ASSESSING CONFIDENCE IN LARGE LANGUAGE MODELS BY CLASSIFYING TASK CORRECTNESS USING SIMILARITY FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty quantification (UQ) provides measures of uncertainty, such as a score of confidence in an LLM's generated output, and is therefore increasingly recognized as a crucial component of trusted AI systems. Black-box UQ methods do not require access to internal model information from the generating LLM and therefore have numerous real-world advantages, such as robustness to system changes, adaptability to choice of LLM including those with commercialized APIs, reduced costs, and substantial computational tractability. In this paper, we propose a simple yet powerful UQ approach that treats confidence estimation as a probabilistic classification task, where one predicts the correctness of a generation using similarities with other generations for the same query as features. This approach requires a small labeled dataset and can be either black-box or white-box, depending on the choice of additional features for the classifier, beyond the similarities. We conduct an empirical study using 6 datasets across question answering and summarization tasks, demonstrating that features based on pairwise similarities generally result in confidence estimates that are better calibrated and more predictive of correctness as compared to the closest baselines.

1 INTRODUCTION

Uncertainty quantification (UQ) approaches in machine learning provide insights into the reliability of model predictions, and can therefore be a critical component for deploying large language models (LLMs) in real-world applications. UQ refers to a broad swathe of techniques that yield measures of uncertainty; in this paper, we are interested in assessing the *confidence* of an LLM's generations for a user-specified task. An important sub-class of UQ techniques are black-box methods, which only assume access to the model being used without requiring other model information such as the weights or even the token log probabilities. Such techniques have numerous practical advantages, as they are robust to the constantly evolving landscape of LLMs and can easily adapt to system changes. Furthermore, they are usually computationally lightweight and can be quickly deployed at inference time. As a result, black-box UQ has become increasingly popular for tasks such as question answering (Kuhn et al., 2022; Lin et al., 2024; Manakul et al., 2023; Cole et al., 2023).

Confidence is not always a well-defined quantity in the UQ literature, but it is typically represented as a number between 0 and 1. We consider tasks where there is a notion of whether an LLM's response to a user query is correct or not, and interpet confidence of a response as the probability that it is correct. Extending this idea, we propose an approach where confidence estimation is regarded as a probabilistic classification task, where we predict whether a generation is correct and view the probability of correctness as our confidence. Importantly, we suggest generating multiple samples from the LLM through some sampling procedure, computing pairwise similarities between samples using any similarity metric of choice, and then using the pairwise similarities between a generation of interest and other generations as features for a probabilistic classifier. Such an approach is black-box if it does not use additional features such as those arising from the token logits, but we also explore white-box extensions for our experiments.

Our approach can be considered a particular type of *consistency-based* UQ, where the idea is to use the consistency between a generation and other sampled generations as a proxy for our confidence in its correctness. The implicit underlying assumption behind consistency-based approaches is that when a generated response is more different from others, it is more likely to be incorrect, implying that responses that are consistently similar are more likely to be correct; this assumption has been explored for various use cases involving self-consistency (Mitchell et al., 2022; Wang et al., 2023; Chen et al., 2024). In our work, we use similarity features as a signal for correctness, as opposed to unsupervised machine learning approaches that cluster generations together (Kuhn et al., 2022; Lin et al., 2024); such methods are not designed to provide information about the correctness of generations. In contrast, our supervised learned surrogate classifier is directly trained to output the degree of correctness. We also highlight that instead of aggregating verbalized confidences (Xiong et al., 2024), we aggregate pairwise similarities between generations, thereby avoiding empirically observed concerns around overconfidence when asking LLMs for probabilities (Hu & Levy, 2023; Xiong et al., 2024).

Confidence estimates can be evaluated in various ways, depending on how they will be used by the system builder, system, or end user. We are primarily interested in approaches yielding confidences that are well *calibrated*, as gauged by how closely they align with the empirical accuracy of the predictions (Murphy & Epstein, 1967; Dawid, 1982). We also evaluate our proposed approaches based on whether confidence estimates are used to select from a set of generations, or to predict whether a generation is correct or not. Classifying task correctness using similarity features is generally shown to perform well on all chosen metrics as compared to baselines, particularly those measuring calibration error.

Our **contributions** are summarized as follows:

- We propose a UQ approach that treats confidence estimation as classification; specifically, the objective is to estimate the probability of a generation being correct for a given task and query, using pairwise similarities with other generations from the LLM for the same query as features.
- We conduct an empirical investigation using 6 datasets – 3 each for question answering and summarization tasks. We demonstrate that using similarity features for classification yields confidence estimates with more desirable properties around calibration and predictive capability, as compared to the closest baselines.

2 Related Work

Procedures for uncertainty quantification (UQ) estimate measures like the variability or confidence of LLM outputs. These methods are categorized as either white-box or black-box. White-box methods assume access to the LLM's internal components, such as model weights, logits, or embeddings. In contrast, black-box methods rely only on outputs, inferring confidence through alternative means, such as consistency across paraphrased inputs. An orthogonal distinction is between verbalized and non-verbalized methods. Verbalized methods prompt the LLM to express uncertainty in natural language, such as using phrases ("I don't know" or "most probably") or quantitative indicators ("low" or "50%" or "90%").

White-Box Methods. Common approaches to estimating an LLM's confidence include considering the minimum or average token-level probabilities (logits) or entropy (Huang et al., 2023; Vazhentsev et al., 2023) coupled with a normalization mechanism to ensure consistency over outputs of different lengths (Murray & Chiang, 2018). Linguistic semantics such as token-level or sentence-level relevance can also be incorporated into these schemes to yield more effective confidence estimators (Duan et al., 2024). Kuhn et al. (2022) propose semantic entropy based clustering on multiple samples generated from the model and then estimating confidence estimates by summing the token-level probabilities in each cluster. Kadavath et al. (2022) suggest a verbalized method where the LLM first generates responses and then evaluates them as either True or False; the probability the model assigns to the generated token (True or False) determines the confidence level.

Other approaches consider the LLM's internal state such as embeddings and activation spaces. For instance, Ren et al. (2023) compute embeddings for both inputs and outputs in the training data, fit them to a Gaussian distribution, and estimate the model's confidence by computing the distance of the evaluated data pair from this Gaussian distribution. Some methods probe the model's attention layers to discriminate between correct and incorrect answers (Kadavath et al., 2022; Burns et al., 2023; Li et al., 2023; Azaria & Mitchell, 2023). Although these methods provide insights into the model's linguistic understanding, they require supervised training on specially annotated data.

Black-Box Methods. One strand of research considers verbalized black-box methods, such as using an LLM to evaluate the correctness of its own generated answers in a conversational agent scenario (Mielke et al., 2022). Xiong et al. (2024) conduct an empirical study on UQ for reasoning tasks, showing that LLMs tend to be overconfident when verbalizing their own confidence in the correctness of the generated answers and align poorly with the likelihood of factual correctness, which may pose significant safety risks in real-world deployments of LLMs. Other related work includes that of Lin et al. (2022) around fine-tuning GPT-3 to verbalize the uncertainty associated with the generated answers. Analysis in Hu & Levy (2023) reveals that LLMs' meta-linguistic judgments are less reliable than quantities derived directly from the model's token-level probabilities.

Many black-box methods use similarity between multiple generations given an input question, where common choices of metrics are natural language inference (NLI) scores (Kuhn et al., 2022), Jaccard index (Qurashi et al., 2020), or embedding-based similarity such as Sentence-BERT (SBERT) (Reimers, 2019). Such similarity metrics can be used to extend clustering algorithms for uncertainty quantification of LLMs (Kuhn et al., 2022; Ao et al., 2024; Da et al., 2024; Jiang et al., 2024). Another promising direction of work assumes that a model's lack of confidence correlates with various responses, often leading to hallucinatory outputs. In this case, confidence is typically estimated by analysing the consistency among various responses of the model. Specifically, Manakul et al. (2023) propose a simple sampling-based approach that uses consistency among generations to find potential hallucinations. Lin et al. (2024) calculate the similarity matrix between generations and then estimate the uncertainty based on the analysis of the similarity matrix, such as the sum of the eigen-values of the graph Laplacian, the degree matrix, and the eccentricity. Recent methods have also explored combining white-box and black-box approaches (Chen & Mueller, 2024; Shrivastava et al., 2023).

Our proposed approach falls within the non-verbalized UQ category and can be either black-box or white-box, depending on the information that is leveraged. The key aspect is that the method relies on assessing the consistency among generations and uses similarities between generations as the basis for features for a classifier.

3 CONFIDENCE ESTIMATION AS CONSISTENCY-BASED CLASSIFICATION

We describe our proposed approach, beginning with some basic notation and assumptions.

3.1 NOTATION AND ASSUMPTIONS

Consider an LLM generating output y for some input query x. We assume there is an associated ground truth output y^* for input x as well as a binary reward $r \in \{0, 1\}$ from a reward function $r(x, y, y^*)$. Importantly, we assume there is a way to gauge whether any particular generation (with corresponding ground truth) is correct or not, i.e. whether the reward is 1 or 0. $Y^*(x)$ denotes the set of responses with reward 1. For tasks such as open-ended question answering and summarization, the reward is gauged to be 1 if a text similarity metric (e.g. RougeL) between the ground truth and generated output exceeds a predetermined threshold; this has also been assumed by related prior work (Kuhn et al., 2022; Lin et al., 2024). In this work, we assume that a sampling procedure generates multiple samples/generations $y_1, \cdots, y_m$ for input query x. We are interested in assessing the confidence of any arbitrary generation y_i, $i \in \{1, \cdots, m\}$, which we denote as c_i.

After generating samples, our consistency-based approach relies on access to a similarity metric with which one can compute pairwise similarities $s(y_i, y_j)$ for all sample pairs, all assumed to lie in the interval $[0, 1]$. As shorthand, we denote the similarities between the i^{th} generation and other generations as $\mathbf{s}_i = s_{i,1}, .., s_{i,i-1}, s_{i,i+1}, .., s_{i,m}$, where $s_{i,k} = s(y_i, y_k)$ is the similarity between samples y_i and y_k. For our experiments, we consider the Jaccard coefficient as similarity metric, but also run ablations with variations of the Rouge metric such as Rouge1 and RougeL. We note that any similarity metric can be used, but in our experience, metrics that treat generations as sets of tokens are often most suitable for consistency-based UQ.

3.2 Classifying Task Correctness

We posit that the similarities between an LLM's generation and other generations, through a suitable sampling procedure, can be used as a signal for estimating its confidence. We therefore propose treating confidence estimation as a classification task; specifically, we train a probabilistic classifier for whether a response is correct using similarities are features.

Suppose we are interested in estimating the confidence c_i for generation y_i as response to input query x. Let us denote $\mathbf{s}_i^f$ as the set of similarity features for classifying correctness. Then, the confidence for generation y_i is computed as: $c_i = P(y_i \in Y^*(x)) = f(\mathbf{s}_i^f)$. This method can be generalized further by also including other non-similarity features $\mathbf{o}_i^f$, such as the generative score from the LLM, in which case $c_i = f(\mathbf{s}_i^f, \mathbf{o}_i^f)$. For our experiments, we learn the function $f(\cdot)$ using a random forest as the probabilistic classifier, but other methods are also applicable. Also, we compare variations of our proposed classification approach with different feature sets. For instance, in an important proposed variation, we only use pairwise similarities with other generations as features, in which case $\mathbf{s}_i^f = \mathbf{s}_i$ and $\mathbf{o}_i^f = \emptyset$. Details about the specific variations are provided later, when describing the experiments.

Note that classifying task correctness requires a small training set that includes ground truth responses. If we ensure that the sampling procedure during training is similar to that during test time, one can first train a classifier using similarities as features, and then deploy the trained classifier to predict whether a generation is correct at test time.

4 Empirical Investigation

We conduct an empirical investigation demonstrating the value of leveraging pairwise similarities between generations as features for confidence estimation.

4.1 Experimental Setup

We provide details about our experimental setup in this subsection. Note that we restrict ourselves to using representative open-source LLMs for generation.

Datasets and Models. We study datasets involving question answering (QA) and summarization:

- **QA**: We consider the open-book conversational QA dataset CoQA (Reddy et al., 2019), the closed-book QA dataset TriviaQA (Joshi et al., 2017), as well as the more challenging closed-book QA dataset called Natural Questions (Kwiatkowski et al., 2019). We take the first 1000 questions from the dev sets of each dataset, and generate responses using two open-source models. **Granite 13B** (Mishra et al., 2024) and **LLaMA 2 70B** (Touvron et al., 2023). QA is a widely studied task in the recent literature on UQ.
- **Summarization**: For this task, we experiment with the following datasets: XSum (Narayan et al., 2018), SamSum (Gliwa et al., 2019) and CNN Dailymail (Nallapati et al., 2016) As before, we consider the first 1000 prompts from the validation splits of each dataset and generate summaries using the same open-source models. Note that summarization typically results in longer form generations than those arising from question answering.

Evaluation Metrics. We consider the following 3 evaluation metrics, each capturing a different facet of how confidence estimates may be utilized by the system and/or user:

- As a performance metric, we propose *accuracy from top selection* (**ATS**), which measures the accuracy (fraction of correct instances) in a test set when confidences are used to select one generation from a set of generations for a query. This represents the situation where the system is expected to provide a single response for every query to the user, and confidences are used as scores for selection among generations.
- As a calibration metric, we choose *adaptive calibrated error* (**ACE**), which bins confidence estimates into probability ranges such that each bin contains the same number of data points (Nixon et al., 2019). Formally, $ACE = \frac{1}{KB} \sum_{k=1}^{K} \sum_{b=1}^{B} |acc(b,k) - c(b,k)|$, where $acc(b,k)$ and

Table 1: Comparing different UQ approaches over 3 evaluation metrics on generations from 2 models on the CoQA dataset. Each approach is marked as either black-box (BB) or white-box (WB). Error bars are from max. and min. values over 5 runs, each with a random 50% train / 50% test split.

Model	Llama2 70B			Granite 13B		
	ATS ↑	ACE ↓	AUROC ↑	ATS ↑	ACE ↓	AUROC ↑
Baselines						
always 1 (BB)	$0.11_{\pm 0.01}$	$0.440_{\pm 0.003}$	$0.50_{\pm 0.00}$	$0.66_{\pm 0.01}$	$0.176_{\pm 0.007}$	$0.50_{\pm 0.00}$
avg. log prob (WB)	$0.11_{\pm 0.01}$	$0.707_{\pm 0.007}$	$0.54_{\pm 0.04}$	$0.64_{\pm 0.01}$	$0.317_{\pm 0.017}$	$0.73_{\pm 0.01}$
spec-ecc (BB)	$0.22_{\pm 0.01}$	$0.346_{\pm 0.012}$	$0.26_{\pm 0.04}$	$0.28_{\pm 0.02}$	$0.597_{\pm 0.006}$	$0.17_{\pm 0.01}$
arith-agg (BB)	$0.16_{\pm 0.02}$	$0.233_{\pm 0.006}$	$0.74_{\pm 0.04}$	$0.72_{\pm 0.02}$	$0.056_{\pm 0.016}$	$0.83_{\pm 0.01}$
clf-gen (WB)	$0.16_{\pm 0.01}$	$0.042_{\pm 0.010}$	$0.49_{\pm 0.04}$	$0.64_{\pm 0.02}$	$0.081_{\pm 0.016}$	$0.73_{\pm 0.01}$
Proposed						
clf-mean (BB)	$0.26_{\pm 0.03}$	$0.043_{\pm 0.011}$	$0.74_{\pm 0.01}$	$0.72_{\pm 0.02}$	$0.045_{\pm 0.016}$	$0.82_{\pm 0.02}$
clf-pairs (BB)	$0.45_{\pm 0.01}$	$0.052_{\pm 0.008}$	$\mathbf{0.83}_{\pm 0.03}$	$\mathbf{0.78}_{\pm 0.01}$	$0.056_{\pm 0.010}$	$\mathbf{0.86}_{\pm 0.01}$
clf-mean+gen (WB)	$0.29_{\pm 0.02}$	$\mathbf{0.039}_{\pm 0.004}$	$0.75_{\pm 0.01}$	$0.73_{\pm 0.02}$	$\mathbf{0.042}_{\pm 0.006}$	$0.83_{\pm 0.02}$
clf-pairs+gen (WB)	$\mathbf{0.46}_{\pm 0.02}$	$0.052_{\pm 0.011}$	$\mathbf{0.83}_{\pm 0.04}$	$\mathbf{0.78}_{\pm 0.01}$	$0.060_{\pm 0.011}$	$\mathbf{0.86}_{\pm 0.01}$

$c(b, k)$ are the accuracy and confidence of adaptive calibration bin b for class label k. We prefer using adaptive bin sizes instead of fixed bin sizes, as the latter often results in unbalanced datapoints across bins. We set the # of bins $B = 5$ for all experiments.
- As a prediction metric, we consider the *area under the receiver operating characteristic* (**AUROC**), which computes the area under the curve of the false positive rate vs. true positive rate when confidences are used as a probabilistic classifier for the correctness of generations.

Baselines. We consider the following baselines:

- *always 1* is a naive baseline that always returns confidence of 1.
- *avg. log prob* computes a probability by exponentiating the average logit over generated tokens and is often used in prior work on QA (Kuhn et al., 2022; Lin et al., 2024; Manakul et al., 2023); we denote the generative score for generation y_i as p_i^g.
- *spec-ecc* is a spectral clustering approach for UQ that leverages a graph Laplacian matrix computed from pairwise similarities and uses eccentricity (Lin et al., 2024).
- *arith-agg* is an approach that estimates a generation's confidence by taking the arithmetic mean of pairwise similarities with other generations; it is mathematically equivalent to a spectral clustering approach that uses degree (Lin et al., 2024).
- *clf-gen* is a classification approach where the only feature is the generative score p_i^g (as described above), which is the exponent of the avg. logit across generated tokens.

Proposed Methods. We consider the following variations of our proposed classification approach. In the following, recall that $\mathbf{s}_i^f$ and $\mathbf{o}_i^f$ refer to similarity and other features respectively:

- *clf-mean* is when we only use mean similarity as the sole feature, i.e. $\mathbf{s}_i^f = \bar{\mathbf{s}}_i$, $\mathbf{o}_i^f = \emptyset$.
- *clf-pairs* is when all pairwise similarities are features, i.e. $\mathbf{s}_i^f = \mathbf{s}_i$, $\mathbf{o}_i^f = \emptyset$.
- *clf-mean+gen* includes the generative score with mean similarity, i.e. $\mathbf{s}_i^f = \bar{\mathbf{s}}_i$, $\mathbf{o}_i^f = p_i^g$.
- *clf-pairs+gen* includes the generative score with all pairwise similarities, i.e. $\mathbf{s}_i^f = \mathbf{s}_i$, $\mathbf{o}_i^f = p_i^g$.

All classification approaches use a random forest with a maximum depth of 4 in our experiments.

4.2 Main Results

We investigate the effectiveness of our proposed classification approach using generations from 2 different models on 6 datasets from QA and summarization. For our sampling procedure, we generate 5 samples each over 6 temperatures, from 0.25 to 1.5 in increments of 0.25. Evaluations are performed only on samples from the lower 3 temperatures since the higher temperatures provide generations with lower performance. This captures the realistic scenario where the user wishes

Table 2: Comparing different UQ approaches over 3 evaluation metrics on generations from 2 models on the Samsum dataset. Each approach is marked as either black-box (BB) or white-box (WB). Error bars are from max. and min. values over 5 runs, each with a random 50% train / 50% test split.

Model	Llama2 70B			Granite 13B		
	ATS ↑	ACE ↓	AUROC ↑	ATS ↑	ACE ↓	AUROC ↑
Baselines						
always 1 (BB)	$0.08_{\pm 0.02}$	$0.462_{\pm 0.007}$	$0.50_{\pm 0.00}$	$0.04_{\pm 0.01}$	$0.478_{\pm 0.003}$	$0.50_{\pm 0.00}$
avg. log prob (WB)	$0.08_{\pm 0.02}$	$0.862_{\pm 0.014}$	$0.63_{\pm 0.01}$	$0.03_{\pm 0.01}$	$0.830_{\pm 0.006}$	$0.48_{\pm 0.04}$
spec-ecc (BB)	$0.07_{\pm 0.02}$	$0.180_{\pm 0.005}$	$0.34_{\pm 0.03}$	$0.00_{\pm 0.00}$	$0.524_{\pm 0.009}$	$0.33_{\pm 0.02}$
arith-agg (BB)	$0.08_{\pm 0.01}$	$0.463_{\pm 0.012}$	$\mathbf{0.71}_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.197_{\pm 0.006}$	$0.87_{\pm 0.02}$
clf-gen (WB)	$0.08_{\pm 0.01}$	$0.032_{\pm 0.006}$	$0.58_{\pm 0.02}$	$0.04_{\pm 0.01}$	$\mathbf{0.018}_{\pm 0.005}$	$0.60_{\pm 0.04}$
Proposed						
clf-mean (BB)	$0.08_{\pm 0.01}$	$0.026_{\pm 0.008}$	$0.69_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.023_{\pm 0.007}$	$0.86_{\pm 0.02}$
clf-pairs (BB)	$\mathbf{0.09}_{\pm 0.01}$	$0.024_{\pm 0.010}$	$\mathbf{0.71}_{\pm 0.01}$	$\mathbf{0.12}_{\pm 0.02}$	$0.022_{\pm 0.006}$	$\mathbf{0.91}_{\pm 0.01}$
clf-mean+gen (WB)	$\mathbf{0.09}_{\pm 0.01}$	$0.027_{\pm 0.009}$	$0.70_{\pm 0.01}$	$0.09_{\pm 0.01}$	$0.020_{\pm 0.004}$	$0.89_{\pm 0.01}$
clf-pairs+gen (WB)	$\mathbf{0.09}_{\pm 0.01}$	$\mathbf{0.023}_{\pm 0.010}$	$\mathbf{0.71}_{\pm 0.01}$	$\mathbf{0.12}_{\pm 0.01}$	$0.022_{\pm 0.004}$	$\mathbf{0.91}_{\pm 0.01}$

to obtain confidence estimates for only those samples they will even consider. We split the data randomly into half for train/test sets, and repeat the experiment 5 times so as to study variability of the results. To gauge the correctness of a generation, we use a rougeL threshold of 0.5 for QA datasets and 0.3 for summarization datasets. Furthermore, the Jaccard coefficient is used as a similarity metric for all methods that leverage pairwise similarity.

Tables 1 and 2 compare various baseline and proposed UQ approaches for generations from 2 models for the CoQA and SamSum datasets, respectively. All 3 evaluation metrics are considered in these tables, where a lower ACE is preferred but higher ATS and AUROC are preferred. Error bars are computed over 5 runs where the data is split each time into equally sized train/test sets. Comparing the performance of each UQ method as shown in the rows, separately for each model, we observe that the proposed classification approaches that use similarity features are generally high performing across all metrics. The contrast with baselines is pronounced for the CoQA dataset where the proposed approaches are notably better. The baseline that computes the arithmetic mean of pairwise similarities (*arith-agg*) is a reasonably strong one, particularly for AUROC.

Table 3 compares a smaller set of UQ approaches for generations from a Llama3 model[1] for the 4 other datasets – Natural Questions, TriviaQA, CNN Daily, and XSum. For readability, we only include 2 evaluation metrics – ACE and AUROC – and do not show the error bars in this table. Once again, we note that the proposed approaches generally perform well across these datasets, and that the *arith-agg* baseline is competitive on AUROC.

4.3 Ablations

We conduct a few ablational studies to understand the impact of some our choices for the main experimental study.

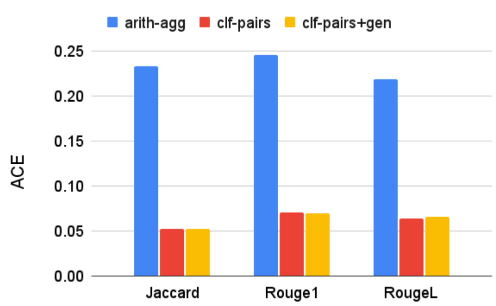
Similarity Metric. The Jaccard coefficient was used as our primary choice of similarity metric for the main experiments. Figure 1 compares 3 similarity-based UQ approaches for the ACE metric, where we consider the Rouge1 and RougeL similarity metrics in addition to Jaccard. The figure shows that the trends generally remain the same regardless of choice of similarity metric – the proposed classification approaches are comparable to each other and better performing than the *arith-agg* baseline on the ACE evaluation metric. Similar trends are noted for the ATS evaluation metric.

RougeL Correctness Threshold. The RougeL threshold is an important parameter in our experiments as it determines whether a generation is correct by comparison with the ground truth. We

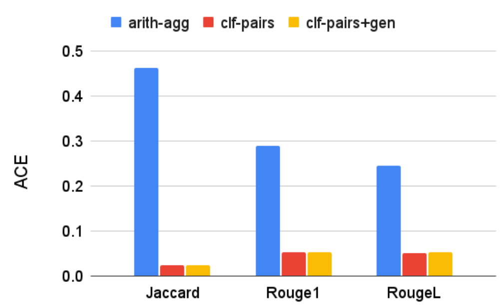
[1]We used the `llama-3.3-70b-instruct` model.

Table 3: Comparing different UQ approaches over 2 evaluation metrics on generations from the Llama3 model on 4 other datasets: Natural Questions (NQ), TriviaQA, CNN, and XSum. Each approach is marked as either black-box (BB) or white-box (WB). We only show the best performing baselines and proposed methods here. Error bars are not shown for readability.

Dataset	NQ		TriviaQA		CNN		XSum	
	ACE ↓	AUROC ↑	ACE ↓	AUROC ↑	ACE ↓	AUROC ↑	ACE ↓	AUROC ↑
Baselines								
avg. log prob (WB)	0.414	0.72	0.174	0.79	0.700	0.60	0.824	0.54
arith-agg (BB)	**0.043**	0.76	0.086	**0.88**	0.319	0.62	0.433	**0.55**
clf-gen (WB)	0.082	0.71	0.071	0.77	0.052	0.59	0.038	0.53
Proposed								
clf-pairs (BB)	0.046	**0.77**	0.044	**0.88**	**0.038**	**0.64**	0.035	0.54
clf-mean+gen (WB)	0.050	0.75	**0.038**	0.87	0.049	0.61	0.035	0.53
clf-pairs+gen (WB)	0.047	**0.77**	0.041	0.87	0.039	**0.64**	**0.034**	0.54

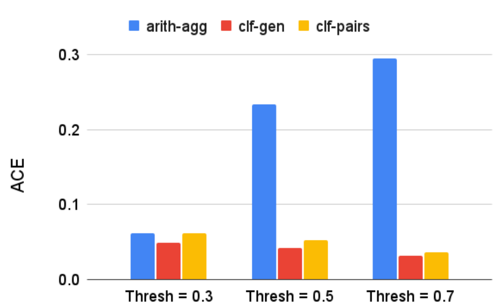


(a) Similarity Metric Ablation: CoQA

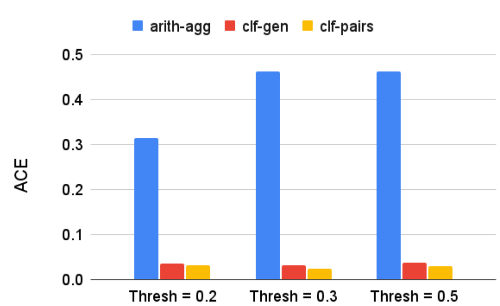


(b) Similarity Metric Ablation: SamSum

Figure 1: Effect of choice of similarity metric on the ACE evaluation metric for the CoQA and SamSum datasets.



(a) RougeL Threshold Ablation: CoQA



(b) RougeL Threshold Ablation: SamSum

Figure 2: Effect of choice of rougeL threshold on the ACE evaluation metric for the CoQA and SamSum datasets.

conduct an ablation to understand the impact of the choice of this threshold. Figure 2 compares 3 UQ approaches (including 2 baselines) for the ACE metric, where this threshold varies in $\{0.3, 0.5, 0.7\}$ for the CoQA dataset and in $\{0.2, 0.3, 0.5\}$ for the SamSum dataset. We observe again that the trends generally remain the same across threshold choices and the 2 datasets. The *clf-gen* baseline is competitive with the proposed *clf-pairs* approach for the ACE evaluation metric.

5 CONCLUSIONS

We propose a simple yet powerful method for estimating the confidence in an LLM's generations. The approach is non-verbalized, as it does not rely on asking an LLM for its confidence about a generation, and can be categorized as consistency-based, since it relies on using consistency between generations as a signal for confidence. Specifically, we view confidence estimation as a probabilistic classification task, where the objective is to predict the correctness of a generation using similarities with other generations for the same query as features. This approach is easily generalizable to include other features that may be relevant for an application. Through an empirical evaluation using 6 datasets addressing the tasks of question answering and summarization, we show that using similarity features results in confidence estimates that perform well on various UQ evaluation metrics, particularly adaptive calibration error.

One limitation of our proposed approach is that it requires a small training set where samples are generated in the same manner across training and testing. In future work, we will conduct further experiments with additional ablations to better understand the impact of similarity as well as other features for similarity aggregation methods.

REFERENCES

Shuang Ao, Stefan Rueger, and Advaith Siddharthan. CSS: Contrastive semantic similarity for uncertainty quantification of LLMs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

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

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, August 2024.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*, 2024.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, 2023.

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

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 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)*, pp. 5050–5063, 2024.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237, 2019. URL `http://arxiv.org/abs/1911.12237`.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040–5060, 2023.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. Look before you leap: An exploratory study of uncertainty measure- ment for large language models. *preprint arXiv: 2307.10236 [cs.CL]*, 2023.

Mingjian Jiang, Yangjun Yangjun, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. Graph-based uncertainty metrics for long-form language model generations. In *Annual Conference on Neural Information Processing Systems*, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611, 2017.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2022.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *preprint arXiv: 2306.03341 [cs.CL]*, 2023.

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

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 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*, pp. 9004–9017, 2023.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. Granite code models: A family of open foundation models for code intelligence. *preprint arXiv: 2405.04324 [cs.CL]*, 2024.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1754–1768, 2022.

Allan H Murphy and Edward S Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755, 1967.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Conference on Machine Translation: Research Papers*, pp. 212–223, 2018.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gu̇lçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028/.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745, 2018. URL http://arxiv.org/abs/1808.08745.

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

Abdul Wahab Qurashi, Violeta Holmes, and Anju P Johnson. Document processing: Methods for semantic text similarity analysis. In *2020 international conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1–6. IEEE, 2020.

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

N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. Llamas know what GPTs don't show: Surrogate models for confidence estimation. *preprint arXiv: 2311.08877 [cs.CL]*, 2023.

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

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 1430–1454, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.