the first response and numerical confidence in the second.

A.3 Claim-Level Methods

While the aforementioned methods operate with the entire generated sequence, it is often desirable to estimate uncertainty for individual claims to pinpoint hallucinations within the generated text. Suppose C denotes a set of token indices corresponding to a particular claim. Many UQ methods can be straightforwardly adopted for the claim level by simply considering only a subset of tokens corresponding to the set C instead of all tokens in the sequence. We consider claim-level generalization for MSP, Mean Token Entropy, Perplexity, and PMI.

P(True), adapting the approach from (Kadavath et al., 2022) to the claim level, quantifies claim uncertainty by prompting the LLM to assess the truthfulness of each generated claim:

$$U_{\text{P}_{\text{True}}}(C \mid \mathbf{x}) = 1 - P(y_1 = \text{``True''} \mid C, \mathbf{x}).$$

Claim-Conditioned Probability (CCP) quantifies uncertainty by evaluating the semantic similarity between the original claim and perturbed versions where each token is replaced with its alternative generations. CCP utilizes a Natural Language Inference (NLI) model to compare the original claim $y_{i\in C, i\leq j}$ with variations where token y_j is replaced with top-K alternatives y_j^k from the model's output distribution:

$$\text{CCP}(y_j \mid \mathbf{y}_{<j}, \mathbf{x}) = \frac{\sum_{k:\texttt{NLI}(y_j^k, y_j)=\texttt{'e'}} P(y_j^k \mid \mathbf{y}_{<j}, \mathbf{x})}{\sum_{k:\texttt{NLI}(y_j^k, y_j)\in\{\texttt{'e'},\texttt{'c'}\}} P(y_j^k \mid \mathbf{y}_{<j}, \mathbf{x})},$$

The resulting uncertainty measure becomes

$$U_{\text{CCP}}(C \mid \mathbf{x}) = 1 - \prod_{j\in C} \text{CCP}(y_j \mid \mathbf{y}_{<j}, \mathbf{x}).$$

Here, $\texttt{NLI}(y_j^k, y_j) = \texttt{'e'}$ denotes that the NLI model predicts an entailment relation between the original claim and the modified claim where y_j is replaced with y_j^k. CCP effectively measures the proportion of high-probability token alternatives that preserve the original claim's semantic meaning according to the NLI model.

B Detailed Description of Uncertainty Normalization Methods

Linear scaling first computes confidence scores by negating uncertainty scores in the calibration set: $c_i = -u_i$. Then, for a new model output with a corresponding uncertainty score u and confidence $c(u) = -u$, the normalized confidence is $c_s(u) = (c(u) - \min_i c_i)/(\max_i c_i - \min_i c_i)$. To ensure the uncertainty scores for tested instances remain within the $[0, 1]$ interval, the confidence scores are clipped accordingly.

Quantile scaling computes confidence using the uncertainty cumulative distribution function estimated using calibration data: $c(u) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(u_i \leq u)$, where N is the number of data points in $\mathcal{D}_{calib}$ and $u_i \in \mathcal{D}_{calib}$. This approach naturally bounds the confidence between 0 and 1 and does this with consideration to the distribution of uncertainty values in the calibration set.

Binned PCC splits the calibration set into non-intersecting bins based on uncertainty values u_i. Thus, a bin is a set of indices $\mathcal{B}^j = \{i\colon b^j_{\min} \leq u_i < b^j_{\max}\}$, where $b^j_{\min}$ and $b^j_{\max}$ are left and right boundaries of the j-th bin respectively.

Then, for a new data point with a raw uncertainty score u, a calibrated confidence score is $c(u) = \frac{1}{|\mathcal{B}'|}\sum_{i\in\mathcal{B}'} q_i$, where $\mathcal{B}'$ is the calibration bin, for which the following holds: $b'_{\min} \leq u < b'_{\max}$.

To remedy this problem, we propose *Isotonic PCC*. In this method the confidence score is obtained as $c(u) = \text{CIR}(u)$, where CIR is a fitted Centered Isotonic Regression that predicts the quality of the response based on its raw uncertainty score, maintaining strict monotonicity, i.e. higher uncertainty always produces lower confidence.

C Detailed Experimental Results for Uncertainty Quantification Methods

UQ Method	CoQA AlignScore	TriviaQA AlignScore	MMLU Accuracy	GSM8k Accuracy	Mean Rank	Mean PRR
Maximum Sequence Probability	0.28±0.01	0.56±0.01	**0.52±0.01**	0.35±0.01	8.88	**0.43**
Perplexity	0.24±0.01	0.55±0.01	0.41±0.01	0.43±0.01	9.25	0.41
Mean Token Entropy	0.26±0.01	0.56±0.01	0.08±0.01	0.45±0.01	12.50	0.34
Pointwise Mutual Information	−0.05±0.01	−0.00±0.01	0.40±0.01	0.13±0.01	20.50	0.12
Conditional Pointwise Mutual Information	−0.14±0.01	−0.20±0.01	0.30±0.01	−0.38±0.01	25.00	−0.11
Rényi Divergence	0.00±0.01	−0.07±0.01	0.41±0.01	−0.10±0.01	19.75	0.06
Fisher-Rao Distance	0.02±0.01	−0.07±0.01	0.41±0.01	−0.06±0.01	19.12	0.08
TokenSAR	0.24±0.01	0.53±0.01	0.41±0.01	0.43±0.01	10.12	0.40
CCP	0.21±0.01	0.61±0.01	0.47±0.01	0.42±0.01	7.38	**0.43**
Monte Carlo Sequence Entropy	0.24±0.01	0.46±0.01	0.17±0.01	0.37±0.01	16.88	0.31
Monte Carlo Normalized Sequence Entropy	0.22±0.01	0.46±0.01	0.17±0.01	0.35±0.01	18.50	0.30
Semantic Entropy	0.26±0.01	0.53±0.01	0.22±0.01	0.40±0.01	14.50	0.35
SentenceSAR	0.34±0.01	0.58±0.01	0.39±0.01	0.19±0.01	10.50	0.38
SAR	0.33±0.01	0.59±0.01	0.33±0.01	0.40±0.01	7.88	0.41
Mahalanobis Distance - Decoder	−0.03±0.01	−0.14±0.01	−0.02±0.01	0.29±0.01	24.38	0.02
RDE - Decoder	−0.02±0.01	−0.04±0.01	0.14±0.01	0.30±0.01	22.25	0.10
Relative Mahalanobis Distance - Decoder	0.00±0.01	−0.14±0.01	−0.13±0.01	0.29±0.01	23.75	0.00
HUQ-MD - Decoder	0.24±0.01	0.55±0.01	0.41±0.01	**0.50±0.01**	8.62	0.42
P(True)	−0.01±0.01	0.13±0.01	−0.29±0.01	0.29±0.01	23.00	0.03
NumSet	0.22±0.01	0.58±0.01	0.26±0.01	0.18±0.01	17.25	0.31
EigValLaplacian NLI Score Entail.	0.35±0.01	0.60±0.01	0.31±0.01	0.21±0.01	10.50	0.37
EigValLaplacian Jaccard Score	0.28±0.01	0.55±0.01	0.27±0.01	0.33±0.01	14.38	0.36
DegMat NLI Score Entail.	**0.37±0.01**	**0.63±0.01**	0.35±0.01	0.27±0.01	8.25	0.40
DegMat Jaccard Score	0.30±0.01	0.57±0.01	0.33±0.01	0.36±0.01	10.25	0.39
Eccentricity NLI Score Entail.	0.33±0.01	0.57±0.01	0.28±0.01	0.36±0.01	10.75	0.38
Eccentricity Jaccard Score	0.30±0.01	0.53±0.01	0.28±0.01	0.32±0.01	14.38	0.36
Lexical Similarity Rouge-L	0.34±0.01	0.59±0.01	0.34±0.01	0.42±0.01	**6.25**	0.42
Lexical Similarity BLEU	0.29±0.01	0.53±0.01	0.34±0.01	0.40±0.01	11.25	0.39

Table 6: PRR↑ 50% with various generation quality metrics for UQ methods in selective QA tasks with the Stable LM 2 12B model. Warmer color indicates better results.

UQ Method	CoQA AlignScore	TriviaQA AlignScore	MMLU Accuracy	GSM8k Accuracy	Mean Rank	Mean PRR
Maximum Sequence Probability	0.33±0.01	0.63±0.01	0.47±0.01	0.47±0.01	6.38	**0.48**
Perplexity	0.29±0.01	0.64±0.01	0.47±0.01	0.31±0.01	10.62	0.43
Mean Token Entropy	0.26±0.01	0.64±0.01	0.45±0.01	0.36±0.01	11.75	0.43
Pointwise Mutual Information	−0.08±0.01	0.03±0.01	0.45±0.01	0.14±0.01	19.75	0.14
Conditional Pointwise Mutual Information	−0.11±0.01	−0.15±0.01	0.45±0.01	−0.08±0.01	21.75	0.03
Rényi Divergence	−0.04±0.01	−0.26±0.01	0.00±0.01	−0.33±0.01	25.75	−0.16
Fisher-Rao Distance	−0.03±0.01	−0.29±0.01	−0.00±0.01	−0.34±0.01	25.88	−0.16
TokenSAR	0.29±0.01	0.62±0.01	0.47±0.01	0.31±0.01	12.00	0.42
CCP	0.29±0.01	0.61±0.01	**0.48±0.01**	**0.53±0.01**	7.50	**0.48**
Monte Carlo Sequence Entropy	0.26±0.01	0.52±0.01	0.35±0.01	0.49±0.01	15.00	0.40
Monte Carlo Normalized Sequence Entropy	0.25±0.01	0.57±0.01	0.36±0.01	0.40±0.01	16.50	0.40
Semantic Entropy	0.29±0.01	0.58±0.01	0.39±0.01	0.48±0.01	12.75	0.43
SentenceSAR	**0.35±0.01**	0.66±0.01	0.43±0.01	0.26±0.01	8.88	0.42
SAR	0.32±0.01	**0.67±0.01**	0.41±0.01	0.50±0.01	**4.88**	**0.48**
Mahalanobis Distance - Decoder	−0.04±0.01	−0.14±0.01	−0.04±0.01	0.38±0.01	22.25	0.04
RDE - Decoder	0.03±0.01	−0.04±0.01	0.23±0.01	0.36±0.01	20.38	0.14
Relative Mahalanobis Distance - Decoder	−0.05±0.01	−0.14±0.01	−0.11±0.01	0.39±0.01	22.50	0.02
HUQ-MD - Decoder	0.28±0.01	0.64±0.01	0.47±0.01	0.48±0.01	7.88	0.47
P(True)	−0.06±0.01	−0.12±0.01	0.07±0.01	0.00±0.01	24.50	−0.03
NumSet	0.21±0.01	0.65±0.01	0.36±0.01	0.24±0.01	16.25	0.36
EigValLaplacian NLI Score Entail.	0.31±0.01	0.65±0.01	0.39±0.01	0.30±0.01	12.00	0.41
EigValLaplacian Jaccard Score	0.25±0.01	0.62±0.01	0.35±0.01	0.39±0.01	16.25	0.40
DegMat NLI Score Entail.	**0.35±0.01**	**0.67±0.01**	0.41±0.01	0.34±0.01	7.62	0.44
DegMat Jaccard Score	0.31±0.01	0.63±0.01	0.40±0.01	0.42±0.01	10.62	0.44
Eccentricity NLI Score Entail.	0.32±0.01	0.64±0.01	0.35±0.01	0.45±0.01	10.75	0.44
Eccentricity Jaccard Score	0.30±0.01	0.60±0.01	0.35±0.01	0.28±0.01	17.12	0.38
Lexical Similarity Rouge-L	0.33±0.01	0.64±0.01	0.40±0.01	0.48±0.01	7.38	0.46
Lexical Similarity BLEU	0.31±0.01	0.61±0.01	0.40±0.01	0.46±0.01	11.12	0.44

Table 7: PRR↑ 50% with various generation quality metrics for UQ methods in selective QA tasks with the Mistral 7B v0.2 model. Warmer color indicates better results.

UQ Method	CoQA AlignScore	TriviaQA AlignScore	MMLU Accuracy	Mean Rank	Mean PRR
NumSet	0.20±0.01	0.54±0.01	0.29±0.00	13.17	0.34
EigValLaplacian NLI Score Entail.	0.34±0.01	**0.60±0.00**	0.32±0.00	4.50	**0.42**
EigValLaplacian Jaccard Score	0.34±0.01	0.58±0.00	0.29±0.00	8.33	0.40
DegMat NLI Score Entail.	0.34±0.01	0.58±0.00	0.30±0.00	6.83	0.41
DegMat Jaccard Score	0.34±0.01	0.59±0.00	0.29±0.00	7.50	0.41
Eccentricity NLI Score Entail.	**0.36±0.01**	**0.60±0.00**	0.31±0.00	**3.67**	**0.42**
Eccentricity Jaccard Score	0.34±0.01	0.59±0.00	0.29±0.00	7.50	0.41
Lexical Similarity Rouge-L	**0.36±0.01**	0.58±0.00	0.29±0.00	7.00	0.41
Lexical Similarity BLEU	0.34±0.01	0.56±0.00	0.29±0.00	9.67	0.40
BB Semantic Entropy	0.31±0.01	0.57±0.00	0.29±0.00	10.50	0.39
Label Prob.	0.31±0.01	0.57±0.00	0.29±0.00	10.50	0.39
BB P(True)	0.01±0.01	0.40±0.01	0.32±0.00	13.83	0.24
Linguistic 1S	0.13±0.01	0.26±0.01	0.23±0.00	17.67	0.21
Verbalized 1S top-1	0.18±0.01	0.42±0.01	0.38±0.00	12.00	0.33
Verbalized 1S top-k	0.23±0.01	0.52±0.01	0.37±0.00	10.00	0.37
Verbalized 2S CoT	0.28±0.01	0.51±0.00	0.40±0.00	9.00	0.40
Verbalized 2S top-1	0.22±0.01	0.55±0.00	0.39±0.00	9.00	0.39
Verbalized 2S top-k	0.19±0.01	0.48±0.01	**0.44±0.00**	10.33	0.37

Table 8: PRR↑ 50% with various generation quality metrics for black-box UQ methods in selective QA tasks with the GPT-4o-mini model. Warmer color indicates better results.

UQ Method	CoQA AlignScore	TriviaQA AlignScore	MMLU Accuracy	Mean Rank	Mean PRR
NumSet	0.23±0.01	0.56±0.01	0.19±0.01	6.33	0.33
EigValLaplacian NLI Score Entail.	**0.34±0.01**	**0.58±0.01**	**0.23±0.01**	**1.33**	**0.38**
EigValLaplacian Jaccard Score	0.29±0.01	0.53±0.01	0.05±0.01	10.50	0.29
DegMat NLI Score Entail.	0.33±0.01	0.56±0.01	0.19±0.01	3.67	0.36
DegMat Jaccard Score	0.31±0.01	0.53±0.01	0.04±0.01	10.17	0.29
Eccentricity NLI Score Entail.	0.33±0.01	0.52±0.01	0.14±0.01	7.00	0.33
Eccentricity Jaccard Score	0.33±0.01	0.51±0.01	0.12±0.01	8.17	0.32
Lexical Similarity Rouge-L	0.31±0.01	0.54±0.01	0.15±0.01	6.50	0.33
Lexical Similarity BLEU	0.28±0.01	0.48±0.01	0.07±0.01	12.33	0.28
BB Semantic Entropy	0.32±0.01	0.52±0.01	0.15±0.01	7.00	0.33
Label Prob.	0.31±0.01	0.50±0.01	0.14±0.01	9.17	0.32
BB P(True)	−0.00±0.01	0.19±0.01	0.16±0.01	13.33	0.12
Linguistic 1S	−0.01±0.01	0.04±0.01	0.10±0.01	16.67	0.04
Verbalized 1S top-1	0.09±0.01	0.42±0.01	0.12±0.01	12.83	0.21
Verbalized 1S top-k	0.15±0.01	0.37±0.01	**0.23±0.01**	9.00	0.25
Verbalized 2S CoT	0.04±0.01	0.14±0.01	0.22±0.01	12.00	0.13
Verbalized 2S top-1	0.02±0.01	0.24±0.01	0.08±0.01	15.33	0.11
Verbalized 2S top-k	0.11±0.01	0.36±0.01	**0.23±0.01**	9.67	0.23

Table 9: PRR↑ 50% with various generation quality metrics for black-box UQ methods in selective QA tasks with the Stable LM 2 12B Chat model. Warmer color indicates better results.

UQ Method	CoQA AlignScore	TriviaQA AlignScore	MMLU Accuracy	Mean Rank	Mean PRR
NumSet	0.21±0.01	0.51±0.01	0.20±0.01	10.50	0.31
EigValLaplacian NLI Score Entail.	0.26±0.01	**0.54±0.01**	0.27±0.01	**3.50**	**0.36**
EigValLaplacian Jaccard Score	0.19±0.01	0.42±0.01	0.21±0.01	15.00	0.27
DegMat NLI Score Entail.	0.26±0.01	0.52±0.01	0.25±0.01	5.00	0.34
DegMat Jaccard Score	0.23±0.01	0.46±0.01	0.22±0.01	10.83	0.30
Eccentricity NLI Score Entail.	0.28±0.01	0.48±0.01	0.23±0.01	6.33	0.33
Eccentricity Jaccard Score	**0.29±0.01**	0.48±0.01	0.25±0.01	4.67	0.34
Lexical Similarity Rouge-L	0.25±0.01	0.44±0.01	0.23±0.01	10.50	0.31
Lexical Similarity BLEU	0.26±0.01	0.44±0.01	0.23±0.01	9.67	0.31
BB Semantic Entropy	0.28±0.01	0.48±0.01	0.22±0.01	7.50	0.33
Label Prob.	0.26±0.01	0.45±0.01	0.22±0.01	10.00	0.31
BB P(True)	−0.02±0.01	0.35±0.01	0.22±0.01	16.83	0.18
Linguistic 1S	0.13±0.01	0.43±0.01	0.25±0.01	11.83	0.27
Verbalized 1S top-1	0.10±0.01	0.48±0.01	0.29±0.01	8.00	0.29
Verbalized 1S top-k	0.10±0.01	0.44±0.01	0.26±0.01	11.00	0.27
Verbalized 2S CoT	0.21±0.01	0.38±0.01	**0.34±0.01**	9.50	0.31
Verbalized 2S top-1	0.08±0.01	0.44±0.01	0.33±0.01	10.50	0.28
Verbalized 2S top-k	0.12±0.01	0.47±0.01	0.25±0.01	9.83	0.28

Table 10: PRR↑ 50% with various generation quality metrics for black-box UQ methods in selective QA tasks with the Mistral 7B v0.2 Instruct model. Warmer color indicates better results.

UQ Method	XSUM		WMT14 Fr-En		WMT19 De-En		Mean Rank	Mean PRR
	ROUGE-L	AlignScore	COMET	AlignScore	COMET	AlignScore		
Maximum Sequence Probability	−0.24±0.01	−0.09±0.00	0.34±0.01	0.09±0.01	**0.50±0.00**	0.11±0.01	12.67	0.12
Perplexity	0.20±0.01	0.02±0.00	0.31±0.01	0.06±0.01	0.38±0.00	0.07±0.01	14.58	0.17
Mean Token Entropy	0.00±0.01	−0.41±0.00	0.37±0.01	0.08±0.01	0.42±0.00	0.10±0.01	14.00	0.09
Pointwise Mutual Information	0.25±0.01	0.12±0.00	0.04±0.01	0.00±0.01	0.05±0.00	−0.01±0.01	17.58	0.08
Conditional Pointwise Mutual Information	**0.61±0.01**	0.04±0.00	−0.18±0.01	−0.03±0.01	−0.20±0.00	−0.05±0.01	20.75	0.03
Rényi Divergence	−0.05±0.01	−0.11±0.00	−0.18±0.01	−0.03±0.01	−0.18±0.00	−0.09±0.01	25.33	−0.11
Fisher-Rao Distance	−0.06±0.01	−0.15±0.00	−0.17±0.01	−0.03±0.01	−0.18±0.00	−0.09±0.01	25.75	−0.11
TokenSAR	−0.01±0.01	−0.45±0.00	0.31±0.01	0.05±0.01	0.40±0.00	0.09±0.01	18.00	0.06
CCP	−0.06±0.01	−0.06±0.00	0.33±0.01	0.11±0.01	0.39±0.00	0.12±0.01	13.50	0.14
Monte Carlo Sequence Entropy	0.17±0.01	0.27±0.00	0.34±0.01	0.10±0.01	0.43±0.00	0.08±0.01	9.08	0.23
Monte Carlo Normalized Sequence Entropy	0.16±0.01	0.15±0.00	0.36±0.01	0.03±0.01	0.43±0.00	0.08±0.01	11.42	0.20
Semantic Entropy	0.17±0.01	**0.30±0.00**	0.35±0.01	0.12±0.01	0.45±0.00	0.13±0.01	5.33	0.25
SentenceSAR	0.05±0.01	0.09±0.00	0.33±0.01	0.11±0.01	0.46±0.00	0.09±0.01	10.08	0.19
SAR	0.18±0.01	0.20±0.00	**0.44±0.01**	0.09±0.01	**0.50±0.00**	0.15±0.01	**4.75**	**0.26**
Mahalanobis Distance - Decoder	0.00±0.01	0.03±0.00	0.00±0.01	0.09±0.01	0.06±0.00	0.04±0.01	19.25	0.04
RDE - Decoder	−0.13±0.01	−0.59±0.00	−0.02±0.01	0.06±0.01	0.04±0.00	0.07±0.01	23.50	−0.09
Relative Mahalanobis Distance - Decoder	0.11±0.01	0.10±0.00	0.03±0.01	0.09±0.01	0.08±0.00	0.10±0.01	14.33	0.08
HUQ-MD - Decoder	−0.05±0.01	−0.29±0.00	0.30±0.01	0.09±0.01	0.38±0.00	0.07±0.01	18.75	0.08
P(True)	−0.10±0.01	−0.25±0.00	0.14±0.01	0.10±0.01	0.09±0.00	0.03±0.01	20.25	0.00
NumSet	0.04±0.01	0.15±0.00	0.01±0.01	0.01±0.01	0.05±0.00	0.07±0.01	18.75	0.06
EigValLaplacian NLI Score Entail.	0.04±0.01	0.29±0.00	0.28±0.01	0.19±0.01	0.38±0.00	0.28±0.00	9.92	0.24
EigValLaplacian Jaccard Score	0.18±0.01	0.07±0.00	0.33±0.01	0.02±0.01	0.43±0.00	0.09±0.01	12.75	0.19
DegMat NLI Score Entail.	0.04±0.01	**0.30±0.00**	0.29±0.01	**0.21±0.00**	0.41±0.00	**0.31±0.00**	8.25	**0.26**
DegMat Jaccard Score	0.18±0.01	0.07±0.00	0.36±0.01	0.03±0.01	0.46±0.00	0.10±0.01	10.08	0.20
Eccentricity NLI Score Entail.	0.02±0.01	0.13±0.00	0.31±0.01	0.20±0.01	0.42±0.00	0.28±0.01	9.17	0.23
Eccentricity Jaccard Score	0.01±0.01	−0.04±0.00	0.34±0.01	0.04±0.01	0.47±0.00	0.10±0.01	12.83	0.15
Lexical Similarity Rouge-L	0.20±0.01	0.07±0.00	0.37±0.01	0.05±0.01	0.47±0.00	0.11±0.01	7.92	0.21
Lexical Similarity BLEU	0.11±0.01	−0.12±0.00	0.30±0.01	0.03±0.01	0.36±0.00	0.09±0.01	17.42	0.13

Table 11: PRR↑ 50% with various generation quality metrics for UQ methods in selective generation tasks with the Stable LM 2 12B model. Warmer color indicates better results.

UQ Method	XSUM		WMT14 Fr-En		WMT19 De-En		Mean Rank	Mean PRR
	ROUGE-L	AlignScore	COMET	AlignScore	COMET	AlignScore		
Maximum Sequence Probability	0.06±0.00	0.06±0.00	0.31±0.02	0.08±0.01	0.45±0.01	0.12±0.01	12.08	0.18
Perplexity	−0.19±0.00	−0.15±0.00	0.37±0.02	0.12±0.01	0.49±0.01	0.14±0.01	12.08	0.13
Mean Token Entropy	0.07±0.00	−0.33±0.00	**0.42±0.02**	0.14±0.01	**0.51±0.01**	0.15±0.01	8.42	0.16
Pointwise Mutual Information	−0.07±0.00	−0.09±0.00	0.01±0.02	0.00±0.01	0.08±0.01	0.00±0.01	23.17	−0.01
Conditional Pointwise Mutual Information	**0.32±0.00**	0.28±0.00	−0.12±0.02	−0.05±0.01	−0.11±0.01	−0.03±0.01	17.92	0.05
Rényi Divergence	−0.04±0.00	0.12±0.00	−0.14±0.02	−0.05±0.01	−0.19±0.01	−0.07±0.01	24.00	−0.06
Fisher-Rao Distance	−0.05±0.00	0.12±0.00	−0.14±0.02	−0.06±0.01	−0.20±0.01	−0.08±0.01	24.92	−0.07
TokenSAR	0.06±0.00	−0.36±0.00	0.35±0.02	0.12±0.01	0.46±0.01	0.14±0.01	11.50	0.13
CCP	−0.09±0.00	0.06±0.00	0.30±0.02	0.09±0.01	0.39±0.01	0.13±0.01	14.33	0.15
Monte Carlo Sequence Entropy	0.18±0.00	0.27±0.00	0.27±0.02	0.06±0.01	0.36±0.01	0.11±0.01	11.67	0.21
Monte Carlo Normalized Sequence Entropy	0.12±0.00	0.14±0.00	0.29±0.02	0.07±0.01	0.43±0.01	0.11±0.01	11.50	0.19
Semantic Entropy	0.18±0.00	0.28±0.00	0.27±0.02	0.10±0.01	0.40±0.01	0.15±0.01	7.58	0.23
SentenceSAR	0.03±0.00	−0.07±0.00	0.23±0.02	0.06±0.01	0.34±0.01	0.08±0.01	18.50	0.11
SAR	0.15±0.00	0.28±0.00	0.34±0.02	0.13±0.01	0.48±0.01	0.18±0.01	**4.42**	**0.26**
Mahalanobis Distance - Decoder	0.03±0.00	0.10±0.00	−0.05±0.02	0.04±0.01	0.04±0.01	0.05±0.01	20.83	0.04
RDE - Decoder	−0.05±0.00	−0.04±0.00	−0.09±0.02	−0.05±0.01	−0.16±0.01	−0.06±0.01	24.58	−0.08
Relative Mahalanobis Distance - Decoder	0.07±0.00	0.05±0.00	−0.02±0.02	0.05±0.01	0.00±0.01	0.06±0.01	19.83	0.04
HUQ-MD - Decoder	0.06±0.00	−0.35±0.00	0.37±0.02	0.12±0.01	0.49±0.01	0.14±0.01	10.67	0.14
P(True)	−0.00±0.00	−0.19±0.00	0.12±0.02	0.05±0.01	0.03±0.01	0.01±0.01	22.17	0.00
NumSet	0.07±0.00	**0.31±0.00**	0.04±0.02	0.04±0.01	0.07±0.01	0.09±0.01	16.25	0.10
EigValLaplacian NLI Score Entail.	0.12±0.00	0.27±0.00	0.17±0.02	0.19±0.01	0.28±0.01	0.24±0.01	9.92	0.21
EigValLaplacian Jaccard Score	0.09±0.00	0.19±0.00	0.20±0.02	0.05±0.01	0.34±0.01	0.10±0.01	15.67	0.16
DegMat NLI Score Entail.	0.13±0.00	0.30±0.00	0.22±0.02	0.22±0.01	0.34±0.01	0.29±0.01	7.67	0.25
DegMat Jaccard Score	0.11±0.00	0.21±0.00	0.28±0.02	0.08±0.01	0.43±0.01	0.12±0.01	10.67	0.20
Eccentricity NLI Score Entail.	0.06±0.00	0.15±0.00	0.28±0.02	**0.24±0.01**	0.39±0.01	**0.30±0.01**	8.75	0.24
Eccentricity Jaccard Score	0.02±0.00	−0.08±0.00	0.33±0.02	0.10±0.01	0.45±0.01	0.11±0.01	13.42	0.16
Lexical Similarity Rouge-L	0.12±0.00	0.28±0.00	0.26±0.02	0.08±0.01	0.42±0.01	0.14±0.01	9.83	0.22
Lexical Similarity BLEU	0.14±0.00	0.03±0.00	0.26±0.02	0.08±0.01	0.36±0.01	0.11±0.01	13.67	0.16

Table 12: PRR↑ 50% with various generation quality metrics for UQ methods in selective generation tasks with the Mistral 7B v0.2 model. Warmer color indicates better results.

D Additional Experimental Results with Uncertainty Normalization

UQ Method	Linear	Quantile	Binned PCC	Isotonic PCC
Maximum Sequence Probability	0.000	0.000	−0.041	0.000
Perplexity	0.000	0.000	0.016	0.001
Mean Token Entropy	0.000	0.000	0.002	0.000
Mean Pointwise Mutual Information	0.000	0.000	−0.398	0.000
Mean Conditional PMI	0.000	0.000	−0.298	0.000
Rényi Divergence	0.000	0.000	−0.121	0.000
Fisher-Rao Distance	0.000	0.000	−0.125	0.000
TokenSAR	0.000	0.000	0.022	0.000
CCP	0.000	0.000	−0.002	−0.001
Monte Carlo Sequence Entropy	0.000	0.000	−0.045	0.000
Monte Carlo Normalized Sequence Entropy	0.000	0.000	0.005	0.000
Semantic Entropy	0.000	0.000	−0.023	0.000
Sentence SAR	0.000	0.000	−0.008	0.000
SAR	0.000	0.000	0.001	0.000
P(True)	0.000	0.000	−0.784	−0.001
NumSet	0.000	0.000	0.375	0.000
EigValLaplacian NLI Score Entail	0.000	0.000	0.009	0.001
EigValLaplacian Jaccard Score	0.000	−0.002	0.000	0.000
DegMat NLI Score Entail	0.000	0.000	0.011	0.001
DegMat Jaccard Score	0.000	0.001	0.060	0.000
Eccentricity NLI Score Entail	0.000	0.000	0.023	0.000
Eccentricity Jaccard Score	0.001	0.002	0.092	0.031
Lexical Similarity Rouge-L	0.000	0.000	0.073	0.000

Table 13: The difference between PRR of raw uncertainty and bounded confidence obtained with various normalization techniques. The lower is better; negative values represent cases when confidence performs better than raw uncertainty scores.

E LLM Text Generation Quality

Model	CoQA	TriviaQA	MMLU	GSM8k	XSum		WMT14 Fr-En		WMT19 De-En	
	AlignScore	AlignScore	Accuracy	Accuracy	Rouge-L	AlignScore	Comet	AlignScore	Comet	AlignScore
Stable LM 2 12B v2	0.77	0.69	0.57	0.55	0.21	0.03	0.87	0.85	0.88	0.84
Mistral 7B v0.2	0.79	0.74	0.64	0.38	0.23	0.07	0.86	0.84	0.86	0.81

Table 14: Generation quality metrics for LLMs without instruction tuning.

Model	Prompt	CoQA	TriviaQA	MMLU
		AlignScore	AlignScore	Accuracy
GPT-4o-mini	Linguistic1S	0.74	0.81	0.75
	Verbalized 1S top-1	0.73	0.81	0.75
	Verbalized 1S top-k	0.70	0.80	0.75
	Verbalized 2S CoT	0.71	0.84	0.79
	Verbalized 2S top-1	0.74	0.81	0.76
	Verbalized 2S top-k	0.69	0.80	0.76
	Base	0.74	0.81	0.76
Stable LM 2 12B Chat	Linguistic1S	0.68	0.74	0.49
	Verbalized 1S top-1	0.69	0.73	0.52
	Verbalized 1S top-k	0.56	0.74	0.55
	Verbalized 2S CoT	0.66	0.74	0.43
	Verbalized 2S top-1	0.69	0.73	0.51
	Verbalized 2S top-k	0.69	0.74	0.56
	Base	0.69	0.73	0.51
Mistral 7B v0.2 Instruct	Linguistic1S	0.69	0.71	0.58
	Verbalized 1S top-1	0.69	0.69	0.58
	Verbalized 1S top-k	0.67	0.70	0.56
	Verbalized 2S CoT	0.60	0.69	0.54
	Verbalized 2S top-1	0.70	0.70	0.57
	Verbalized 2S top-k	0.68	0.70	0.56
	Base	0.70	0.70	0.57

Table 15: Generation quality for instruction-tuned LLMs.