Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models

Jinhao Duan[1] **Hao Cheng**[3] **Shiqi Wang**[2] **Alex Zavalny**[1] **Chenan Wang**[1]
Renjing Xu[3] **Bhavya Kailkhura**[4] **Kaidi Xu**[1]*

[1]Drexel University [2]AWS AI Lab
[3]Hong Kong University of Science and Technology (Guangzhou)
[4]Lawrence Livermore National Laboratory

Abstract

Large Language Models (LLMs) show promising results in language generation and instruction following but frequently "hallucinate", making their outputs less reliable. Despite Uncertainty Quantification's (UQ) potential solutions, implementing it accurately within LLMs is challenging. Our research introduces a simple heuristic: not all tokens in auto-regressive LLM text equally represent the underlying meaning, as "linguistic redundancy" often allows a few keywords to convey the essence of long sentences. However, current methods underestimate this inequality when assessing uncertainty, causing tokens with limited semantics to be equally or excessively weighted in UQ. To correct this, we propose **S**hifting **A**ttention to more **R**elevant (*SAR*) components at both token- and sentence-levels for better UQ. We conduct extensive experiments involving a range of popular "off-the-shelf" LLMs, such as Vicuna, WizardLM, and LLaMA-2-chat, with model sizes extending up to 33B parameters. We evaluate various free-form question-answering tasks, encompassing domains such as reading comprehension, science Q&A, and medical Q&A. Our experimental results, coupled with a comprehensive demographic analysis, demonstrate the superior performance of *SAR*. The code is available at https://github.com/jinhaoduan/SAR.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities in multi-round conversation (Long, 2023; Chen et al., 2023), logical reasoning (Creswell et al., 2022; Pan et al., 2023; Duan et al., 2024), and also disclose great potential in scientific discovery (Birhane et al., 2023). For instance, ChatGPT, BARD, GPT-4, pre-trained on large-scale corpora and carefully aligned to human preferences (Christiano et al., 2017; Ouyang et al., 2022), profoundly shape the range of what AIs could do, and how they communicate with humans.

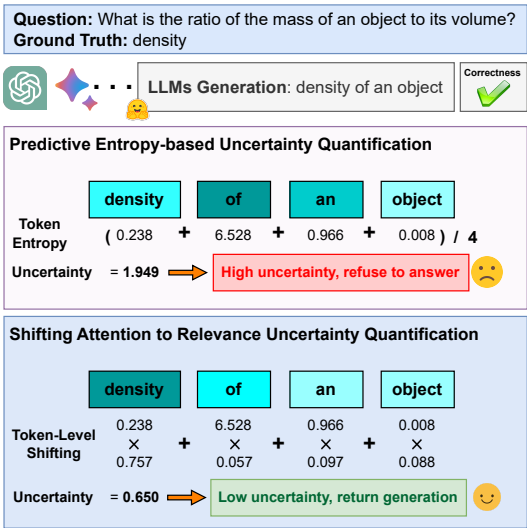

Figure 1: Irrelevant tokens (or sentences) may commit majority uncertainty in free-form generations, such as the token "of" committing extremely large uncertainty misleads the uncertainty quantification of LLMs. We term these observations as generative inequalities and tackle them by shifting attention to more relevant components.

Despite the surprising progress, LLMs are proven to be vulnerable to widely known reliability issues (Yao et al., 2024; Sun et al., 2024; Hong et al., 2024), such as hallucination (Manakul et al., 2023a) and factual errors (Bian et al., 2023; Karpinska and Iyyer, 2023; Gekhman et al., 2023). Uncertainty quantification (UQ) is one of the most popular approaches to answering when humans can trust the generations of LLMs, which is critical for Human-AI interaction applications (e.g., therapy and mental health (Lin et al., 2023; Sharma et al., 2023)) where humans need to densely communicate with LLMs. In these applications, the resulting behaviors will be largely affected by the generations from LLMs.

Unfortunately, UQ remains challenging due to various uncertainty sources (e.g., aleatoric uncer-

* Corresponding author: Kaidi Xu <kx46@drexel.edu>.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5050–5063

tainty and epistemic uncertainty (Kendall and Gal, 2017)). This challenge is particularly pronounced in the context of free-form LLMs, which are characterized by high complexity and an essentially limitless solution space—any output matching the semantic content of the true answer is considered correct. This makes UQ in LLMs markedly distinct from more traditional classification models or models with defined labels, where the solution space is constrained.

Prior works in this direction estimate uncertainty by prompting LLMs to answer confidence (Lin et al., 2022a; Kadavath et al., 2022a) or designing logits- or entropy-based measurements (Malinin and Gales, 2021, 2020; Kuhn et al., 2023). The most recent work proposes *Semantic Entropy* (*SE*) (Kuhn et al., 2023) where generations sharing the same meaning are gathered in a semantic cluster. Then the cluster-wise entropy is calculated as the uncertainty measurement.

Our motivation is derived from an intuitive fact: *tokens are created unequally in presenting semantics*. Namely, some tokens (e.g., nouns, verbs) are more meaningful than other tokens (e.g., definite articles). For example, for a given question "*What is the ratio of the mass of an object to its volume?*" and a model generation "*density of an object*", "density" is the most relevant token in presenting semantics than the rest tokens. We term the former as *relevant tokens* and the rest tokens as *irrelevant tokens*. Prior works treat each token equally when estimating uncertainty, which is counter-intuitive (Figure 1). Therefore, we ask:

Are relevant tokens more critical than irrelevant tokens in uncertainty quantification?

To answer this question, we first investigate how token-level generative inequality affects uncertainty quantification in LLMs. Specifically, we first measure the *relevance score* of each token by comparing the semantic change before and after removing this token from the sentence. A larger semantic change means more relevance for this token and vice versa. Then we quantify the *uncertainty proportions*, i.e., the uncertainty committed by this token. At last, we analyze the correlation between relevance and uncertainty proportion. Our results reveal that large amounts of tokens containing very limited semantics are weighted equally or even heavily in UQ. Similar observations are also observed when generalizing to the sentence-level inequality by assessing relevant sentences and irrelevant sentences.

Based on these observations, we propose a simple attention-shifting method, by jointly examining the relevance of each component and reassigning its attention, from both the token level and the sentence level, termed as **S**hifting **A**ttention to **R**elevance (*SAR*). *SAR* is evaluated on multiple popular instruction-tuned LLMs (e.g., Vicuna (Zheng et al., 2023), LLaMA-2-chat (Touvron et al., 2023b), WizardLM (Xu et al., 2023)), with model size up to 33B, and popular pre-trained LLMs (e.g., OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023a)) with model sizes up to 30b, over cross-domain free-form question-answering tasks, such as the conventional NLP domain (e.g., CoQA (Reddy et al., 2019), TriviaQA (Joshi et al., 2017) and SciQ (Welbl et al., 2017)) and medical domain (e.g., MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022)). Experimental results demonstrate *SAR*'s superior performance. Our contributions can be summarized as the following:

- We disclose that uncertainty quantification is significantly affected by token- and sentence-level generative inequality, i.e., irrelevant tokens or sentences might be over-valued when estimating uncertainty.
- We mitigate the two inequality biases by **S**hifting **A**ttention to **R**elevance (*SAR*), which jointly examines the relevance of each token and sentence, and reassigns attention when estimating uncertainty.
- We conduct experiments over "off-the-shelf" instruction-tuned LLMs and popular pre-trained LLMs, across various free-form question-answering tasks. Experimental results demonstrate that *SAR* outperforms previous state-of-the-art by a large margin.

2 Related Works

Uncertainty Quantification in Conventional NLP Tasks. Uncertainty Quantification of machine translation (MT) has been studied for years to evaluate the performance of MT better. (Ott et al., 2018) access uncertainty by comparing multiple model outputs to multiple references with inter-sentence BLEU. (Glushkova et al., 2021) measure uncertainty through techniques of Monte Carlo dropout (Gal and Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017). (Fomicheva et al., 2020) use uncertainty quantification methods to improve probability estimates in

neural networks. (Lahlou et al., 2021) proposed Direct Epistemic Uncertainty Prediction, a model-agnostic framework, for estimating epistemic uncertainty in machine learning models. For regression tasks, (Wang et al., 2022) use uncertainty quantification to address both data uncertainty and model uncertainty, and (Malinin et al., 2020) proposes a method for uncertainty quantification using Prior Networks to obtain interpretable measures of uncertainty at a low computational cost. For Natural Language Understanding tasks, (Talman et al., 2023) use uncertainty quantification by applying Bayesian uncertainty modeling using Stochastic Weight Averaging-Gaussian.

Uncertainty Quantification in LLMs. Although uncertainty quantification has been thoroughly examined in models with distinct labels, such as classification models (Ulmer et al., 2022; Vazhentsev et al., 2022), it is still under-explored for popular free-form LLMs, e.g., GPT (Radford et al., 2019), OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023a). These models present a unique challenge in uncertainty quantification as their solution domains are flexible and effectively infinite, i.e., any generation can be deemed correct as long as the semantics align consistently with the real answer.

(Xiao et al., 2022) conducts large-scale empirical evaluations on how the configuration (e.g., model size, architecture, training loss) of LLMs affect uncertainty. (Lin et al., 2022a; Kadavath et al., 2022a) propose to quantify uncertainty by directly prompting the language models to answer the uncertainty with respect to their generations. (Manakul et al., 2023b) measures the faithfulness of generations by quantifying the consistency of generations, i.e., generations should be consistent if the model really captured the concept. (Malinin and Gales, 2021) examines the uncertainty of free-form LLMs by calculating the accumulative predictive entropies over multiple generations. Recently, Semantic Entropy (*SE*) (Kuhn et al., 2023) is presented to tackle the "semantic equivalence" difficulty in uncertainty quantification. *SE* gathers generations sharing the same semantics into clusters and performs cluster-wise predictive entropy as the uncertainty measurement.

We aim to design metrics from multiple generations to characterize the uncertainty of LLMs. Our work focuses on the token- and sentence-level generative inequalities, which are not explored by prior works in uncertainty quantification.

3 Generative Inequality in Uncertainty Quantification

Tokens are created unequally in reflecting the meaning of the generation yet they are treated equally when estimating uncertainty. We term these inequalities as *generative inequalities* and investigate how they affect uncertainty quantification.

3.1 Preliminaries

LLMs normally output generations in a free-form and auto-regressive manner, i.e., progressively predicting the probability distribution of the next token. We denote by $\boldsymbol{x}$ the input (or the prompt) and $\boldsymbol{s}$ the sentence consisting of N tokens. Here, we take a sentence $\boldsymbol{s}$ as a completion regarding prompt $\boldsymbol{x}$. Then, for a given LLM, the probability of generating z_i as the i-th token can be described as $p(z_i|\boldsymbol{s}_{<i}, x)(1 \leq i \leq N)$, where $\boldsymbol{s}_{<i}$ refers to the previously generated tokens $\{z_1, ..., z_{i-1}\}$.

Baseline. We use the popular Predictive Entropy (*PE*), described in (Kadavath et al., 2022b), as the baseline and investigate how it is affected by generative inequalities in this section. The Predictive Entropy (*PE*) is defined as the entropy over the whole sentence $\boldsymbol{s}$:

$$PE(\boldsymbol{s}, \boldsymbol{x}) = -\log p(\boldsymbol{s}|\boldsymbol{x}) = \sum_i -\log p(z_i|\boldsymbol{s}_{<i}, \boldsymbol{x}). \quad (1)$$

It can be interpreted as the accumulation of the token-wise entropy.

3.2 Token-Level Generative Inequality

Generative inequality refers to an observation: tokens containing limited semantics are equally valued when estimating the uncertainty of a sentence, which is counter-intuitive. To outline this, we specify two quantities for each token: how much semantics the token contains, i.e., the *relevance*, and how much uncertainty the token committed, i.e., the *uncertainty proportion*.

For a given prompt $\boldsymbol{x}$ and the sentence $\boldsymbol{s}$ consisting of N tokens, i.e., $\boldsymbol{s} = \{z_1, z_2, ..., z_N\}$:

Relevance. To measure how important z_i is in reflecting the semantics of $\boldsymbol{s}$, we compare the semantic change before and after removing this token:

$$R_T(z_i, \boldsymbol{s}, \boldsymbol{x}) = 1 - |g(\boldsymbol{x} \cup \boldsymbol{s}, \boldsymbol{x} \cup \boldsymbol{s} \setminus \{z_i\})|, \quad (2)$$

where $g(\cdot, \cdot)$, calculating the similarity between two sentences on a scale of 0 to 1, can be any semantic similarity measurement. In our experiments, we leverage the Cross-Encoder (Reimers

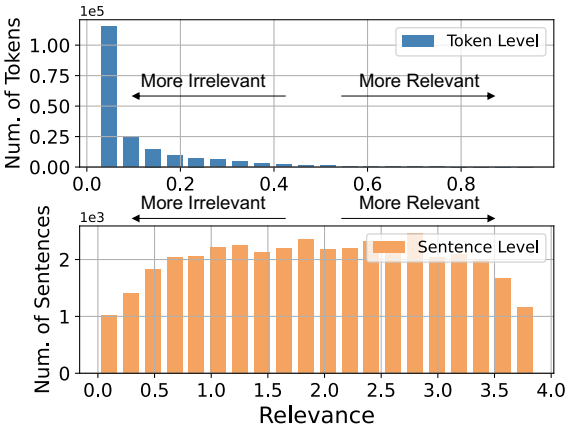

Figure 2: Distributions of relevance scores in both token-level and sentence-level situations. It is shown that irrelevant tokens and sentences take considerable proportions.

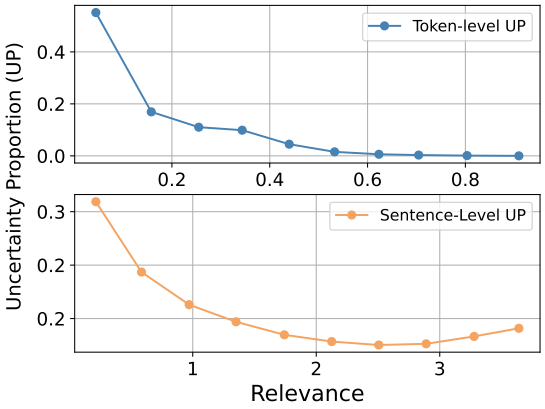

Figure 3: Correlations between relevance scores and uncertainty proportions in both token-level and sentence-level situations. Irrelevant tokens and sentences dominate the total volume of uncertainty quantification.

and Gurevych, 2019a)-RoBERTa-large (Liu et al., 2019) as this measurement since it is one of the most powerful sentence similarity evaluation models provided by the popular SentenceTransformers Library (Reimers and Gurevych, 2019b). Generally, larger $R_T(z_i, \boldsymbol{s}, \boldsymbol{x})$ means removing z_i will lead to significant semantic changing, indicating that z_i is more relevant.

Uncertainty Proportion. To measure the proportion of uncertainty committed by z_i, we simply derive the ratio from Eq. (1):

$$UP_T(z_i, \boldsymbol{s}, \boldsymbol{x}) = \frac{-\log p(z_i|\boldsymbol{s}_{<i}, \boldsymbol{x})}{PE(\boldsymbol{s}, \boldsymbol{x})}. \quad (3)$$

Larger $UP_T(z_i, \boldsymbol{s}, \boldsymbol{x})$ means z_i commits more uncertainty when estimating the uncertainty of sentence $\boldsymbol{s}$; vice versa.

3.3 Sentence-Level Generative Inequality

It has been widely shown that involving multiple sentences benefits estimating uncertainty (Kadavath et al., 2022b). For instance, *PE* will usually be the arithmetic mean of multiple sentences in practice, i.e., $\frac{1}{K}\sum_k PE(\boldsymbol{s}_k, \boldsymbol{x})\,(1 \le k \le K)$ where $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_K\}$ consisting of K sentences regarding $\boldsymbol{x}$ and $\boldsymbol{s}_k \in S$ is the k-th sentence. Following Section 3.2, for a given sentence $\boldsymbol{s}_i$, we define the sentence-level relevance of $\boldsymbol{s}_i$ as the probability-weighted semantic similarity with other sentences:

$$R_S(\boldsymbol{s}_i, S, \boldsymbol{x}) = \sum_{j=1, j\neq i} g(\boldsymbol{s}_i, \boldsymbol{s}_j) p(\boldsymbol{s}_j|\boldsymbol{x}), \quad (4)$$

where $1 \le i, j \le K$ and $p(\boldsymbol{s}_j|\boldsymbol{x})$ is the generative probability of $\boldsymbol{s}_j$. It is out of an intuitive assumption that sentences are more convincing if they are semantically consistent with other sentences. Namely, a sentence that is semantically close to other sentences is considered more representative. Besides, the generative probability $p(\boldsymbol{s}_j, \boldsymbol{x})$ provides more confidence to $\boldsymbol{s}_j$ when measuring relevance, i.e., higher $p(\boldsymbol{s}_j, \boldsymbol{x})$ makes $\boldsymbol{s}_j$ more acceptable.

Similar to the token-level situation, the sentence-level uncertainty proportion of $\boldsymbol{s}_i$ is defined as:

$$UP_S(\boldsymbol{s}_i, S, \boldsymbol{x}) = \frac{PE(\boldsymbol{s}_i, \boldsymbol{x})}{\sum_k PE(\boldsymbol{s}_k, \boldsymbol{x})}, \quad (5)$$

where $1 \le k \le K$. It is the proportion of uncertainty committed by $\boldsymbol{s}_i$,

3.4 Analytical Insights

We leverage the defined *relevance* and *uncertainty proportion* to characterize the generative inequality observations in this section. We utilize CoQA as the dataset and OPT-13b as the model to be examined. For each prompt in CoQA, we generate 10 sentences, i.e., $K = 10$ in Eq. (4) and Eq. (5). More details of generative hyper-parameters can be found in Appendix A.

We first quantify the histograms of token-level relevance scores and sentence-level relevance scores. Results are summarized in Figure 2. For token-level relevance, it is clear that most of the tokens are irrelevant tokens, i.e., low relevance scores, indicating that linguistic redundancy exists widely. In terms of the sentence-level situation, although the distribution is smoother than the token-level situation, the irrelevant sentences still take a considerable amount over all the sentences.

We further investigate the correlations between relevance and uncertainty proportions, i.e., how much uncertainty is committed by tokens and sentences under various relevance scores. Specifically, we first group tokens and sentences into 10 bins with uniform relevance ranges and then average/sum the uncertainty proportions committed by tokens or sentences grouped in the same bin. Since irrelevant tokens take majority proportions over all the tokens, averaging the uncertainty proportions in each bin may hide the real effect of irrelevant tokens. Therefore, we report the sum of uncertainty proportions in each bin in the token-level situation. Results are summarized in Figure 3.

It is clear that irrelevant tokens/sentences commit significantly more uncertainty than relevant sentences in both token-level and sentence-level situations. These observations demonstrate the existence of sentence inequalities and also the uncertainty quantification is highly affected by these inequalities.

4 Shifting Attention to Relevance

A natural hypothesis derived from Section 3.4 is that shifting the attention to those relevant components may benefit uncertainty quantification. In this section, we introduce the proposed Shifting Attention to Relevance (*SAR*) in detail.

4.1 Notations

We reuse the notations defined in Section 3.1 where we denote by $\boldsymbol{x}$ the prompt and S the K sentences regarding $\boldsymbol{x}$. There will be N_j tokens for each sentence $\boldsymbol{s}_j \in S\,(1 \leq j \leq K)$.

4.2 Relevance Discovery and Shifting

SAR corrects generative inequalities by reviewing the relevance of each token and/or sentence and emphasizing uncertainty quantification attention to those more relevant components. Here we introduce token-level shifted measurement and sentence-level shifted measurements:

Token-Level Shifting. For a sentence $\boldsymbol{s}_j$ regarding prompt $\boldsymbol{x}$, $\boldsymbol{s}_j = \{z_1, z_2, ..., z_{N_j}\}$ contains N_j tokens. We first calculate the normalized relevance score for each token $z_i\,(1 \leq i \leq N_j)$ based on Eq. (2), i.e., $R_T(z_i, \boldsymbol{s}_j, \boldsymbol{x})$:

$$\tilde{R}_T(z_i, \boldsymbol{s}_j, \boldsymbol{x}) = \frac{R_T(z_i, \boldsymbol{s}_j, \boldsymbol{x})}{\sum_n^{N_j} R_T(z_n, \boldsymbol{s}_j, \boldsymbol{x})} \tag{6}$$

Then we enlarge the uncertainty proportions of relevant tokens by re-weighting token entropy according to their normalized relevance scores:

$$E_T(z_i, \boldsymbol{s}_j, \boldsymbol{x}) = -\log p(z_i|\boldsymbol{s}_{<i}, \boldsymbol{x})\tilde{R}_T(z_i, \boldsymbol{s}_j, \boldsymbol{x}). \tag{7}$$

The token-level shifted predictive entropy defined over $\boldsymbol{s}_j$ can be formulated as:

$$\text{TOKEN}\textit{SAR}(\boldsymbol{s}_j, \boldsymbol{x}) = \sum_i^{N_j} E_T(z_i, \boldsymbol{s}_j, \boldsymbol{x}). \tag{8}$$

The reason we normalize relevance score in Eq. (6) is two-fold: a) to make tokens comparable across sentences; b) to mitigate the bias posed by the length of sentence, such as the length normalization in Length-normalized Predictive Entropy (*LNPE*) (Malinin and Gales, 2020). In this way, the uncertainty proportions of tokens containing strong relevance will be enlarged when estimating uncertainty.

Sentence-Level Shifting. As mentioned in Section 3.3, sentences that have higher relevance scores, i.e., semantically consistent, are more convincing than others. Therefore, we simply reduce sentence uncertainty by enlarging its generative probability with a relevance-controlled quantity:

$$\begin{aligned} E_S(\boldsymbol{s}_j, S, \boldsymbol{x}) &= -\log(p(\boldsymbol{s}_j|\boldsymbol{x}) + \frac{1}{t}R_S(\boldsymbol{s}_j, S, \boldsymbol{x})) \\ &= -\log(p(\boldsymbol{s}_j|\boldsymbol{x}) + \underbrace{\frac{\sum_{k\neq j} g(\boldsymbol{s}_j, \boldsymbol{s}_k)p(\boldsymbol{s}_k|\boldsymbol{x})}{t}}_{\text{sentence relevance}}), \end{aligned} \tag{9}$$

where $p(\boldsymbol{s}_j|\boldsymbol{x}) = \prod_i p(z_i|\boldsymbol{s}_{<i}, \boldsymbol{x})$ is the generative probability of $\boldsymbol{s}_j$ and t is the temperature used to control the scale of shifting. Then, the sentence-level shifted predictive entropy over K sentences can be formulated as:

$$\text{SENT}\textit{SAR}(S, \boldsymbol{x}) = \frac{1}{K}\sum_k E_S(\boldsymbol{s}_k, S, \boldsymbol{x}). \tag{10}$$

Note that Eq. (9) shares a similar form with *SE* (Kuhn et al., 2023), i.e., reducing the uncertainty of semantically consistent sentences. Differently, *SE* achieves this with bi-directional entailment prediction and we achieve this with weighted relevance scores. With manual examination, we found that around 36.7% of the entailment predictions are undesirable, over the long sentences that have more than 20 tokens on average (120 questions in total). Instead, our SENT*SAR* leverages the more "soft" sentence similarity to calculate the relevance score, which is more desirable for long and complex sentences.

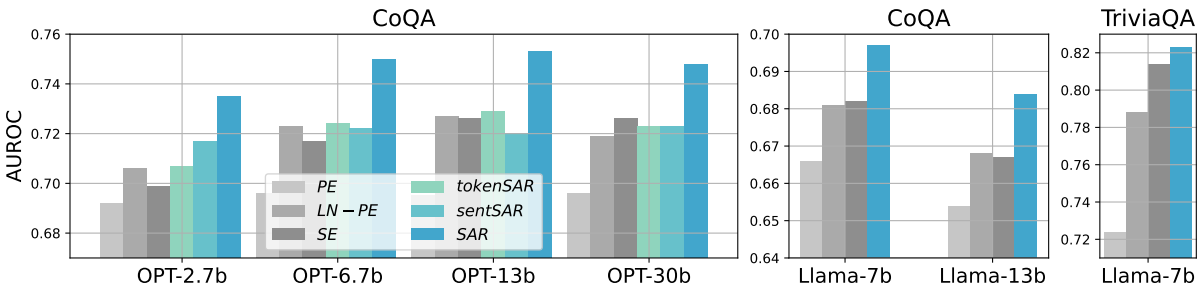

Figure 4: The AUROCs of TOKEN*SAR*, SENT*SAR*, *SAR*, and baseline methods, across various "off-the-shelf" LLMs and datasets (e.g., CoQA, and Trivia QA). Rouge-L with a threshold of 0.5 is used as the correctness metric. The proposed *SAR* substantially outperforms existing methods across all the scenarios.

Models & Datasets	*LS*	*PE*	*LN-PE*	*SE*	TOKEN*SAR* (ΔSE)	SENT*SAR*(ΔSE)	*SAR*(ΔSE)
Vicuna-13b w./ 5 generations are generated for each question							
Trivia QA	0.560	0.690	0.624	0.630	0.692 (+6.2%)	0.745 (+11.5%)	**0.749** (+11.9%)
SciQ	0.589	0.708	0.668	0.675	0.706 (+3.1%)	**0.745** (7.0%)	0.741 (+6.6%)
Vicuna-33b w./ 5 generations are generated for each question							
Trivia QA	0.565	0.644	0.639	0.651	0.652 (+0.1%)	**0.715** (+6.4%)	0.710 (5.9%)
SciQ	0.584	0.665	0.668	0.674	0.665 (-0.9%)	**0.717** (+4.3%)	0.710 (+3.6%)
WizardLM-13b w./ 5 generations are generated for each question							
Trivia QA	0.519	0.647	0.615	0.634	0.657 (+2.3%)	0.743 (+10.9%)	**0.744** (+11.0%)
SciQ	0.574	0.677	0.638	0.649	0.681 (+3.2%)	**0.719** (+7.0%)	0.707 (+5.8%)
LLaMA-2-13b-chat w./ 5 generations are generated for each question							
Trivia QA	0.504	0.647	0.615	0.622	0.654 (+3.2%)	0.698 (+7.6%)	**0.704** (+8.2%)
SciQ	0.578	0.718	0.688	0.692	0.718 (+2.6%)	**0.737** (+4.5%)	0.725 (+3.3%)
Average	0.555	0.675	0.644	0.653	0.678 (+2.5%)	**0.727** (+7.4%)	0.724 (+7.1%)

Table 1: Uncertainty quantification AUROCs of TOKEN*SAR*, SENT*SAR*, *SAR*, and baseline methods, across various instruction-tuned open-source LLMs, over different datasets (e.g., SciQ, and Trivia QA). The threshold of Rouge-L is set to 0.5. Underline means the second best method.

4.3 Overall Measurement

Token-level shifting and sentence-level shifting are conceptually different as they emphasize different perspectives. However, they are orthogonal and can be naturally combined to shift attention from both token-level and sentence-level, resulting in more effective uncertainty quantification. To achieve that, we simply replace the generative probabilities in Eq. (9), i.e., $p(\boldsymbol{s}_i|\boldsymbol{x})$ and $p(\boldsymbol{s}_j|\boldsymbol{x})$, with the token-shifted probability derived from Eq. (8), i.e. $p'(\boldsymbol{s}_i|\boldsymbol{x}) = e^{-\text{TOKEN}SAR(\boldsymbol{s}_i,\boldsymbol{x})}$ and $p'(\boldsymbol{s}_j|\boldsymbol{x}) = e^{-\text{TOKEN}SAR(\boldsymbol{s}_j,\boldsymbol{x})}$:

$$E_{T,S}(\boldsymbol{s}_j, S, \boldsymbol{x}) = -\log(p'(\boldsymbol{s}_i|\boldsymbol{x}) + \frac{\sum_{k\neq j} g(\boldsymbol{s}_j, \boldsymbol{s}_k) p'(\boldsymbol{s}_j|\boldsymbol{x})}{t}). \tag{11}$$

Then the token- and sentence-level shifted predictive entropy over K sentences can be defined as $SAR = \frac{1}{K}\sum_k E_{T,S}(\boldsymbol{s}_k, S, \boldsymbol{x})$.

We denote TOKEN*SAR*, SENT*SAR*, and *SAR* as the token-shifted predictive entropy, sentence-shifted predictive entropy, and both token- and sentence-shifted predictive entropy respectively, in the rest of this paper.

5 Empirical Evaluations

5.1 Experimental Settings

Baselines. We consider 4 baseline methods in our experiments, including Lexical Similarity (Lin et al., 2022b), Semantic Entropy (*SE*) (Kuhn et al., 2023), Predictive Entropy (*PE*) (Kadavath et al., 2022b), and Length-normalized Predictive Entropy (*LN-PE*) (Malinin and Gales, 2020). Lexical Similarity considers the similarities among multiple sentences. *SE* introduces the "semantic equivalence" difficulty in the uncertainty quantification of free-form LLMs and tackles this issue by gathering sentences containing the same meaning into clusters and calculating cluster-wise entropy. *LN-PE* is

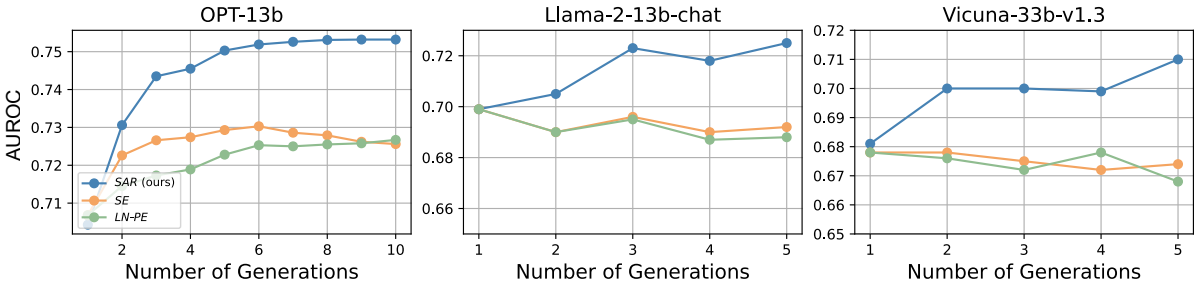

Figure 5: The performance of *SAR* and baseline methods over various numbers of generations.

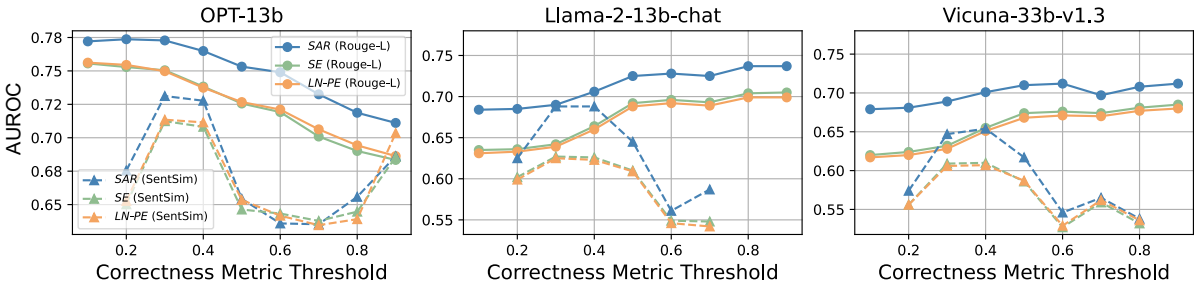

Figure 6: The performance of *SAR* over various Rouge-L and Sentence Similarity thresholds.

the length normalized *PE*, i.e., divided by sentence length N: $\textit{LN-PE}(\boldsymbol{s}, \boldsymbol{x}) = \frac{1}{N}\textit{PE}(\boldsymbol{s}, \boldsymbol{x})$.

Models. We conduct experiments over popular "off-the-shelf" LLMs, including instruction-tuned LLMs (e.g., Vicuna (Zheng et al., 2023), LLaMA-2-chat (Touvron et al., 2023b), WizardLM (Xu et al., 2023)) and pre-trained LLMs (e.g., OPT (Zhang et al., 2022) and LLaMA (Touvron et al., 2023a)), with model size up to 33B. We leverage greedy search for the most likely generations which are used to evaluate correctness, and multinominal sampling for reference generations which are used to estimate uncertainty. More details of generative hyper-parameters can be found in A

Datasets. We consider 5 free-form question-answering datasets: CoQA (Reddy et al., 2019), Trivia QA (Joshi et al., 2017), SciQ (Welbl et al., 2017), MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). More details of the used datasets and the splittings can be found in B.

Correctness Metrics. We adopt Rouge-L (Lin, 2004) and sentence similarity as the correctness metrics when evaluating the correctness of LLMs' generations. We set the threshold of Rouge-L and sentence similarity as 0.5, i.e., generations having above 0.5 semantic similarities/Rouge-L scores with the ground truth are correct. Sentence similarity is measured by DistillRoBERTa (Sanh et al., 2019) in SentenceTransformers (Reimers and Gurevych, 2019b). The sensitivity of *SAR* to these thresholds will be studied in Section 5.4.

Evaluation Metric. Following prior work (Kuhn et al., 2023), we evaluate uncertainty quantification by predicting the correctness of the model's generations regarding a given question. The area under the receiver operator characteristic curve (AUROC) indicates the probability that a random correct generation has a lower uncertainty than a random incorrect generation, predicted by uncertainty quantification methods. AUROC equals 0.5 meaning the assigned uncertainty is no better than random guessing, i.e., they can not differentiate between correct and incorrect generations. AUROC equals 1 meaning all the correct generations are assigned lower uncertainty than all incorrect generations.

Hyperparameters. For OPT-2.7b/6.7b/13b, we generate 10 generations for each question, i.e. K=10. For other models, we generate 5 generations. The temperature t is set to 0.001. The generative settings can be can be found in Appendix A. All the experiments are conducted on a server with one Intel(R) Xeon(R) Platinum 8358 CPU and two NVIDIA A100 GPUs.

5.2 UQ for pre-trained LLMs

We compare *SAR*, TOKEN*SAR*, and SENT*SAR* with state-of-the-art methods. Results are summarized in Figure 4. Generally, our methods significantly outperform prior methods in most of the settings. For instance, *SAR* outperforms other methods by at

OPT Size	*SE*	*SAR* w. sentence similarity RoBERTa	MiniLM	MPNet	OPT-13b
2.7b	0.699	**0.735**	0.723	0.723	0.716
6.7b	0.717	**0.750**	0.740	0.739	0.731
13b	0.725	**0.753**	0.741	0.740	0.733
30b	0.726	**0.748**	0.738	0.739	0.734

Table 2: Sensitivity of *SAR* to sentence similarity measurements. We consider popular models from Sentence-Transformers (Appendix D) and also the target LLMs as the sentence similarity measurement.

most 3.6% AUROC over the CoQA dataset, measured by Rouge-L 0.5. The results of setting Rouge-L to 0.3, which is the same as in (Kuhn et al., 2023), can be found in Appendix C.4.

Also, the synergy of TOKEN*SAR* and SENT*SAR* achieves remarkable improvements. For instance, TOKEN*SAR* and SENT*SAR* achieve 0.723 AUROC in the OPT-30b-CoQA setting yet combining them results in 0.748 AUROC. It indicates that TOKEN*SAR* and SENT*SAR* are compatible and can be incorporated effectively.

5.3 UQ for Instruction-Tuned LLMs

We estimate the uncertainty of powerful instruction-tuned LLMs, including Vicuna-13b/33b, LLaMA-2-chat-13b, and WizardLM-13b. All these models are obtained from Huggingface, without any further modifications. Results are summarized in Table 1. It is shown that *SAR* consistently beat baseline methods in most situations. For example, *SAR* outperforms *SE* by 7.1% AUROC on average, evaluated by Rouge-L 0.5.

5.4 Ablation Studies

Number of Generations. The effects of the number of generations are summarized in Figure 5. It is shown that our *SAR* is generation-efficient, i.e., it achieves 0.750 AUROC with only 5 generations and it can be consistently boosted with more generations, while other methods may even drop slightly when more generations are provided.

Sensitivity to Sentence Similarity. We investigate the sensitivity of SAR to sentence similarity measurements. As shown in Table 2, general-purpose sentence similarity models are desirable and more effective than the target LLMs (last column of Table 2). This is because LLMs are not specifically designed for sentence similarity.

Sensitivity to Correctness Metrics. Empirical results are presented in Figure 5. Higher thresholds mean the correctness standards are more harsh. It is shown that the performances of uncertainty quantization will be affected as the metrics are getting harsh. However, our methods significantly outperform baseline methods in most cases.

Method	# Generation	Time (s)	avg. AUROC SciQ/Trivia QA
PE	5	5.28	0.692/0.657
LN-PE	5	5.28	0.666/0.623
SE	5	6.78	0.673/0.634
SENT*SAR*	2	**2.64**	**0.708/0.685**

Table 3: Efficiency comparisons between 2-generations *SAR* and 5-generations baseline methods.

Model	Dataset	*LN-PE*	*SE*	*SAR*
Vicuna-13b	MedQA	0.572	**0.599**	0.598
	MedMCQA	0.649	0.685	**0.717**
LLaMA-2-13b-chat	MedQA	0.562	0.609	**0.616**
	MedMCQA	0.647	0.655	**0.702**
WizardLM-13b	MedQA	0.609	0.620	**0.635**

Table 4: The performance of *SAR* and baseline methods over medical Q&A datasets. Our method achieves better performances for most settings.

Efficiency Comparison. In Appendix C.5, we provide a detailed computational cost analysis, regarding the time consumed by each operation. We provide the results of 2-generations SENT*SAR* with 5-generations baseline methods in Table 3 over instruction-tuned LLMs. Our *SAR* still surpasses the baseline methods while consuming less than half the time, demonstrating its greater generation efficiency.

5.5 UQ in Medical Domain

We evaluate *SAR* over the AI for science scenarios, such as medical domains. As shown in Table 4, we perform experiments over MedQA (Jin et al., 2020) and MedMCQA (Pal et al., 2022) datasets and our methods achieve better performance for most of the settings. This indicates the potential impacts of our methods on the real world.

6 Conclusion

In this paper, we disclose the generative inequality observation in uncertainty quantification: tokens and generations are created unequally in reflecting semantics yet they are treated equally when estimating uncertainty, which is counter-intuitive. We propose to tackle these inequalities by Shifting Attention to Relevance (*SAR*) from both token-level (TOKEN*SAR*) and sentence-level (SENT*SAR*). Experiments over "off-the-shelf" LLMs demonstrate the superior performances of *SAR*.

7 Ethical Considerations

Our proposed method has the potential to impact the credibility and reliability of LLMs, particularly in the context of reducing misinformation. LLMs have the potential to generate highly plausible but false information. Uncertainty quantification techniques can help distinguish between accurate and misleading outputs. Success in adequately addressing this issue can contribute to the prevention spread of misinformation and its potential societal consequences

8 Limitations

Our method will introduce sentence similarity calculations and comparisons. We tackle this issue by leveraging a small backbone in our implementation but it still might bring additional latency in practice. In addition, our methods require access to token logits. Although token logits are widely supported by commercial LLM providers, this still might restrict the potential application of our methods in black-box scenarios.

Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 23-ERD-030 (LLNL-CONF-851171). This work was partially supported by the National Science Foundation under Grant No.2319242.

References

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink makes a million think: The spread of false information in large language models.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5:277 – 280.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*.

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.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, et al. 2024. Decoding compressed trust: Scrutinizing the trustworthiness of efficient llms under compression. *arXiv preprint arXiv:2403.15447*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal

Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022a. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

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

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

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

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2021. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

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

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Kush R. Varshney. 2023. Towards healthy ai: Large language models need therapists too.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

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

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022b. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. *Findings of the Association for Computational Linguistics: ACL 2022*.

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*.

Jieyi Long. 2023. Large language model guided tree-of-thought.

Andrey Malinin, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. 2020. Regression prior networks. *arXiv preprint arXiv:2006.11590*.

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

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

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023a. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023b. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Aarne Talman, Hande Celikkanat, Sami Virpioja, Markus Heinonen, and Jörg Tiedemann. 2023. Uncertainty-aware natural language inference with stochastic weight averaging. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 358–365.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in nlp: A study on the impact of method & data scarcity. *arXiv preprint arXiv:2210.15452*.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

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

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *ArXiv*, abs/1707.06209.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Appendix

A Details of LLMs Generation

OPT models. We will generate 1 most likely generation with the greedy search for all the OPT models. This generation will be used to evaluate the correctness. For OPT-2.7b/6.7b/13b, we will generate 10 sentences for each question with multinomial sampling for uncertainty quantification. For OPT-30b, we will generate 5 sentences. The temperature of generation is fixed at 0.5 for all models. For OPT-2.6b/6.7b/13b, the max length of each generation is set to 256 tokens for the CoQA dataset and SciQ dataset and is set to 128 tokens for the Trivia QA dataset. For OPT-30b, the max length of each generation is set to 128 tokens for all the datasets.

LLaMA/Vicuna/WizardLM. We will generate 1 most likely generation with the greedy search and 5 sentences with multinomial sampling for all these models. The max length of each generation is set to 128 tokens. The temperature of generation is set to 0.5.

B Datasets

CoQA (Reddy et al., 2019) is a large-scale conversational QA task, with more than 127,000 questions. Each question is equipped with a passage to provide contextual information. **Trivia QA** (Joshi et al., 2017) is a high-quality reading comprehension dataset that contains over 650k question-answer pairs. These questions are obtained from trivia enthusiasts and answers from Wikipedia. **SciQ** (Welbl et al., 2017) dataset is a science-related QA dataset aimed at developing models' capabilities of understanding complex scientific texts. It consists of approximately 13,679 crowdsourced science questions. **MedQA** (Jin et al., 2020) is a free-form multiple-choice OpenQA dataset for solving medical problems, collected from the professional medical board exams. **MedMCQA** (Pal et al., 2022) is a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions.

Following (Kuhn et al., 2023), we randomly select around 8,000 questions from the training split of Trivia QA as the questions to be examined. For instruction-tuned experiments, we use 2,000 questions of Trivia QA. We utilize the full validation set (1,000 questions) of SciQ and the development split (7,983 questions) of CoQA. For MedQA and MedMCQA, we also utilize their full validation sets.

C Additional Experimental Analysis

C.1 Effects of *SAR* Temperature t

The hyperparameter t introduced in Eq. (9) is used to control the scale of sentence shifting. The effects of t is provided in Table 5. It is shown that t marginally affects the performance of *SAR*.

t	OPT-13b		LLaMA-7b	
	CoQA	SciQ	CoQA	TriviaQA
1×10^{-3}	0.753/0.720	0.737/0.784	0.697/0.658	0.823/0.815
1×10^{0}	0.752/0.719	0.739/0.786	0.695/0.656	0.822/0.816
1×10^{1}	0.743/0.714	0.729/0.786	0.686/0.658	0.813/0.812

Table 5: Effects of temperature t in Eq. (9). Results are evaluated by Rouge-L with 0.5 as the threshold. Results are obtained from *SAR*/TOKEN*SAR*.

C.2 Generation Efficiency

The generation-efficiency of *SAR* on LLaMA-7b-Trivia QA setting is presented in Figure 7.

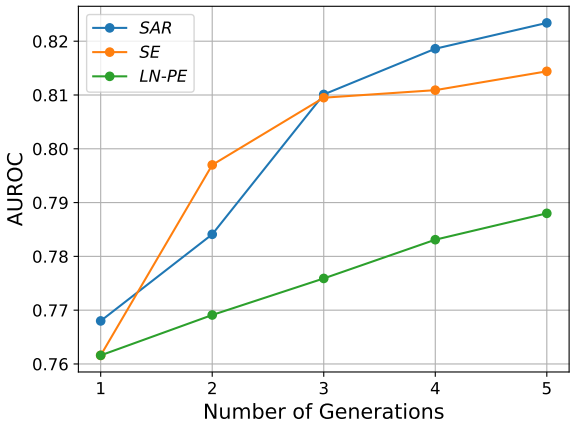

Figure 7: The performance of *SAR* over various numbers of generations. Results are obtained from the LLaMA-7b model over the Trivia QA dataset.

C.3 Sensitivity to Sentence Length.

To study how the *SAR* is affected by sentence length, we quantify the uncertainty rank change for each sentence, caused by *SAR* and SENTß*SAR*. Assume a sentence has a rank of i among all the sentences, evaluated by LN-PE and has a rank of j evaluated by *SAR*, then the uncertainty rank change is $|i - j|$. The correlations between average uncertainty rank change and sentence length are presented in Figure 8. It is shown that our methods tend to conclude medium- and long-length sentences.

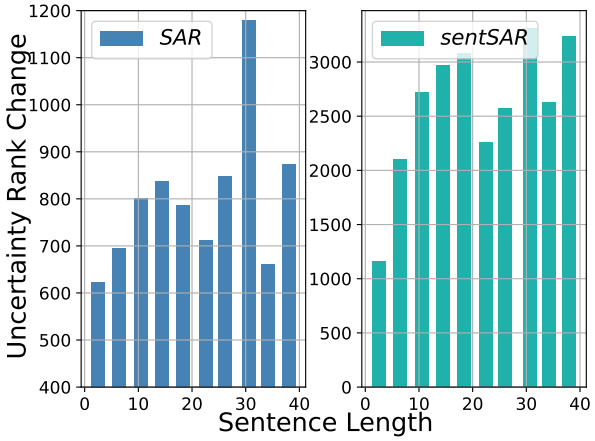

Figure 8: Demographic analysis of sentence length. Uncertainty Rank Change between (**Left**) *SAR* and *LN-PE*, and between (**Right**) SENT*SAR* and *LN-PE*. It is shown that *SAR* and SENT*SAR* are more tend to affect medium- or long-length sentences.

C.4 Different Correctness Metric Threshold

We report the results of Rouge-L (0.3) (same as (Kuhn et al., 2023) in Table 6.

C.5 Computational Costs Analysis

SAR is more generation-efficient. It surpasses baseline methods under significantly smaller computational constraints. We have quantified the time consumed for each step in the overall uncertainty quantification pipeline. This includes sequence generation, computing logits, semantic clustering for SE, and sentence similarity for *SAR*. We exclude the time taken for aggregating logits/scores as it is negligible (less than 0.001 seconds for all methods). The average time consumed per question, based on an evaluation of 1000 questions from the Vicuna-13b + SciQ dataset, is provided. These measurements were taken using an AMD EPYC 7302 16-Core CPU and a 1xA40 GPU server. Results are summarized in Table 7.

D Sentence Similarity Measurement

The following is the sentence similarity measurement models we leveraged in Table 2:

- RoBERTa: cross-encoder/stsb-roberta-large
- MiniLM: sentence-transformers/all-MiniLM-L6-v2
- MPNet: sentence-transformers/all-mpnet-base-v2

Dataset	Model	*LS*	*PE*	*LN-PE*	*SE*	TOKEN*SAR*	SENT*SAR*	*SAR*
CoQA	OPT-2.7b	0.573	0.666	0.719	0.712	0.719	0.689	**0.742**
	OPT-6.7b	0.588	0.671	0.745	0.741	0.746	0.696	**0.768**
	OPT-13b	0.588	0.666	0.750	0.751	0.752	0.690	**0.773**
	OPT-30b	0.550	0.671	0.742	0.751	0.746	0.698	**0.767**
	LLaMA-7b	0.511	0.646	0.673	0.672	0.672	0.635	**0.686**
	LLaMA-13b	0.522	0.617	0.653	0.652	0.653	0.610	**0.665**
Trivia QA	LLaMA-7b	0.533	0.713	0.783	0.814	0.793	0.800	**0.818**
	LLaMA-13b	0.655	0.492	0.627	**0.758**	0.635	0.749	0.716
Average		0.565	0.643	0.712	0.731	0.715	0.696	**0.742**

Table 6: Uncertainty estimation AUROCs of TOKEN*SAR*, SENT*SAR*, *SAR*, and baseline methods, across various "off-the-shelf" LLMs and datasets (e.g., CoQA, and Trivia QA). Rouge-L with a threshold of 0.3 is used as the correctness metric.

Method	Num. of Generations	Generation	Logits Computing	Semantic Clustering	Sentence Similarity	Sum
PE	5	4.09s	1.19s	0s	0s	5.28s
LN-PE	5	4.09s	1.19s	0s	0s	5.28s
SE	5	4.09s	1.19s	1.5s	0s	6.78s
SENT*SAR*	5	4.09s	1.19s	0s	2.58s	7.86s
SENT*SAR*	2	1.64s	0.48s	0s	0.52s	**2.64s**

Table 7: Computational costs of *SAR* and baseline methods. We report both SENT*SAR* with 5 and 2 generations.