Label-Confidence-Aware Uncertainty Estimation in Natural Language Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) display formidable capabilities in generative tasks but also pose potential risks due to their tendency to generate hallucinatory responses. Uncertainty Quantification (UQ), the evaluation of model output reliability, is crucial for ensuring the safety and robustness of AI systems. Recent studies have concentrated on model uncertainty by analyzing the relationship between output entropy under various sampling conditions and the corresponding labels. However, these methods primarily focus on measuring model entropy with precision to capture response characteristics, often neglecting the uncertainties associated with greedy decoding results, the sources of model labels, which can lead to biased classification outcomes. In this paper, we explore the biases introduced by greedy decoding and propose a label-confidence-aware (LCA) uncertainty estimation based on Kullback-Leibler (KL) divergence bridging between samples and label source, thus enhancing the reliability and stability of uncertainty assessments. Our empirical evaluations across a range of popular LLMs and NLP datasets reveal that different label sources can indeed affect classification, and that our approach can effectively capture differences in sampling results and label sources, demonstrating more effective uncertainty estimation.

1 Introduction

Large language models (LLMs) have demonstrated formidable capabilities in natural language processing tasks such as machine translation (Fomicheva et al., 2020), abstract text summarization (Brown et al., 2020), and question-answering (Touvron et al., 2023). Techniques such as In-context Learning (ICL) (Dong et al., 2022) and Chain-of-Thought (COT) (Wei et al., 2022) have further enhanced model performance on complex reasoning tasks and scenarios involving unseen data, consistently setting new benchmarks. However, despite their proficiency under scaling laws (Kaplan et al., 2020), these models underperform on more challenging tasks like mathematical problems (Luo et al., 2023). A significant concern is that, rather than refusing to answer, models are more likely to generate answers that include illusory reasoning processes and hallucinations. Uncertainty estimation and measurement have become essential tools in machine learning aiding in determining the extent to which humans can trust AI-generated content and deciding when to intervene with manual assistance. Previous research works in this field have involved prompting LLMs to self-assess the confidence of their own answers or employing confidence assessments based on model outputs using logits or entropy. Recent development *Semantic Entropy* (SE) (Kuhn et al., 2023) has introduced semantic-based entropy prediction schemes in that account for the synonym phenomena inherent in language models, performing answer aggregation in semantic space. Duan et al. (2023) and Bakman et al. (2024) propose schemes *SAR* and *MARS* based on semantic importance weighting, focusing on more precisely measuring the information content in the model's latent space to offer viable approaches to align the sampling entropy more closely with the actual value. However, we observe that the confidence and semantic alignment of the answers which serve as label sources, as well as their deviations from the distribution space, significantly impact the entropy's classification performance, an aspect overlooked by these schemes.

As shown in Figure 1, when given a question, in the beam search multi-sampling strategy, three out of the five answers generated by the LLM are correct, but due to the high overall entropy value, the LLM may be marked as unable to answer this question. Such an error is caused by the entropy threshold used in the evaluation only considering

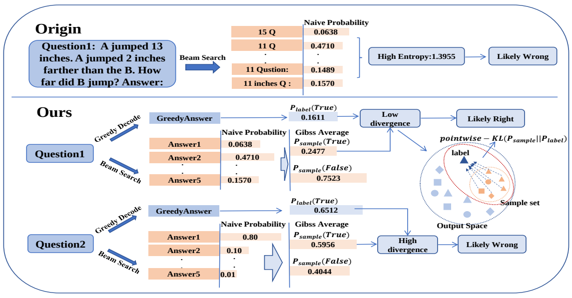


Figure 1: Ignoring the probability information of the label answer in Free-form may lead to incorrect uncertain classification. We term it as label confidence unawareness, and integrate the omitted information into our method.

the absolute value, such as the common $-\log(0.5)$, and ignoring the distribution of the model itself for the question, that is, the greedy decoding probability is lower than the probability corresponding to the sample entropy value, which is 0.1661 as shown below.

To mitigate this issue, as shown in Figure 1, we propose a label-confidence-aware (LCA) uncertainty estimation based on Kullback-Leibler divergence (KLD) bridging between samples and label source, thus enhances the reliability and stability of uncertainty assessments. We first sample answers of question as well as the output probabilities for calculating entropy of sample set. We then obtain an average probability stand for the samples and merge it with labeled answer probability by KLD to measure their difference, and use the integrated information to classify whether the model could answer the question or whether the answer can be trusted.

Our work contributes in the following ways:

- We conduct experiments on 5 models and 5 datasets on recently popular methods, identifying and reporting biases in the uncertainty measurement methods when assessing different answers and sample sizes, as well as analyze the reasons behind these biases based on semantic probabilities.

- We introduce a novel method for estimating uncertainty, termed Label-Confidence-Aware (LCA), which is based on what we refer to as Gibbs probability. This method explicitly accounts for the discrepancies between the sampling outcomes and the observed results when quantifying uncertainty.

- We evaluate multiple important free-form question-answering datasets on the currently popular pre-trained LLMs. Results demonstrate that our LCA based on KLD surpass baseline methods. Furthermore, through hyperparameter ablation experiments, we show how the variables in our method affect the final results.

2 Related Work

Verbalization and logit-based or entropy-based methods play a crucial role in addressing uncertainty in the field of Natural Language Processing (NLP). The verbalization methods which prompt models to output confidence levels for their generated content, first introduced by Lin et al. (2022), unfortunately often result in overconfident outputs. Enhancements such as COT reasoning (Xiong et al., 2023) and multi-round dialogue cross models (Cohen et al., 2023) encourage models to stimulate multi-steps reasoning for a more convincing scores. Fine-tuning methods transforms model confidence outputs into assessments of answer correctness in a designed format and tuning the models with specially crafted data (Kapoor et al., 2024; Han et al., 2024). Logit-based and entropy-based methods assess model confidence and uncertainty by focusing on the logits during the output process. Kadavath et al. (2022) add a classification head to the model's final layer, mapping logits to the probability of the "True" token, thus estimating the model's confidence in its responses. Huang et al. (2023) combine token-level probabilities and one-sentence entropy to evaluate the uncertainty in model-generated content. Jiang et al. (2021) proposes to mitigate the miscalibration of token probability caused by linguistic synonymy through data augmentation training and temperature fine-tuning and Farquhar et al., 2024 suggests that aggregates probabilities of synonymous sentences at the sentence-level in the multi-sampling process for better hallucination detection

3 Background

Total uncertainty includes aleatoric uncertainty —measuring the ambiguity inherent in the problem itself, and epistemic uncertainty -measuring the uncertainty in predictions due to a lack of knowledge within the models. It can be understood as the entropy of the model's predictions, Predictive Entropy (PE). For a given input x and output space Y, the predictive entropy is calculated as following:

$$PE(x) = -\int P(y|x) \log P(y|x) dy, \quad (1)$$

where $P(y|x)$ is the conditional probability of generation y.

The higher $PE(x)$ is, the closer the model's output probabilities are to a uniform distribution, indicating lower confidence in any specific output y out of the output space Y, and thus greater model uncertainty.

In Bayesian networks, the sampling space for a model with a vocabulary of K tokens generating sequences of length L is exponentially large, specifically $|K|^L$, posing computational challenges. To mitigate these challenges, we can employ Monte Carlo sampling (Gal and Ghahramani, 2016), which introduces random factors to approximate the sampling process.

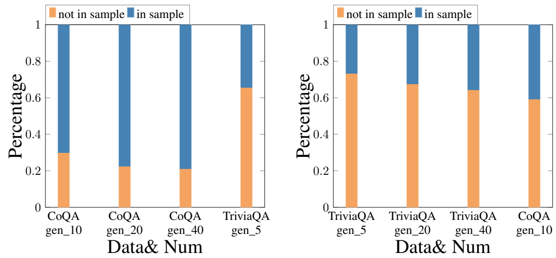
Under the condition of sufficient sampling quantity, an unbiased estimate of entropy can be:

$$\begin{aligned} PE(x) &= -\frac{1}{|N|} \sum_{y \sim Y} \log P(y|x) \\ &= -\log \prod_{y \sim Y} P(y|x)^{\frac{1}{|N|}} = -\log \tilde{P}. \end{aligned} \quad (2)$$

So we get $\tilde{P} = e^{-PE(x)}$. This form resembles the Gibbs factor, which represents the overall probability of system in physics. We refer to this value as "Gibbs probability", a probability estimation for the sampled outcome distribution of the problem. Besides, the probability derived from a corresponding greedy decoded answer is termed the observed probability.

As probabilities tend to decrease with increasing length, length-normalization method (Malinin and Gales, 2020), replacing probability of y with $\frac{1}{N}\sum_i^N \log P(y_i|y_{<i})$, could be used to scale the conditional probabilities of sentences of different lengths to the same magnitude and has been successfully applied in machine translation scenarios (Murray and Chiang, 2018).

While in natural language generation tasks for sequence prediction, different sentences may express the same meaning, thus sharing a common semantic space. SE introduced an effective UQ method in the level of semantic cluster in which uncertainty is the average of each cluster entropy.



(a) Falcon-7B (b) Mistral-7B

Figure 2: Percentage of Falcon-7B and Mistral-7B w. & w/o label answers in sample on CoQA and TriviaQA.

The formula is expressed as follows:

$$\begin{aligned} SE(x) &= -\sum_{c \in C} P(c|x) \log P(c|x) \\ &= -\sum_{c \in C} ((\sum_{s \in c} P(s|x) \log (\sum_{s \in c} P(s|x)))) \\ &\approx -|C|^{-1} \sum_{i=1}^{|C|} \log P(C_i|x). \end{aligned} \quad (3)$$

Similar to prior works, in our study, we also normalize the entropy values obtained through different methods based on length.

4 Entropy Bias in Evaluating Different Subjects

Uncertainty Quantification calculate a value about information content of high-probability samples. The higher the total probability of the sampling results, the closer it approximates the true distribution. Then such a value is then evaluated on the effectiveness of priorly representing the quality of greedy decoded answer.

To analyze the representativeness of the greedy decoded label, we evaluated the relationship between the greedy decoded label and the sampled results. The datasets and models we used here are the same as those described in experiment section 6. Specifically, we first measured the ROUGE-L score between the labeled answer and the sampled answers. Denoting sample set as $\mathcal{S}$ and the greedy decoded answer as $\mathcal{G}$, $\mathcal{G}$ is considered to be in $\mathcal{S}$ if at least one $Rouge-L(\mathcal{S}_i, \mathcal{G})$ exceeds a predefined threshold α:

$$\text{sim}(\mathcal{S}, \mathcal{G}) = \begin{cases} 1 & if\ \exists\, \text{Rouge}(\mathcal{S}_i, \mathcal{G}) > \alpha \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Figure 2 illustrates the occurrence of greedy decoded results within the sampled outcomes for

Table 1: Uncertainty estimation AUROCs of *LNPE* & *SE* with and without labeled answers in sample set.

model	data	num	*LNPE* in	*LNPE* not in	*SE* in	*SE* not in
Falcon-7B	CoQA	10	**0.7332**	0.5466	**0.7394**	0.5344
	CoQA	20	**0.7245**	0.6820	**0.7121**	0.6663
	TriviaQA	5	0.5225	**0.5547**	**0.7117**	0.6197
Mistral-7B	CoQA	10	**0.7473**	0.4233	**0.7720**	0.3834
	TriviaQA	5	**0.6408**	0.4720	**0.7492**	0.5098
	TriviaQA	20	**0.6392**	0.5322	**0.7662**	0.4622
	avg		**0.6601**	0.5507	**0.7256**	0.5771

Falcon-7B and Mistral-7B over CoQA and TriviaQA (refer to the Appendix A for more results). Our results indicate that in many cases, the greedy results do not appear within the sampled set. Even when we increase the number of samples per question to 20 or 40, such a phenomenon is not significantly alleviated. This observation aligns with results from SE (Kuhn et al., 2023), that performance improvements tend to plateau once the number of samples reaches five. This indicates that, although we hope the sampled outcomes would effectively represent the entire semantic space, current sampling strategies often fail to meet this objective.

We further grouped the test data according to whether it is in or not in sample set to analyze the impact on the classification performance of the set entropy. We used the Area Under the Receiver Operating Characteristic (AUROC) metric to evaluate performance. The algorithm is shown below: We

Algorithm 1 Comparison between groups

Require: model M, questions Q, answer G, threshold α, sets A, B, LA, LB, greedy-decoded answer g, samples S, label L

```
for each q ∈ Q do
    Generate g and samples S using model M
    L = 1 if Rouge-L(G, g) > α else 0
    for each s ∈ S do
        Calculate β = Rouge-L(g, s)
        if β > α then
            A ← A ∪ {g}, LA ← LA ∪ {L}
        else
            B ← B ∪ {g}, LB ← LB ∪ {L}
        end if
    end for
end for
Calculate AUC(A, LA) and AUC(B, LB)
```

conduct experiments on LNPE (Malinin and Gales, 2020) scheme and SE scheme. The models and the datasets remain the same as those mentioned above.

We present the experimental results in Table 1. In most cases, when the greedy decoded answer is in the sampled results, the entropy of the sampled results aligns with the quality of labeled answer well and the performance drops significantly when this is not the case. We focus on bridging between those two circumstances to mitigate the misclassification.

5 Method

Based on the previous experimental conclusions, we believe that introducing label answers into the sample set may improve performance. An intuitive method is to group labeled answer based on sim$(\mathcal{S}, \mathcal{G})$, however it not only incurs significant additional computational costs but also becomes effective only when the greedy answer introduces new answers. Additionally, when a label source answer is merged into the sampled set, its inherent confidence level should still be considered as a vital piece of information. Our label-confidence-aware (LCA) method, designed to effectively link answers from any label source to the sampled results, shifts the focus to probabilities. By integrating the overall probability of the joint sampling distribution which derived from the entropy-based Gibbs probability with observed outcomes, it identifies a more efficient and stable metric for measurement.

For a given problem x, we first use multinomial beam search to sample M sequences from $P(Y|x)$, resulting in a sample set $\{s_1, s_2, ...s_M\}$. We then compute the semantic implications between each sentence and categorize them into $|C|$ clusters using RoBERTa-Large (Liu et al., 2019), The conditional probability of a cluster containing N sequences is the sum of the probabilities of the sentences. At the cluster level, we calculate the entropy E_x and the corresponding Gibbs probability. Then we greedily decode a represent answer of which probability is P_{greedy}. We consider the aggregated probability of the sampling results as a measure of confidence, representing the model's perceived probability of a set to be able to provide an answer, considered as $P(True)$. Similarly, we view the probability of the greedy results as the observed probability that can provide an correct answer, considered as $P^{'}(True)$.

5.1 Pointwise KL-Divergence

When we introduce a new labeled answer to measure the overall probability of the calculation, this answer will introduce epistemic uncertainty. We used Kullback-Leibler divergence (KLD) to quantifiy the information lost when one distribution is used to approximate another and to messure the new uncertainties arising from noisy labels. In our study, we employ KLD between distributions of sampling results and observed outcomes as a metric to measure model uncertainty. This can help us analyze to what extent the greedy decoding labels may be overconfident or underestimated. Specifically, we use the pointwise KL divergence between these two distributions, as described by Robert (2014), focusing solely on the probability differences between tokens within the distributed answers:

$$\text{Differ}_{KLD}(\mathcal{S}, \mathcal{G}) = \tilde{P} \log \frac{\tilde{P}}{P_{\mathcal{G}}}. \tag{5}$$

5.2 Why Gibss probability?

The Expected Pairwise KL Divergence (EPKL) is another measure of uncertainty that quantifies the total bidirectional divergence between each pair of samples in the model. We derive that our method is calculated from a geometric mean perspective, integrating information from all sampled answers in one direction and smoothing out some details, making it more suitable for an overall assessment of the entire sampling distribution, while EPKL is based on the arithmetic mean, which leads to numerical instability when there is significant variance among sample results. More details refer to Appendix E.2.

6 Experiments

Baselines. We chose vinilla Length Normalizaiton Predictive Entropy (LNPE) (Malinin and Gales, 2020), Semantic Entropy (SE) (Kuhn et al., 2023), and Shift Attention Towards Relevance (SAR) (Duan et al., 2023) as baselines, and enhancing them with aggregation methods to compare performance. Detailed implementations are available in Appendix B.

Models. Following experimental methodologies in the SE and SAR studies, we conduct experiments using open-source LLMs, including models from the Llama 2 (Touvron et al., 2023), OPT (Zhang et al., 2022), Falcon (Penedo et al., 2023), and Mistral (Jiang et al., 2023) series, ranging in size from 2.7B to 13B parameters. Detailed experimental configurations can be found in Appendix C.

Datasets. We conduct experiments on several free-form text generation tasks in NLP, including CoQA (Reddy et al., 2019), Natural Questions (NaturalQA) (Kwiatkowski et al., 2019) , TriviaQA (Joshi et al., 2017) , SciQ (Welbl et al., 2017) and SVAMP (Patel et al., 2021). CoQA is a machine reading comprehension task, SciQ, NaturalQA and TriviaQA are open domain tasks, and SVAMP focuses on mathematical problems. Details regarding the composition of the test sets can be found in Appendix D.

Correctness Metric We employ the ROUGE-L metric to determine the labels, which serve as a classification result for whether the model can answer the question. The datasets we focus on are primarily concerned with sentence-level generation, making ROUGE-L the most commonly used evaluation metric for these types of tasks. Unless specifically stated otherwise, we set the default ROUGE threshold to 0.5, as this is a commonly accepted value.

Evaluation Metric Following the prior works, we used AUROC as a evaluation metric, which is popular in binary classification tasks. Furthermore, We calculated the Pearson correlation coefficient to analyze the performance of our method in the case of continuous classification.

Hyperparameters. For the CoQA dataset, we generated 10 answers per question, while for others, we generated 5 answers per question. We set the generation temperature at 0.5 which works best. In the SAR experiments, the parameter t was set to 10. To be consistent with prior works, we employed greedy search to generate the most probable answers for evaluating correctness labels and utilized multinomial sampling to produce reference generations. All experiments were carried out using two NVIDIA A40 GPUs.

7 Results Analysis

In Table 2, we provide a detailed performance comparison between our LCA method and the baselines across evaluation datasets using models including OPT-2.7B, Falcon-7B, Mistral-7B, Llama2-7B and OPT-13B. In the majority of cases, our metric outperforms the baseline. Our LCA method, in the average results of all data, has an AUROC that exceeds the SAR method by 5.5%, the TokenSAR method by 6.8%, the SE method by 8.5%, and the LNPE method by 12%. Even when the OPT-13B model achieves a high AUROC score of 0.8514 on the SciQ dataset on LNPE, LCA method still enhances its performance further, reaching 0.9033. On the challenging SVAMP, our method significantly outperforms baselines by effectively analyz-

Table 2: Uncertainty estimation AUROCs of our LCA method with different methods as backbone and baselines across datasets.

model	data	*LNPE*		*SE*		TokenSAR		SAR	
		base	LCA	base	LCA	base	LCA	base	LCA
OPT-2.7B	CoQA	**0.7377**	0.6934	0.7037	**0.7048**	0.7006	**0.7055**	0.7116	**0.7165**
	TriviaQA	0.7418	**0.9304**	0.7477	**0.8499**	0.7524	**0.8042**	0.7540	**0.8011**
	NaturalQA	0.7573	**0.7670**	0.8488	**0.8617**	**0.8673**	0.8624	**0.8675**	0.8661
Mistral-7B	CoQA	0.6217	**0.8629**	0.6206	**0.7652**	0.6227	**0.7377**	0.6215	**0.7180**
	TriviaQA	0.5928	**0.8803**	0.6189	**0.8030**	0.6272	**0.7433**	0.6257	**0.7244**
	NaturalQA	0.5461	**0.6521**	0.5716	**0.5959**	0.5662	**0.5944**	0.5695	**0.5932**
	SciQ	0.5933	**0.8640**	0.6720	**0.8237**	0.6980	**0.7808**	0.6972	**0.7731**
	SVAMP	0.6385	**0.7902**	0.5734	**0.8291**	0.5781	**0.8309**	0.5773	**0.8039**
Falcon-7B	CoQA	**0.7674**	0.7137	**0.7472**	0.7448	0.7384	**0.7415**	0.7485	**0.7519**
	TriviaQA	0.6098	**0.7637**	0.6902	**0.7715**	**0.6953**	0.6799	**0.6969**	0.6828
	NaturalQA	0.4800	**0.5365**	0.5815	**0.5918**	0.5916	**0.5993**	0.5949	**0.6033**
	SciQ	0.7136	**0.8812**	0.7200	**0.8294**	0.7046	**0.7330**	0.7109	**0.7350**
	SVAMP	0.6793	**0.8441**	0.6701	**0.8342**	0.6696	**0.8304**	0.6699	**0.8220**
Llama2-7B	CoQA	0.7636	**0.8602**	0.7465	**0.8146**	0.7333	**0.7886**	0.7475	**0.7917**
	TriviaQA	0.5720	**0.8064**	0.6336	**0.7660**	0.6289	**0.7071**	0.6287	**0.7013**
	NaturalQA	0.5500	**0.5990**	0.6267	**0.6437**	0.6215	**0.6473**	0.6247	**0.6476**
	SciQ	0.5827	**0.8054**	0.6150	**0.7468**	0.6133	**0.6922**	0.6153	**0.6892**
	SVAMP	0.6242	**0.8737**	0.5319	**0.8804**	0.5368	**0.8803**	0.5401	**0.8172**
OPT-13B	CoQA	**0.7438**	0.7250	0.7309	**0.7337**	0.7277	**0.7340**	0.7376	**0.7436**
	TriviaQA	0.5839	**0.8285**	0.6897	**0.7995**	0.6934	**0.7100**	0.6949	**0.7098**
	NaturalQA	0.6990	**0.7429**	0.7428	**0.7562**	**0.7515**	0.7456	0.7489	**0.7523**
	SciQ	0.8514	**0.9033**	0.6824	**0.7725**	0.7214	**0.7675**	0.7280	**0.7620**
avg		0.6568	**0.7874**	0.6711	**0.7690**	0.6745	**0.7420**	0.6778	**0.7364**

Table 3: Pearson correlation coefficient results of experiments.

model	SE		LNPE		TokenSAR		SAR	
	base	LCA	base	LCA	base	LCA	base	LCA
OPT-2.7B	0.202	**0.286**	0.210	**0.298**	0.053	**0.254**	0.220	**0.255**
Falcon-7B	0.208	**0.306**	0.191	**0.288**	0.124	**0.237**	0.214	**0.233**
Mistral-7B	0.135	**0.372**	0.123	**0.409**	**0.462**	0.315	0.138	**0.231**
Llama2-7B	0.147	**0.278**	0.146	**0.291**	**0.309**	0.234	0.154	**0.205**
OPT-13B	0.174	**0.249**	0.160	**0.243**	0.066	**0.198**	0.187	**0.202**
avg	0.173	**0.298**	0.166	**0.306**	0.203	**0.248**	0.183	**0.225**

ing the relationship between the probability divergence among the sample sets and observed results

We also calculated the average Pearson correlation coefficients performance of different methods on 5 datasets on 5 models. Results are shown in Table 3. These results show that our proposed metric has a stronger correlation with ROUGE-L and performs better as a priori representation of NLG answer quality, surpassing metrics designed only for classification tasks.

We further explored the impact of introducing perturbations to the label sources and probabilities. By using labels derived from different answer strategies, we aimed to more deeply analyze the importance and effectiveness of establishing a connection between the two probabilities. This was achieved by comparing the overall model performance and the associated uncertainty. We employed various strategies for replacing labels. On LNPE, we chose the highest probability sample from the sampling set, denoted as $LNPE_{sample}$, as the label source. On SE, we chose the sample with the highest probability from the largest semantic cluster, denoted as SE_{sample}. Additionally, in both experiments, we randomly pick samples from the sets, $LNPE_{random}$ and SE_{random} to get new labels for evaluation. On SE, we add a control group that integrates the greedy decoded answers into a sample set based on semantic similarity. Specially, if the semantic similarity between the greedy-decoded answer and s_i is the highest and exceeds 0.5, $\mathcal{G}$ is assigned to the semantic cluster containing s_i. Otherwise, $\mathcal{G}$ is assigned to a new semantic cluster.

Our results in Table 4 show that, in both LNPE and SE experiments, labels from sampled answers significantly surpass the baseline in AUROC. We attribute this observation to the fact that samples, as part of the sampled set, exhibit a stronger correlation with the Gibbs probability of the set. The probability of a sample, to some extent, reflects the contribution of its label within the set—a stronger contribution often implies that its label is more representative of the overall labels. Additionally, as the highest probability in the entire semantic space or within the largest semantic cluster of the sam-

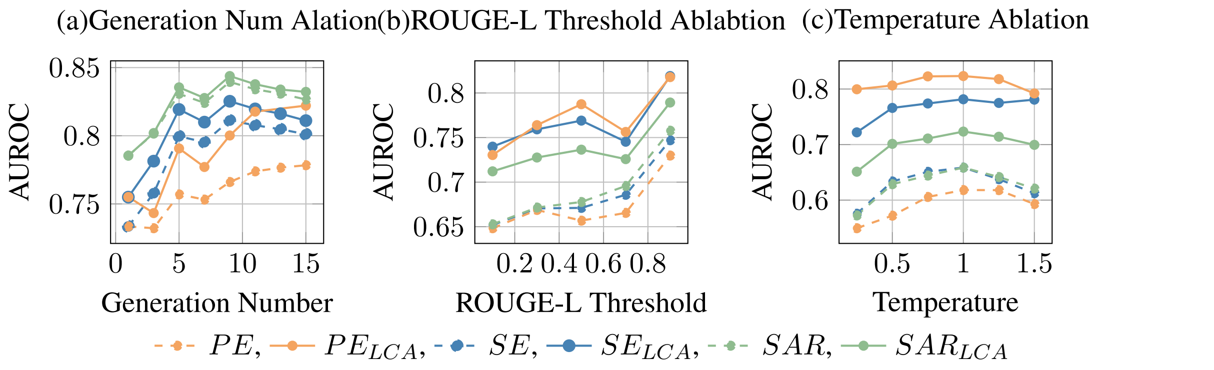


Figure 3: Ablation results. (a):Num of generation ablation. As number rises, AUROCs increase and then levels off.(b)ROUGE-L threshold ablation. As the higher threshold is, a stricter critirion it is and the better result we get. (c)TriviaQA temperature ablation on Llama2-7B. As the temperature rises, AUROCs first increase and then decrease

Table 4: Uncertainty estimation AUROCs of *LNPE* & *SE* with labels from different strategies. TQ stands for TriviaQA, sp stands for Sample, and rd stands for random.

model &data	num	*LNPE* base	*LNPE* sp	*LNPE* rd	*SE* base	*SE* sp	*SE* rd	*SE* merge
Falcon-7B								
CoQA	10	0.747	**0.748**	0.734	0.747	**0.772**	0.748	0.746
CoQA	20	**0.737**	0.736	0.719	0.721	**0.747**	0.734	0.718
TQ	5	0.549	**0.589**	0.479	0.690	0.729	0.623	**0.761**
Mistral-7B								
CoQA	10	0.608	**0.777**	0.746	0.620	0.802	0.770	**0.833**
TQ	5	0.567	**0.678**	0.649	0.619	0.808	0.730	**0.818**
TQ	20	0.578	**0.680**	0.621	0.620	**0.811**	0.6798	0.748
avg		0.631	**0.701**	0.658	0.670	**0.778**	0.714	0.771

Table 5: Uncertainty estimation AUROCs of baseline and LCA method in different datasets. Results are averaged from all our test models.

data	baseline	*not in sample* base	*not in sample* LCA	*in sample* base	*in sample* LCA	*merge* base	*merge* LCA
CoQA	0.717	0.466	0.588	0.745	0.748	0.780	**0.788**
NaturalQA	0.640	0.420	0.612	0.645	0.673	0.697	**0.703**
SCiQ	0.691	0.559	0.733	0.692	0.793	0.764	**0.794**
TriviaQA	0.648	0.595	0.789	0.659	0.759	0.786	**0.818**
SVAMP	0.617	0.536	**0.864**	0.566	0.681	0.839	0.840
avg	0.663	0.515	0.717	0.661	0.731	0.773	**0.789**

pled space, its label possesses higher representativeness. The AUROC of randomly selected labels surpasses the baseline but remains significantly lower than the highest score, which indirectly supports our hypothesis that randomly picked labels are less robust as representatives of the set. Furthermore, when integrating the greedy decoded answer with the sampled results, the performance exceeds that of randomly picked labels but slightly falls short of SE_{sample}, indicating that the greedy decoded answer is not always the most probable one. We provide a probabilistic analysis of how it impacts the results in Appendix E.1.

We also evaluated the improvements brought by our method when the labeled answer is either included in or excluded from the sample set, across different data sets Table 5 presents a comparison result using SE as a backbone method. Our method consistently outperforms baselines in both scenarios to varying degrees. Furthermore, in the scenario where the greedy answer is semantically integrated into the sample set, we still achieves a 1.6% increase in the score compared to the baseline (refer to Appendix F for more data). This demonstrates that even when we group the labeled answer semantically to enhance the entropy representiveness, the confidence of label still need to be concerned about. As SVAMP is harder, models tend to be wrong even when label probability is high, and the correct answer of this type of problem tends to come from the beam search sampling. After merging it into the sample, the entropy value is reduced, resulting in the correct answer result being opposite to the label. It shows that the label selection strategy is also an issue worthy of attention.

8 Ablation Study

8.1 Number of Generation

The impact of the number of samples on the performance of our method with LNPE, SE and SAR methods as backbone is illustrated in Figure 3(a). Even though the SAR method significantly surpass others, we get higher scores. Taking the performance of the OPT-2.7 model on the NaturalQA (NQ) dataset as an example, the AUROC increases with the number of samples, reaches its peak and

Table 6: The performance of KLD-based method and R-KLD-based method on each backbone. All the results are obtained by averaging results of all models on all datasets.

backbone	baseline	KLD	R-KLD	SAD
LNPE	0.6568	**0.7874**	0.6856	0.4096
SE	0.6711	**0.7690**	0.6018	0.6607
TOKENSAR	0.6745	**0.7420**	0.6553	0.6235
SAR	0.6778	**0.7364**	0.6363	0.6711
avg	0.6701	**0.7587**	0.6447	0.5912

stabilizes with more samples and almost constant diversity, which is similar to results proposed by SE. These results suggest that further optimizing the model's decoding strategy to enhance its diversity could potentially improve the method's performance.

8.2 Sensitivity to Rouge-L Threshold

We use the mean of all experimental results to show the effect of the change in ROUGE-L threshold on the performance of KLD-based method in Figure 3(b). As the Rouge threshold increases, the correctness judgment becomes more stringent. Our experimental results show that as the Rouge-L threshold increases, the performance of different methods in judging model uncertainty increases accordingly. Across all thresholds our methods are always better than the baselines.

8.3 Temperature

We show the effect of temperature on performance in Figure 3(c). Following SE, we conduct experiments on TriviaQA using the Llama2-7B. A smaller temperature will make the token probability sharper and reduce the diversity of model generation. As the temperature increases, after the temperature exceeding 0.5, the performance of the model decreases as the temperature increases. We speculate that this is because although the model diversity has increased, the difference between tokens in vocabulary, thus the probability divergence of the final sampling set and greedy decoding results has become flatter and more difficult to distinguish.

8.4 Different Integrate Methods

We compare the use of KL-divergence (KLD) with methods that use sample average deviation (SAD) (Rivera et al., 2024) and Reverse KL-divergence (R-KLD) (Malinin and Gales, 2019) as aggregation methods, where:

$$\text{Differ}_{SAD}(\mathcal{S}, \mathcal{G}) = |\tilde{P} - P_{\mathcal{G}}|, \tag{6}$$

$$\text{Differ}_{R-KLD}(\mathcal{S}, \mathcal{G}) = P_{\mathcal{G}} \log \frac{P_{\mathcal{G}}}{\tilde{P}}. \tag{7}$$

Our results, shown in Table 6 results indicate that when we treat the sampling results as the "correct" distribution and view greedy sampling as the prediction, divergence calculations help us better identify when the model is more likely to be able to answer. However, with R-KLD, it shows a poor simulator of the actual distribution, only winning in LNPE. As for SAD, it shows that directly comparing the probabilities would even mislead our classification in LNPE.

8.5 Effectiveness on Multi-fact Generation

Multi-fact generation tasks represent a common category within natural language generation (NLG). To evaluate the performance of LCA method on such tasks, we took summarization task as a representative. We utilized the Llama3-8B model to conduct experiments on the XSum (Narayan et al., 2018) dataset. Generations with ROUGE-L greater than the threshold will be assigned a label of 1, otherwise it will be assigned a label of 0. The results of these experiments are presented in Table 7. Our LCA consistently enhances performance across various methods, achieving a maximum improvement of 0.09 on LNPE backbone. Notably, the method of LNPE performs the best. We attribute this to the presence of multiple facts in the generated text. Specifically, sequence-level clustering employed by other semantic-level methods tends to overlook the independence of individual facts within generations.

Table 7: Results of Llama3-8B on Xsum under different Rouge Threshold.

Thres	**SE**		**PE**		**TokenSAR**		**SAR**	
	base	LCA	base	LCA	base	LCA	base	LCA
0.3	0.529	**0.555**	0.543	**0.614**	0.529	**0.557**	0.527	**0.543**
0.2	0.499	**0.531**	0.525	**0.616**	0.500	**0.532**	0.502	**0.521**
0.15	0.517	**0.552**	0.548	**0.636**	0.518	**0.552**	0.514	**0.536**

9 Conclusion

In this paper, we reveal the impact of biases between label sources and samples in uncertainty estimation and propose our LCA method to aggregate the confidence of them. Results demonstrate that our method surpasses the state-of-the-art performance. Further ablation results show the impact of various parameters on method performance.

10 Limitations

We recognize that there are several areas where our approach can be further enhanced: (1) Model Capability: In Section 6, we utilized Roberta to assess semantic relevance. Employing a more powerful model, or fine-tuning Roberta specifically on the test domain, could yield superior sampling results for semantic clustering and would significantly boost the performance of our uncertainty measurement. (2) Similarity Calculation in Multi-fact Scenarios: Our experiments on the xsum dataset reveal that sequence-level similarity calculations can detract from the method's performance in multi-fact contexts. Implementing more refined similarity calculations in these scenarios would likely enhance overall model performance.

11 Ethics Statement

In our research and experimental endeavors, we uphold rigorous ethical standards to ensure that our development and application of artificial intelligence technology are conducted responsibly. Throughout our research process, we have avoided using data that relies on personal information or manual annotations. Additionally, we have utilized open-source models for our experiments without any additional training, thereby ensuring that we do not introduce bias or other harmful knowledge into them. We have also made our code and data publicly available on GitHub. We hope this transparency allows the community to verify the performance of our proposed method and to further enhance it.

References

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. *arXiv preprint arXiv:2402.11756*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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.

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

Haixia Han, Tingyun Li, Shisong Chen, Jie Shi, Chengyu Du, Yanghua Xiao, Jiaqing Liang, and Xin Lin. 2024. Enhancing confidence expression in large language models through learning from past experience. *arXiv preprint arXiv:2404.10315*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Andrey Malinin and Mark Gales. 2019. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in neural information processing systems*, 32.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. *arXiv preprint arXiv:1808.10006*.

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. *arXiv preprint arXiv:1808.08745*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

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.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. *arXiv preprint arXiv:2401.08694*.

Christian Robert. 2014. Machine learning, a probabilistic perspective.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Results For Preliminary Experiments

we will further show the total distributions of models about the num of greedy answer is in/not in sample set and the ratio value on our evaluation datasets. As shown in Table 8, we see that there are 53% questions with, 47% questions without greedy decoded answers in their sample sets, suggesting that our multinomial beam search sampling can search a larger retrieval space. On the other hand, it also shows that our greedy decoding answer is not the maximum decoding probability in a broad sense. We may need to choose a better decoding result as our label source. Such a distribution has obvious deviations in different data and different models. For example, the overall in_rate is higher in the CoQA and NaturalQA datasets, indicating that the diversity of answers in these two datasets is relatively small, and most of the greedy decoded results in the overall sampling space belong to relatively high probability answers. However, the TriviaQA, SciQ and SVAMP datasets show the opposite result, that is, the answer diversity of these questions is relatively large. In this case, we are often more likely to find the correct answer to the problem in the sampling set. For example, in the SVAMP dataset, the accuracy of the label source is low.

Table 8: Distributions about whether greedy decoded answer is in sample set.

data	model	In_num	NotIn_num	In_rate
CoQA	Falcon-7B	5614	2369	0.71
	Llama2-7B	7103	880	0.89
	Mistral-7B	3275	4708	0.41
	OPT-2.7B	6562	1421	0.82
	OPT-13B	6913	1070	0.87
NaturalQA	Falcon-7B	3450	160	0.96
	Llama2-7B	3524	86	0.98
	Mistral-7B	3519	91	0.97
	OPT-2.7B	3086	524	0.85
	OPT-13B	3400	210	0.94
SciQ	Falcon-7B	177	823	0.18
	Llama2-7B	172	828	0.17
	Mistral-7B	284	716	0.28
	OPT-2.7B	39	961	0.04
	OPT-13B	56	944	0.06
SVAMP	Falcon-7B	280	717	0.28
	Llama2-7B	110	887	0.11
	Mistral-7B	141	856	0.14
	OPT-2.7B	122	875	0.12
TriviaQA	Falcon-7B	2769	5234	0.35
	Llama2-7B	1602	6401	0.20
	Mistral-7B	2155	5848	0.27
	OPT-2.7B	1348	6655	0.17
	OPT-13B	1083	6920	0.14
Total		56784	51181	0.53

B Details Of Baselines

B.1 TOKENSAR

TokenSAR considers the different semantic importances of tokens during generation, adjusting the contribution of different tokens in the overall sentence probability. This importance is measured by the similarity between the token and the sentence. That is:

$$W(s_{i,j}, s_i, x) = 1 - |g(x \bigcup s_i, x \bigcup s_i \backslash s_{i,j})|, \tag{8}$$

with $g(,)$ calculates the similarity before and after removing the corresponding token $s_{i,j}$. The more relevant the token, the greater the semantic change it will cause, thus assigning it a higher weight. The uncertainty measure of the entire sentence becomes:

$$\begin{aligned} \text{TOKENSAR}_{s_i} = \sum_{j=0}^{|s_i|} - \log P(s_{i,j}|x, s_{i,<j}) \cdot (1 \\ - |g(x \bigcup s_i, x \bigcup s_i \backslash s_{i,<j})|) \end{aligned} \tag{9}$$

B.2 SAR

The SAR method combines TokenSAR and SentSAR, where SentSAR considers the relevance of individuals within the beam search set to others, calculated by their similarity $\text{sim}(s_i, s_j)$, shown below:

$$\begin{aligned} \text{SentSAR}_{s_i} &= -\log(P(s_i|x) \\ &+ \frac{1}{t} \sum_{j=0}^{|s|\ \&\ j!=i} g(s_i, s_j) P(s_j|x)), \end{aligned} \tag{10}$$

with t as a hyperparameter for temperature. Replace $P(s_i|x)$ in SentSAR_{s_i} with $e^{-\text{TOKENSAR}_{s_i}}$, and we will get SAR_{s_i}.

C Details Of Models

To enhance the generalizability of our experimental results, we employ a diverse range of models, spanning from 2.7B to 13B parameters, including both pre-trained and instruction-tuned variants. Building upon models used in prior studies, we select the OPT, Falcon, Mistral, and Llama series for our evaluation. Specifically, we test pre-trained models such as OPT-2.7B, OPT-13B, and Llama2-7B, as well as instruction-tuned models like Mistral-7B and Falcon-7B. No additional fine-tuning on evaluation datasets is applied to these models.

D Details Of Datasets

D.1 CoQA

CoQA is a dialogue comprehension dataset spanning multiple domains, with each entry comprising a story relevant to the posed questions as well as multi-turn human conversations. We conduct inference tests on the entire validation set, which includes 500 dialogues and a total of 7,983 questions. For each question, we concatenate the background story and the conversation history, which serves as a reference for the model's responses, in the following format:

```
[The Provided Background Story]
[History Conversations]
Q: [Question for the model]
```

D.2 SciQ

SciQ is a question-answering dataset focused on the scientific domain, aiming at improving the performance of natural language models in science-related tasks. We perform inference tests on the entire validation set, which includes a total of 1,000 questions.

D.3 TriviaQA

TriviaQA is an open-domain, closed-book question-answering dataset that spans a broad spectrum of topics and knowledge areas. We utilize the Question-Answer pairs, where the questions can be answered by the model without access to the associated documents. From the TriviaQA validation set, which consists of 17,944 entries, we select about 8,000 for evaluation to maintain consistency in dataset size with COQA.

Following the SE paper, we evaluate SciQ and TriviaQA using a 10-shot prompt format, constructed from 10 randomly selected questions from the validation set. Below is an example:

```
This is a bot that correctly answers
questions.
Question: {Question1} Answer: {Answer1}
Question: {Question2} Answer: {Answer2}
Question: {Question3} Answer: {Answer3}
Question: {Question4} Answer: {Answer4}
Question: {Question5} Answer: {Answer5}
Question: {Question6} Answer: {Answer6}
Question: {Question7} Answer: {Answer7}
Question: {Question8} Answer: {Answer8}
Question: {Question9} Answer: {Answer9}
Question: {Question10} Answer: {Answer10}
Question: {Question for model} Answer:
```

D.4 Natural Questions

Natural Questions (NaturalQA) is an open-domain question-answering dataset derived from real user queries entered into a search engine, providing a closer approximation to real-world scenarios. We utilize NQ-Open, a simplified derivative of the original dataset, and conduct testing on the entire validation set, comprising 3,610 questions. We construct a 2-shot prompt using two randomly selected examples, with the data formatted as follows:

```
Answer these questions:
Question: What is the capital city of
Australia?
Answer: The capital city of Australia
is Canberra.
Question: Who painted the famous artwork
"Starry Night"?
Answer: "Starry Night" was painted by
Vincent van Gogh.
Question: {Question for model}?
Answer:
```

D.5 SVAMP

SVAMP is a dataset designed for mathematical reasoning tasks, requiring models to comprehend and solve math problems described in natural language. This dataset is specifically created to challenge models with complex reasoning, testing their ability to perform multi-step arithmetic operations accurately. SVAMP also features problems with varying levels of difficulty, making it a comprehensive benchmark for evaluating the mathematical reasoning capabilities of natural language models. We randomly select 3 problems from the validation set to construct a 3-shot prompt, which is then used to evaluate 997 test questions. Below is an example:

```
Q: Winter is almost here and most
animals are migrating to warmer
countries. There are 41 bird families
living near the mountain. If 35 bird
families flew away to asia and 62 bird
families flew away to africa How many
more bird families flew away to africa
than those that flew away to asia? A:
27 Q: Paige raised 7 goldfish and 12
catfish in the pond but stray cats loved
eating them. Now she has 15 left. How
many fishes disappeared? A: 4 Q: Marco
and his dad went strawberry picking.
Together they collected strawberries
that weighed 22 pounds. On the way back
Marco'dad found 30 more pounds of
strawberries. Marco's strawberries now
weighed 36 pounds. How much did his dad'
s strawberries weigh now? A: 16 Q: Debby
bought 200 water bottles and 256 soda bottles
when they were on sale. If she drank 312
water bottles and 4 soda bottles a day How
```

Table 9: Uncertainty estimation AUROCs for experiments that contain the greedy decoded answer within the sample set.

data	model	PE		SE		TOKENSAR		SAR	
		base	LCA	base	LCA	base	LCA	base	LCA
COQA	Falcon-7B	0.7534	0.6898	0.7394	0.7352	0.7330	0.7257	0.7457	0.7420
	Llama2-7B	0.7417	0.8073	0.7305	0.7861	0.7160	0.7776	0.7343	0.7823
	Mistral-7B	0.7723	0.7517	0.7720	0.7954	0.7632	0.7829	0.7742	0.7889
	OPT-13B	0.7270	0.6898	0.7244	0.7242	0.7230	0.7213	0.7359	0.7358
NaturalQA	Falcon-7B	0.4696	0.5255	0.5786	0.5891	0.5912	0.6067	0.5947	0.6067
	Llama2-7B	0.5609	0.6110	0.6436	0.6609	0.6382	0.6566	0.6418	0.6583
	Mistral-7B	0.5377	0.6443	0.5683	0.5923	0.5635	0.5860	0.5668	0.5852
	OPT-2.7B	0.7670	0.7499	0.8452	0.8492	0.8620	0.8637	0.8629	0.8671
	OPT-13B	0.7283	0.7536	0.7489	0.7586	0.7568	0.7605	0.7541	0.7575
SciQ	Falcon-7B	0.5871	0.6969	0.6703	0.7098	0.7684	0.7897	0.7766	0.7948
	Llama2-7B	0.5209	0.7249	0.5833	0.6920	0.5989	0.6992	0.6042	0.6996
	Mistral-7B	0.6060	0.7614	0.6914	0.7954	0.7266	0.8219	0.7308	0.8161
	OPT-13B	0.9636	1.0000	0.9091	0.9636	0.9273	0.9818	0.9455	0.9818
SVAMP	Falcon-7B	0.6701	0.6165	0.6752	0.6779	0.6779	0.6785	0.6789	0.6831
	Llama2-7B	0.6566	0.6951	0.5280	0.7612	0.5317	0.7576	0.5335	0.7392
	Mistral-7B	0.5262	0.7084	0.3376	0.6283	0.3259	0.4794	0.3288	0.4193
TriviaQA	Falcon-7B	0.5552	0.6417	0.7117	0.7399	0.7493	0.7708	0.7512	0.7705
	Llama2-7B	0.5383	0.6424	0.6672	0.7155	0.6685	0.7167	0.6682	0.7147
	Mistral-7B	0.6728	0.6728	0.7492	0.7552	0.7555	0.7622	0.7492	0.7586
	OPT-2.7B	0.7010	0.8789	0.7417	0.8268	0.7461	0.8273	0.7484	0.8242
	OPT-13B	0.5453	0.8899	0.7072	0.8305	0.7270	0.8400	0.7301	0.8350
avg		0.6477	0.7215	0.6820	0.7423	0.6929	0.7431	0.6979	0.7410

Table 10: Uncertainty estimation AUROCs for experiments that exclude the greedy decoded answer within the sample set.

data	model	PE		SE		TOKENSAR		SAR	
		base	LCA	base	LCA	base	LCA	base	LCA
COQA	Falcon-7B	0.5251	0.6001	0.5344	0.5478	0.5158	0.5271	0.5056	0.5148
	Llama2-7B	0.4981	0.7238	0.4786	0.5229	0.4857	0.5371	0.4762	0.5171
	Mistral-7B	0.4345	0.8055	0.3834	0.4870	0.3978	0.4853	0.4022	0.4723
	OPT-13B	0.4345	0.5652	0.4388	0.4550	0.4398	0.4557	0.4349	0.4471
NaturalQA	Llama2-7B	0.0941	0.3294	0.0353	0.0706	0.0353	0.1294	0.0235	0.0941
	OPT-2.7B	0.7089	0.8720	0.8765	0.9578	0.9053	0.9674	0.9053	0.9610
	OPT-13B	0.1911	0.7572	0.6106	0.6875	0.6466	0.7212	0.6418	0.7212
SciQ	Falcon-7B	0.4672	0.7672	0.4366	0.6279	0.4643	0.5771	0.4649	0.5665
	Llama2-7B	0.5668	0.8060	0.5990	0.7436	0.5902	0.7404	0.5901	0.7321
	Mistral-7B	0.5264	0.8356	0.5397	0.6884	0.5376	0.6927	0.5351	0.6811
	OPT-13B	0.9343	0.8176	0.4040	0.5779	0.4825	0.5864	0.4931	0.5737
SVAMP	Falcon-7B	0.5855	0.9137	0.5547	0.8304	0.5487	0.8239	0.5495	0.8054
	Llama2-7B	0.5626	0.8936	0.5092	0.8818	0.5078	0.8278	0.5082	0.8134
	Mistral-7B	0.5344	0.7898	0.4668	0.8723	0.4870	0.8528	0.4875	0.8338
TriviaQA	Falcon-7B	0.6246	0.7919	0.6197	0.7191	0.5675	0.6750	0.5619	0.6637
	Llama2-7B	0.5316	0.7729	0.5937	0.7312	0.5898	0.7301	0.5865	0.7207
	Mistral-7B	0.4982	0.8428	0.5098	0.6964	0.5094	0.6949	0.5046	0.6671
	OPT-2.7B	0.6736	0.9288	0.6515	0.8001	0.6617	0.8018	0.6592	0.7923
	OPT-13B	0.5831	0.8228	0.6676	0.7834	0.6663	0.7818	0.6666	0.7774
avg		0.5250	0.7703	0.5216	0.6674	0.5284	0.6636	0.5261	0.6503

`many days would the soda bottles last? A:`

E Probabilistic Analysis

E.1 Merge Greedy Decoded Answer into Samples

We denote the probability of an individual in a sampling set with N samples as P_i and the probability of the greedy decoded answer as P_{greedy}. When considering merging the greedy decoded answer into the sampling set based on semantic similarity, the impact on the overall entropy will differ depending on whether the greedy answer has already appeared in the sampling set. The entropy of the samples can be calculated as:

$$E_{sample} = -\sum_i^N P_i \log P_i, \tag{11}$$

If the greedy answer belongs to $cluster_i$ within the sampling domain, the entropy remains unchanged since the answer has already been sampled, avoiding repeated calculations of the same answer that would bias the entropy value. If the greedy answer is outside the sampling domain, the entropy changes to:

$$E_{sample} = -(\sum_i^N P_{i'} \log P_{i'} + P_{greedy} \log P_{greedy}), \tag{12}$$

where $P_{i'} = \frac{P_i}{\sum_{i=1}^N P_i + P_{greedy}}$ Since $P_{i'} < P_i$, the entropy increases, further widening the gap between the expected probability and the observed value. Thus, when the greedy decoded answer has not appeared in the sampling set, adopting a merging strategy will make the overall distribution more closely approximate the true distribution.

E.2 Gibss Probability and EPKL

Expected Pairwise KL-divergence (EPKL) is another uncertainty measurement that calculate total divergence between each sample from model:

$$\mathrm{EPKL}[y, \theta | x, D] = \mathbb{E}_{q(\theta)q(\tilde{\theta})}\Big[\mathbb{E}_{p(y|x,\theta)}\Big[\ln P(y|x,\theta) - \ln P(y|x,\tilde{\theta})\Big]\Big]. \tag{13}$$

where θ, $\tilde{\theta}$ represent either Bayesian network parameters or randomness injected via Monte Carlo sampling. As mentioned above, we treat Gibss probability and "observed probability" as $P(True)$ and $P'(True)$, standing for confidence level. We use the divergence between distributions of pairwise sampling results as a measure of the network's uncertainty. Instead of calculating the average KL divergence between the set of sampled answers and the labeled answer, denoted as $\frac{1}{|\mathcal{S}|}\sum_i^{|\mathcal{S}|} P_{\mathcal{S}_i} \log \frac{P_{\mathcal{S}_i}}{P_{\mathcal{G}}}$ (Malinin and Gales, 2020), we use "Gibbs Probability". When the number of samples is sufficient, the sum of sample probabilities $\sum P_{\mathcal{S}_i}$ approaches 1, providing the following unbiased estimate:

$$\frac{1}{|\mathcal{S}|}\sum_i^{|\mathcal{S}|} P_{\mathcal{S}_i} \log \frac{P_{\mathcal{S}_i}}{P_{\mathcal{G}}} = \frac{1}{|\mathcal{S}|}(\sum_i^{|\mathcal{S}|} P_{\mathcal{S}_i} \log P_{\mathcal{S}_i} - \sum_i^{|\mathcal{S}|} P_{\mathcal{S}_i} \log P_{\mathcal{G}}) \approx \frac{\sum_1^{|\mathcal{S}|} P_{\mathcal{S}_i}}{|\mathcal{S}|}(\log \tilde{P} - \log P_{\mathcal{G}}) \tag{14}$$

$$\tilde{P} \log \frac{\tilde{P}}{P_{\mathcal{G}}} = \tilde{P}(\log \tilde{P} - \log P_{\mathcal{G}}) \approx \prod_1^{|\mathcal{S}|} P_{\mathcal{S}_i}^{\frac{1}{|\mathcal{S}|}} (\log \tilde{P} - \log P_{\mathcal{G}}), \tag{15}$$

Eq. 15 calculates from a geometric mean perspective integrating information from all sampled answers in one direction, smoothing out some details, making it more suitable for an overall assessment of the entire sampling distribution, while Eq. 14 is based on the arithmetic mean leading to numerical instability when there is significant variance among sample results.

F Results Of Experiments

In our experiments, we present the average performance of different models across three scenarios: when "the greedy decoded answer is present in the sample set", when "the greedy decoded answer is absent from the sample set", and when "the greedy decoded answer is merged into the sample set" across various datasets. In this subsection, we provide a detailed comparison of our LCA method against the baseline in these three scenarios. When we group the data according to the experimental strategy in the paper, in some cases, the AUROC will be 0 because all the answers to the corresponding group of the question are wrong. We remove this part of the data before displaying it, and only display the cases where the AUROC is greater than 0.

F.1 Label Answer In Sample Set

Table 9 presents the AUROC results for the experiment with greedy decoded answer in sample set.

Table 11: Uncertainty estimation AUROCs for experiments that merge the greedy decoded answer into the sample set.

data	model	SE	Merge baseline	Merge LCA
CoQA	Falcon-7B	0.7472	0.7456	0.7402
	Llama2-7B	0.7465	0.8074	0.8178
	Mistral-7B	0.6206	0.8327	0.8573
	OPT-13B	0.7309	0.7343	0.7366
NaturalQA	Falcon-7B	0.5815	0.5899	0.5988
	Llama2-7B	0.6267	0.6572	0.6572
	Mistral-7B	0.5716	0.6050	0.6263
	OPT-2.7B	0.8488	0.8686	0.8609
	OPT-13B	0.7428	0.7617	0.7713
SciQ	Falcon-7B	0.7200	0.7926	0.8143
	Llama2-7B	0.6150	0.7686	0.7923
	Mistral-7B	0.6720	0.8316	0.8496
	OPT-13B	0.6824	0.6633	0.7209
SVAMP	Falcon-7B	0.6701	0.7069	0.7066
	Llama2-7B	0.5319	0.9254	0.9255
	Mistral-7B	0.5734	0.8869	0.8886
TriviaQA	Falcon-7B	0.6902	0.7614	0.7810
	Llama2-7B	0.6336	0.7747	0.8043
	Mistral-7B	0.6189	0.8181	0.8412
	OPT-2.7B	0.7477	0.8015	0.8530
	OPT-13B	0.6897	0.7720	0.8119
avg		0.6696	0.7669	0.7836

In most instances, our LCA method surpasses the baseline method to varying extents, with an improvement of 8% on PE method, 6% on SE, and 5% on TOKENSAR and SAR. Specifically, when using the OPT-13B model on the SciQ dataset, the baseline method achieves an AUROC of 0.9636, while our LCA approach further enhances this to a perfect score of 1. Moreover, it is evident that in most cases, when the label answer is present in the sample set, there is a strong correlation between the entropy value of the set and the final label. Notably, only 4 out of 168 experimental groups exhibit an AUROC below 0.5, which indicates a negative correlation between the entropy value and the classification label. In 3 of these 4 cases, our LCA method successfully corrects these discrepancies, resulting in AUROCs greater than 0.5.

F.2 Label Answer Not In Sample Set

Table 10 presents the AUROC results for the experiment without greedy decode answer in sample set. In this part of the experiment, the AUROC scores are generally low, but our LCA solution can still achieve good performance, improving 25% on the PE solution and 13% on the SE, TOKENSAR and SAR methods. In most cases, the correlation between entropy and corresponding label is low, and in 1/3 of the cases, the AUROCs are lower than 0.5. However, in these serious misclassification cases, 4/5 of which our LCA solutions can optimize and prompt AUROCs to a higher level.

F.3 Merge Label Answer To Sample Set

In Table 11, we show the AUROC changes when the greedy decoded answer is semantically merged into the sampling set, and the AUROCs further increase when our LCA solution is applied on this basis. We can see that except for a slight decrease in the baseline score of OPT-13B on the SciQ dataset, and mostly the correlation between entropy value and labels after merging have been improved, with an overall improvement of 9.7%. Combined with our previous experimental analysis, this is because we have expanded the diversity of the sampling space (because the greedy answer does not appear in the sampling set in half of the cases), and the distributions are closer to the true one. Our LCA method further improves 1.7% on this basis, which is 11.4% higher than the original solution in average. This result shows that label confidence awareness can still play a role when the label answer is merged into the sampling set.

G Additional Overhead

When our solution is integrated on the backbone method, no additional computational overhead is introduced except for calculating the KL divergence.