LUQ: Long-text Uncertainty Quantification for LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable capability in a variety of NLP tasks. Despite their effectiveness, these models are prone to generate nonfactual content. Uncertainty Quantification (UQ) is pivotal in enhancing our understanding of a model's confidence in its generated content, thereby aiding in the mitigation of nonfactual outputs. Existing research on UQ predominantly targets short text generation, typically yielding brief, word-limited responses. However, real-world applications frequently necessitate much longer responses. Our study first highlights the limitations of current UQ methods in handling long text generation. We then introduce LUQ, a novel sampling-based UQ approach specifically designed for long text. Our findings reveal that LUQ outperforms existing baseline methods in correlating with the model's factuality scores (negative coefficient of -0.85 observed for Gemini 1.0 Pro). With LUQ as the tool for UQ, we investigate behavior patterns of several popular LLMs' response confidence spectrum and how that interplays with the response' factuality. We identify that LLMs lack confidence in generating long text for rare facts and a factually strong model (i.e. GPT-4) tends to reject questions it is not sure about. To further improve the factual accuracy of LLM responses, we propose a method called LUQ-ENSEMBLE that ensembles responses from multiple models and selects the response with the least uncertainty. The ensembling method greatly improves the response factuality upon the best standalone LLM.

1 Introduction

Large Language Models (LLMs) have demonstrated significant prowess across a wide range of NLP tasks and are increasingly being used in various downstream applications (Zhao et al., 2023; Chang et al., 2023). However, existing LLMs are susceptible to hallucination, often resulting in the generation of nonfactual or fabricated content (Manakul et al., 2023; Zhang et al., 2023). One way to predict the factuality of an LLM's output without resorting to resource-intensive fact-checking procedures is by examining its uncertainty over a user query. Moreover, accurate measurement of a model's confidence in its generated responses can enable the rejection of answers with high uncertainty, potentially reducing hallucinations and improving the factuality of the output (Geng et al., 2023; Wang et al., 2023).

Although Uncertainty Quantification (UQ) is a well-researched area in the broader field of machine learning (Gawlikowski et al., 2023), its application in the context of LLMs remains relatively underexplored. One primary limitation is that previous studies on UQ mostly require access to a model's internal states (*e.g.* logits) (Murray and Chiang, 2018; Kuhn et al., 2022; Vazhentsev et al., 2023; Duan et al., 2023). However, many of the most advanced LLMs, such as GPT-4 (OpenAI, 2023), Gemini 1.0 Pro (Gemini Team, 2023), and Claude 2.1 (Anthropic, 2023), are closed-source and only accessible via API calls. This limits the ability to directly analyze their internal processes. Another challenge is that existing research on modeling uncertainty predominantly focuses on short responses, typically less than 10 words in length (Kuhn et al., 2022; Duan et al., 2023; Lin et al., 2023). This is in stark contrast to the more common use cases of LLMs, where responses to queries often far exceed this length, sometimes reaching hundreds of words. Such disparity points to a need for new UQ methods tailored for long-form text generated by LLMs. Therefore, in this study we aim to answer the following research questions: **RQ1:** Are existing UQ methods still effective in the context of long-text generation? **RQ2:** If not, how can we effectively quantify LLMs' uncertainty for long-form answers? **RQ3:** In what ways can uncertainty scores be utilized to enhance the factuality of model

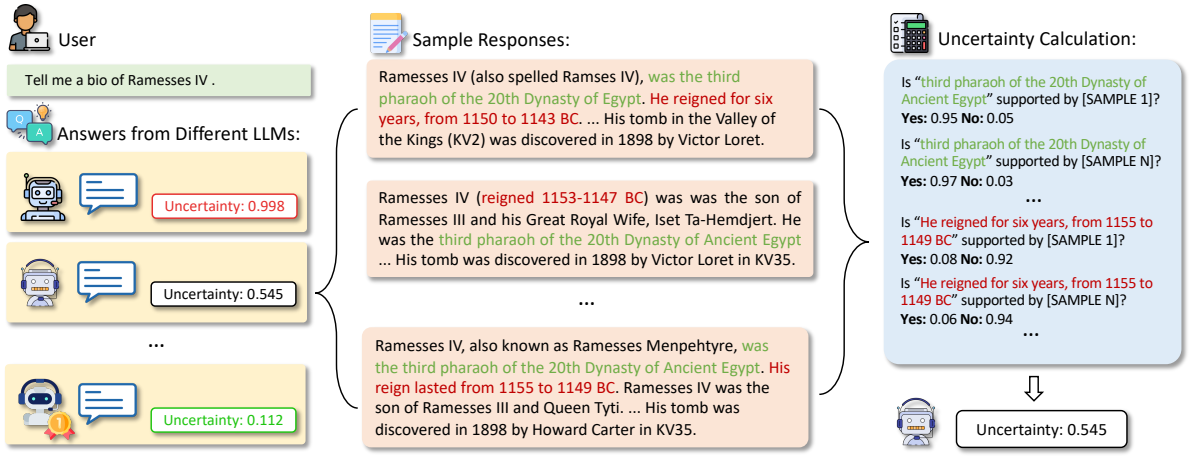

Figure 1: The illustration of the LUQ and LUQ-ENSEMBLE framework. Given a question, various LLMs exhibit differing levels of uncertainty. We generate n sample responses from each LLM and then assess the uncertainty based on the diversity of these samples (the LUQ metric). Green highlights indicate consistency across responses (low uncertainty) and red highlights discrepancies (high uncertainty). The LUQ-ENSEMBLE method selects the response from the LLM with the lowest uncertainty score as the final answer.

outputs?

Our experiments primarily focus on black-box LLMs, with an emphasis on using factuality as the key metric to evaluate the models' performance. The main contributions of this paper are:

- We first highlight the limitations of existing UQ methods for long text generation and subsequently propose LUQ (Long-text Uncertainty Quantification; pronounced as "luck"), a novel UQ method that computes sentence-level consistency in long text scenarios.
- Through extensive experiments on the original FACTSCORE dataset and our newly proposed FACTSCORE-DIS dataset in medical domain, we demonstrate that our proposed LUQ consistently shows strong negative correlations with the responses' factuality over 6 popular LLMs, outperforming all the baseline methods.
- We propose an ensemble modeling approach that selects responses from the model exhibiting the lowest LUQ uncertainty score, observing an improvement of up to 5% in the overall factuality scores. Additionally, we enhance the model's uncertainty awareness by implementing a selective answering strategy.

2 Background

2.1 Uncertainty and Confidence

Confidence and uncertainty in the context of machine learning models pertain to the level of assurance or certainty associated with a prediction or decision (Geng et al., 2023). While many studies treat *confidence* and *uncertainty* as antonyms and use them interchangeably (Xiao et al., 2022; Chen and Mueller, 2023), Lin et al. (2023) provide a clear distinction: uncertainty denotes the dispersion of potential predictions for a given input, whereas confidence pertains to the degree of confidence in a specific prediction or output. We will adopt this terminology in the following sections.

Currently, a formal and universally accepted definition of uncertainty levels in language generation tasks remains elusive. Common practice in existing literature involves measuring uncertainty through the entropy of predictions, akin to approaches in classification tasks (Kuhn et al., 2022; Lin et al., 2023). Predictive entropy is formally expressed as $H(Y \mid x) = -\int p(y \mid x) \log(p(y \mid x))dy$, which captures the uncertainty associated with a prediction for a given input x. In the context of NLG, where $\mathbf{R}$ denotes all possible generations and $\mathbf{r}$ is a specific response, the uncertainty score can thus be conceptualized as:

$$U(x) = H(\mathbf{R} \mid x) = -\sum_{\mathbf{r}} p(\mathbf{r} \mid x) \log(p(\mathbf{r} \mid x))$$

Similarly, the concept of confidence within NLG frequently adopts methodologies from classification tasks. In these tasks, confidence for a specific prediction y is quantified using the predicted probability, represented as $\hat{p}(Y = y \mid x)$ (Geifman and El-Yaniv, 2017; Hendrycks and Gimpel, 2016). In the context of NLG, the confidence score for a

given response $\mathbf{r}$ is represented by the joint probability of the tokens in the response:

$$C(x, \mathbf{r}) = \hat{p}(\mathbf{r} \mid x) = \prod_i \hat{p}\left(r_i \mid r_{<i}, x\right).$$

2.2 Uncertainty for Long Text Generation

In our study, we adopt a more flexible approach to defining uncertainty and confidence in long text generation. We focus on the ability of UQ methods to effectively rank responses, differentiating between correct and incorrect predictions. This approach aligns with the concept of *relative confidence* as discussed by Geng et al. (2023). Our objective diverges from the orthogonal research direction about models' calibration, which requires models to precisely reflect their true accuracy in practical scenarios (Lin et al., 2023). We argue that while short-answer questions may be straightforwardly assessed using metrics such as accuracy or exact match, these standards are often unrealistic for long text generation, given the complexities of real-life probabilities.

From a practical perspective, we aim for the uncertainty score to serve as a reliable indicator of the model's performance. This performance encompasses several dimensions of generation quality, including factuality, coherence, and creativity. Notably, our study prioritizes factuality and the truthfulness of responses, adopting these as our primary metrics. The factuality of the responses $\mathbf{R}$ given a specific query x is denoted as $F\left(\mathbf{R} \mid \mathbf{x}\right)$. Considering two inputs x_i and x_j, we explore the relationship between the model's uncertainty, denoted as $\mathrm{U}\left(\mathbf{x}\right)$, and the factuality. Our goal is to have:

$$\mathrm{U}\left(\mathbf{x}_i\right) \leq \mathrm{U}\left(\mathbf{x}_j\right) \Longleftrightarrow F\left(\mathbf{R} \mid \mathbf{x}_i\right) \geq F\left(\mathbf{R} \mid \mathbf{x}_j\right)$$

Correspondingly, for a given input x, the model's confidence in generating a specific response r is represented as $\mathrm{C}\left(\mathbf{x}, \mathbf{r}\right)$. Thus, we aim to establish the following relationship:

$$\mathrm{C}\left(\mathbf{x}, \mathbf{r}_i\right) \leq \mathrm{C}\left(\mathbf{x}, \mathbf{r}_j\right) \Longleftrightarrow F\left(\mathbf{r}_i \mid \mathbf{x}\right) \leq F\left(\mathbf{r}_j \mid \mathbf{x}\right)$$

3 LUQ

In this section, we introduce our LUQ method for estimating uncertainty in long text generation. The framework is illustrated in Figure 1. Our underlying assumption posits that the greater the model's uncertainty regarding a given question x, the more diverse its responses to question x will be. For instance, as shown in Figure 1, the term "*third pharaoh of the 20th Dynasty of Egypt*" is frequently supported by other sample responses, indicating the model's high confidence in this information. However, the samples suggest different reign periods for Ramesses IV; the inconsistency shows the model is uncertain about this information. We begin by highlighting the limitations of previous UQ methods on dealing with long text, then formally define our LUQ method that overcomes these issues.

Motivation Following the generation of n responses, traditional UQ methods for short text commonly calculate the pairwise similarity among the responses, employing either Jaccard Similarity or calculating Natural Language Inference (NLI) scores (Kuhn et al., 2022; Lin et al., 2023). These pairwise similarity scores indicate the consistency between a pair of responses and play a vital role in subsequent uncertainty estimation.

However, responses to certain queries are brief, making it relatively straightforward to distinguish between different answers. For example, for the question *"What is the capital of France?"*, the answers often consists of just a few words and can be easily examined. In contrast, answers to other questions such as *"Give me an introduction of ..."* and *"Tell me something about ..."* may extend to hundreds of words. Longer text leads to an unexpected high similarity across all response pairs when applying previous methods. To address this issue and achieve a more nuanced similarity assessment, we propose the LUQ uncertainty measurement with sentence-level similarity computation. Inspired by the hallucination detection method in Manakul et al. (2023), we split each response to sentences, and check whether each sentence can be supported by other samples.

Notation Let r_a represent the response generated by a LLM to a user query x. We generate an additional n stochastic LLM sample responses $R = \{r_1, r_2, \ldots, r_n\}$ using the same query. The set $R' = \{r_a, r_1, r_2, \ldots, r_n\}$ encompasses all outputs from the model.

For any given response $r_i \in R'$, the first objective is to determine how often it is supported (or entailed) by other samples. To this end, we employ a NLI classifier to assess the similarity between r_i and each $r' \in R' \setminus \{r_i\}$. The output from an NLI classifier normally includes classifications of entailment, neutral, and contradiction, along with their respective logit values. It is important to note that

we focus exclusively on the "entailment" and "contradiction" classes, as sentences labeled as "neutral" generally do not impact the overall factuality of a response. We calculate the NLI score for each sentence s_j within a response r, and then average these scores. Formally, the similarity score $\mathcal{S}(r_i, r')$ between r_i and r' is defined as:

$$\mathcal{P}\left(\text{ entail } \mid s_j, r'\right) = \frac{\exp\left(l_e\right)}{\exp\left(l_e\right) + \exp\left(l_c\right)}$$

$$\mathcal{S}(r_i, r') = \frac{1}{n}\sum_{j=1}^{n} P\left(\text{ entail } \mid s_j, r'\right)$$

where l_e and l_c are the logits of the "entailment" and "contradiction" classes, respectively. We opt to calculate $\mathcal{P}(\text{entail} \mid s_j, r')$ over $\mathcal{P}(\text{contradict} \mid s_j, r')$ because non-contradictory responses can still be largely irrelevant, indicating higher uncertainty (Lin et al., 2023). The model's confidence in response r_i and the overall uncertainty is therefore defined as:

$$C(x, r_i) = \frac{1}{n}\sum\nolimits_{r' \in R' \setminus \{r'\}} \mathcal{S}(r_i, r')$$

$$U(x) = \frac{1}{n+1}\sum\nolimits_{r_i \in R'} (1 - \mathcal{C}(x, r_i))$$

Unlike Kuhn et al. (2022)'s method of applying an off-the-shelf DeBERTa model, we apply the DeBERTa-v3-large model (He et al., 2022), fine-tuned on the MultiNLI (Williams et al., 2018) dataset. This choice is due to our input being a concatenation of short hypothesis (sentence s) and a comparatively longer premises (reference response r'). The format of our input aligns with the task in MultiNLI dataset, ensuring an effective assessment of consistency among the responses.

4 Experiments

4.1 Dataset, Metric, and LLM Selection

Dataset We employ FACTSCORE (Min et al., 2023) to evaluate the factuality of our generated text. FACTSCORE offers automated assessment with a low error rate (below 2%), enabling scalable application to diverse LLMs without requiring manual annotation. To supplement the extensive reliability testing of FACTSCORE conducted by its creators, we performed a smaller-scale human annotation study. Our findings demonstrate a strong Pearson correlation of 0.88 between FACTSCORE ratings and human factuality judgments, suggesting FACTSCORE being a reliable reference for factuality. For a comprehensive description of our validation process, please refer to Appendix E.

The original FACTSCORE dataset (denoted as FACTSCORE-BIO) includes 500 individuals' biographies from Wikidata with corresponding Wikipedia entries. To evaluate the applicability of UQ methods across different domains, we additionally developed a dataset, FACTSCORE-DIS, focusing on disease entities. Details of this dataset can be found in Appendix D.

Metrics For each generated response, FACTSCORE calculates a factuality score (FS). It first breaks down the response into a series of atomic facts, which are concise statements that capture a single piece of information. It then assigns a binary label to each atomic fact and calculate the precision as the response's factuality score. We apply FACTSCORE for the first generated response (r_a). As the LLMs may sometime refuse to answer certain questions, to have a fair comparison, we introduce a penalized factuality score (PFS) and penalized uncertainty score (PUS). To calculate PFS and PUS, we assign a factuality score of zero and uncertainty score of one to questions that models opt not to answer.

We then proceed to calculate both the Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SCC) between the factuality scores and uncertainty scores. Following the criteria proposed by Schober et al. (2018), we classify the correlation coefficients into five categories based on their absolute values: over 0.9 indicates a very strong correlation; 0.7 to 0.9 signifies strong; 0.5 to 0.7 suggests moderate; 0.3 to 0.5 denotes weak; 0.1 to 0.3 implies very weak; and below 0.1 means negligible correlation.

LLMs We selected six top-performing LLMs from the Arena Leaderboard (Zheng et al., 2023) for our experiments. Within our access rights, we chose three closed-sourced models: GPT-4 (OpenAI, 2023), GPT-3.5 (OpenAI, 2022), and Gemini 1.0 Pro (Gemini Team, 2023); and three open-sourced models: Yi-34B-Chat (01.ai, 2023), Tulu-2-70B (Ivison et al., 2023), and Vicuna-33B (Zheng et al., 2023). For each LLM, we include the following baseline UQ methods for comparison. Our implementation is based on the LM-Polygraph framework as proposed by Fadeeva et al. (2023). More details are provided in Appendix A.

		LexSim	Ecc	NumSets	EigV	Deg	SelfCheck	LUQ
FACTSCORE-BIO								
GPT-4	PCC	-45.25	-24.83	-8.24	-36.94	-3.78	-53.12	**-60.37**
	SCC	-35.97	-12.74	4.18	-18.72	6.73	-41.79	**-45.28**
GPT-3.5	PCC	-67.75	-10.61	-11.90	-30.30	-22.44	-65.14	**-71.30**
	SCC	-52.45	-26.46	-17.02	-34.60	-22.85	-61.07	**-66.62**
Yi-34B-Chat	PCC	-70.09	-27.63	-25.69	-48.97	-39.82	-70.34	**-73.78**
	SCC	-68.23	-44.97	-31.34	-51.12	-38.94	-72.69	**-74.59**
Tulu-2-70B	PCC	-55.69	-2.13	-20.68	-50.08	-53.39	-75.63	**-77.58**
	SCC	-61.84	10.12	-18.07	-50.29	-53.97	**-76.89**	-75.39
Gemini Pro	PCC	-67.24	-50.27	-52.98	-72.73	-64.37	-84.49	**-85.09**
	SCC	-63.73	-57.79	-57.02	-69.68	-67.69	**-82.37**	-81.29
Vicuna-33B	PCC	-38.28	-18.72	-20.04	-60.49	-58.25	-66.82	**-71.79**
	SCC	-50.62	-13.98	-16.65	-61.74	-62.38	-66.48	**-70.78**
FACTSCORE-DIS								
GPT-3.5	PCC	-41.75	-27.92	-7.81	-38.75	-13.50	-58.97	**-67.31**
	SCC	-39.37	-25.96	-6.94	-36.93	-16.33	-59.11	**-65.30**
Yi-34B-Chat	PCC	-63.61	-19.27	-11.23	-40.59	-26.45	-65.14	**-70.48**
	SCC	-58.68	-21.52	-16.34	-38.41	-22.11	-67.81	**-72.39**

Table 1: Pearson and Spearman correlation coefficients (expressed as percentages) between different LLMs and various UQ methods on the FactScore dataset. We use the original factuality scores instead of the penalized ones.

	FS	PFS	US	PUS	RR
GPT-4	80.76	72.37	20.75	28.98	86.62
GPT-3.5	68.25	68.25	25.71	25.71	100
Yi-34B-Chat	55.71	55.71	41.25	41.25	100
Tulu-2-70B	47.19	47.19	55.83	55.83	100
Gemini Pro	43.20	42.73	61.74	62.17	98.90
Vicuna-33B	42.47	42.47	55.31	55.31	100

Table 2: Results on FACTSCORE-BIO dataset; FS and PFS stand for average factuality score and penalized factuality score; US and PUS stand for average uncertainty score and penalized uncertainty score computed by LUQ; RR means the respond rate. All numbers are in percentages.

Baselines for UQ We use the following black-box UQ methods as baselines: Lexical similarity (LexSim) (Fomicheva et al., 2020), Numer of semantic sets (NumSets) (Lin et al., 2023), Sum of eigenvalues of the graph Laplacian (EigV) (Lin et al., 2023), Degree matrix (Deg) (Lin et al., 2023), Eccentricity (Ecc) (Lin et al., 2023), SelfCheckNLI (SelfCheck) (Manakul et al., 2023). We mainly use the library LM-Polygraph (Fadeeva et al., 2023) for the UQ methods. The details of these methods can be found in Appendix B.

4.2 Uncertainty Quantification Results

Effectiveness of LUQ Table 1 and Figure 2 illustrate the correlation between factuality scores and uncertainty scores. The results highlight LUQ's effectiveness as an indicator of model factuality in long text generation tasks. LUQ demonstrates a strong negative correlation for GPT-3.5, Gemini 1.0 Pro, Yi-34B-Chat, Vicuna-33B, and Tulu-2-70B, with the strongest Pearson correlation being -0.8509.

For the baseline methods, LexSim emerges as a robust baseline offering lower computational demands. EigV have competing performance with LexSim in models like Gemini 1.0 Pro and Vicuna-33B. The confidence-based SelfCheck method demonstrates the best Spearman correlation in models such as Gemini 1.0 Pro and Tulu-2-70B. Other baselines such as Eccentricity, NumSets and Deg yield unsatisfactory results, occasionally exhibiting even positive correlations.

We also observe that LUQ is better suited for models with relatively lower factuality and a lack of self-awareness regarding uncertainty. For models with high factuality capabilities, such as GPT-4, LUQ only demonstrates a moderate correlation with factuality scores. As indicated in Table 2 and Figure 2a, among all models, GPT-4 exhibits the highest overall factuality scores and the lowest average uncertainty scores. The data points are tightly clustered with only few instances of uncertainty. Moreover, GPT-4's self-awareness of uncertainty may also impact external UQ efficiency, as it tends to abstain from answering questions more often compared to other models, highlighting its heightened internal uncertainty awareness. This observation does not influence the effectiveness of our method, as in real life models with lower factual-

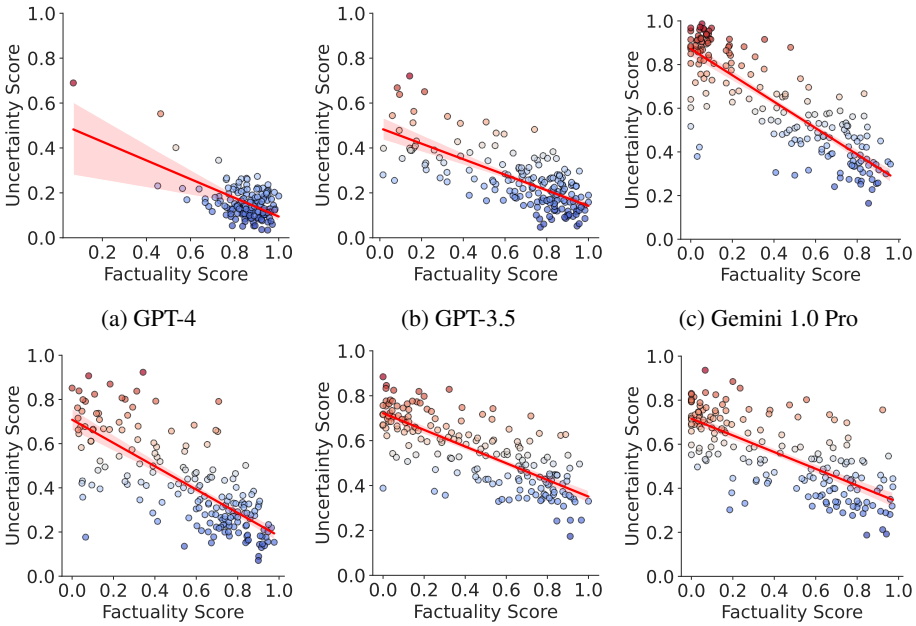

Figure 2: Scatter plot illustrating the relationship between factuality scores (x-axis) and uncertainty scores (y-axis) for different LLMs. Each point symbolizes an item in the FactScore dataset, with a red line highlighting the Pearson correlation. The distribution suggests a pattern where higher factuality correlates with lower uncertainty.

ity and unable to express uncertainty are in greater need of external uncertainty measurements.

LUQ in FACTSCORE-DIS We test one closed-source LLM, GPT-3.5, and one open-source LLM, Yi-34B-Chat in our newly proposed FACTSCORE-DIS. Our LUQ model consistently surpasses the performance of baseline models, thereby demonstrating its effectiveness on the newly proposed dataset within the medical domain.

Higher frequency leads to higher factuality and lower uncertainty. In Figure 3, we compare the factuality and uncertainty scores across different entity frequencies. The original FACTSCORE dataset provides the frequency of each entity in Wikipedia, categorizing them based on page views and co-occurrence within the training set (Min et al., 2023). Frequencies are classified into five categories, ranging from "very rare" to "very frequent." Our observations suggest that questions associated with higher entity frequencies tend to yield more factual responses, alongside decreased model uncertainty. Notably, GPT-4 demonstrates consistent performance regarding uncertainty and factuality across varying frequencies, potentially attributable to its selective response strategy. Although it answers all the questions in the "very frequent," "frequent," and "medium" categories, it refuses to answer around 25% of "rare" questions and 30% of "very rare" questions.

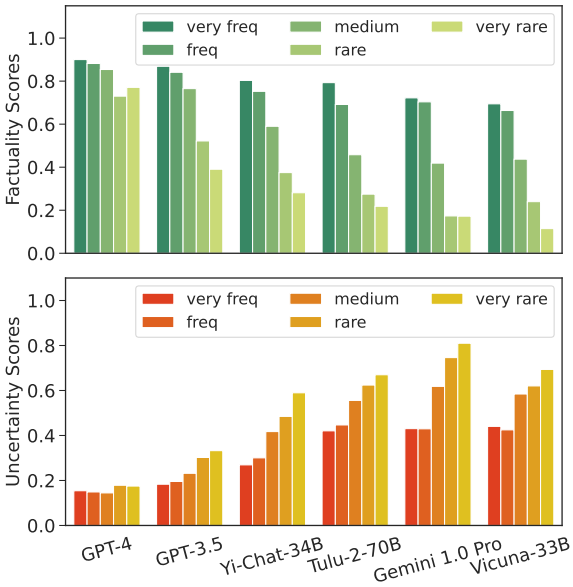

Figure 3: Comparison of factuality and uncertainty scores across different frequencies on FACTSCORE-BIO.

4.3 LUQ-ENSEMBLE

Given the variance in training corpus, different LLMs may possess varying levels of knowledge for a specific question, leading to discrepancies in uncertainty levels across models. Utilizing the LUQ uncertainty score as a reliable indicator of fac-

Percentile	GPT-3.5		Yi-34B-Chat		Tulu-2-70B		Vicuna-33B		Gemini 1.0 Pro	
	FS	US	FS	US	FS	US	FS	US	FS	US
0	68.25	25.71	55.71	41.25	47.19	55.83	42.47	55.31	43.20	61.74
2.5	69.80	24.07	56.91	40.24	48.28	53.86	43.59	54.41	44.28	60.05
5	70.82	23.40	57.99	39.28	49.35	53.13	44.54	53.76	45.20	59.23
7.5	71.52	22.66	58.86	38.12	50.27	52.57	45.54	52.97	46.34	58.19
10	72.30	22.18	60.21	36.79	51.45	51.87	46.14	52.34	47.25	57.73
12.5	74.13	21.61	61.69	35.04	52.13	51.31	46.53	51.58	48.41	56.29
15	75.04	21.17	62.91	34.18	53.29	50.63	47.52	50.96	49.51	55.40

Table 3: Selective question answering results on FACTSCORE-BIO (expressed as percentage). The percentile indicates the percentage of questions for which answers were abstained.

Methods	PFS	PUS	AD
Tulu-2-70B	47.19	55.83	42.08
Gemini 1.0 Pro	42.73	62.17	29.51
Vicuna-33B	42.47	58.06	28.41
LUQ-ENSEMBLE	**52.83**	**45.83**	100
Yi-34B-Chat	55.71	41.25	66.12
Tulu-2-70B	47.19	55.83	21.31
Gemini 1.0 Pro	42.73	62.17	12.57
LUQ-ENSEMBLE	**58.75**	**37.60**	100
GPT-3.5	67.31	25.71	92.35
Gemini 1.0 Pro	42.73	62.17	1.64
Vicuna-33B	42.47	58.06	6.01
LUQ-ENSEMBLE	**67.37**	**24.75**	100
GPT-4	72.11	28.98	60.11
GPT-3.5	67.31	25.71	32.79
Yi-34B-Chat	55.71	41.25	7.10
LUQ-ENSEMBLE	**76.61**	**17.27**	100

Table 4: Results of different ensemble strategies on FACTSCORE-BIO (expressed as percentage). The Answer Distribution (AD) metric indicates the percentage of final answers generated by each component model.

tuality, we enhance overall performance through an ensemble approach. In this method, the model exhibiting the lowest LUQ score for a given question is chosen as the final answer. Experimental results (Table 4) affirm the superiority of the LUQ-ENSEMBLE over its constituent counterparts.

Ensembling models with similar factuality scores can notably enhance performance. Our findings suggest that ensembling models with similar factuality scores can significantly enhance performance. For instance, in the combination of Tulu-2-70B, Gemini 1.0 Pro, and Vicuna-33B, the PFS increases by 5% compared to the originally top-performing Tulu-2-70B, which scored 47.19%. Additionally, ensembling models with comparable performance leads to a more balanced distribution of answers. In contrast, integrating a model with substantially superior performance, as seen in the combination of GPT-3.5, Gemini 1.0 Pro, and Vicuna-33B, predominantly favors answers from GPT-3.5 (92.35%), leading to marginal improvement (0.06%) in the ensemble method.

To provide a detailed analysis, Table 4 includes an ensemble of the top three models with the highest factuality scores: GPT-4, GPT-3.5, and Yi-34B-Chat. The results indicate that the majority of answers are generated by GPT-4 and GPT-3.5, accounting for 92.90%. However, the improvement in PFS primarily stems from questions that GPT-4 refuses to answer. They are given a factuality score of zero and an uncertainty score of one for penalty. This underscores a pivotal decision point for practitioners regarding model behavior in the face of uncertainty: whether the model should attempt to answer even though it is uncertain or opt to abstain.

4.4 Selective Question Answering

From Table 2, it is observed that while GPT-4 opts not to respond to some queries, other models generally attempt to answer all questions. The limited refusal by Gemini 1.0 Pro primarily stems from considerations of sensitive content and regulatory constraints, rather than uncertainty[1]. Therefore, we investigate the application of the LUQ score to equip these models with the capability for selective question answering—that is, to enable them to decline responses when uncertain. Contrary to the traditional aim of responding correctly to every question, the objective in a selective question answering framework is to preserve accuracy while maximizing the number of questions answered (Kamath et al., 2020; Cole et al., 2023; Yang et al., 2023).

Table 3 presents the results of selective question answering. The models are permitted to refrain from answering questions with high uncertainty. The *percentiles* indicate the proportion of ques-

[1]Gemini 1.0 Pro API returns the reasons of refusing to answer certain questions.

tions each model abstained from answering. The findings demonstrate that adopting a selective answering approach enhances the models' factuality by allowing for more question rejections. By declining to answer a similar proportion of questions (approximately 15%) as GPT-4, the models typically achieve an improvement of over 5% in overall factuality scores. Notably, when implementing this selective answering strategy in practical applications, it is essential for practitioners to tailor the uncertainty thresholds to the specific models and tasks at hand.

5 Related Work

UQ in Machine Learning Prior to LLMs, UQ has been extensively explored within the field of machine learning (Gawlikowski et al., 2023). According to the source of uncertainty, it is typically categorized into two types: aleatoric and epistemic uncertainty(Hora, 1996; Der Kiureghian and Ditlevsen, 2009). Aleatoric uncertainty, also known as statistical uncertainty, pertains to the inherent randomness in experimental outcomes due to stochastic effects (Hüllermeier and Waegeman, 2021). In contrast, epistemic uncertainty stems from incomplete knowledge, potentially including uncertainties in a machine learning model's parameters or the lack of certain training data (Hüllermeier and Waegeman, 2021; Huang et al., 2023). Our focus is primarily on epistemic uncertainty.

UQ in LLMs In contrast to discriminative models, which readily provide probability scores for specific categories, uncertainty estimation in generative LLMs presents unique challenges: (1) There is an exponential increase in the output space as sentence length grows, rendering the evaluation of all possible predictions impractical (Geng et al., 2023; Wang et al., 2023). (2) The significance of semantic nuances and their inherent uncertainties, which diverges from the fixed category labels typical of discriminative models, complicates matters further (Kuhn et al., 2022). Generally, UQ methods for LLMs can be categorized based on the accessibility of the model's internal states, distinguishing between black-box and white-box approaches. White-box LLMs often rely on logit-based evaluations, assessing sentence uncertainty through token-level probabilities or entropy (Murray and Chiang, 2018; Kuhn et al., 2022; Vazhentsev et al., 2023; Duan et al., 2023).

However, as access to LLMs increasingly relies on API calls, research has pivoted towards black-box methods. These can be further categorized into: (i) *verbalized methods*, which prompt LLMs to articulate their uncertainty in the output, using phrases like "I am sure" or "I do not know" (Mielke et al., 2022). Nonetheless, a practical mismatch between the expressed and actual uncertainty levels has been noted (Lin et al., 2022; Xiong et al., 2023). Xiong et al. (2023) highlight that LLMs often display excessive confidence when verbalizing their certainty. (ii) *Consistency-based (sampling-based)* estimation premises on the assumption that increased uncertainty in a model corresponds to greater diversity in its outputs, frequently resulting in hallucinatory outputs (Manakul et al., 2023; Lin et al., 2023). Our proposed method, LUQ, follows this consistency-based approach. There are also efforts on integrating verbalized methods with consistency-based approaches (Xiong et al., 2023; Rivera et al., 2024).

6 Conclusion

In this work, we first identify that existing UQ methods are ineffective on long text generation. We therefore introduce LUQ, a novel UQ method tailored for long-form text generation in LLMs. It overcomes the limitation of previous methods by calculating sentence level consistency. We conduct extensive experiments over six popular LLMs, such as GPT-4 and Gemini 1.0 Pro. We extend the existing FACTSCORE dataset with human validation and annotations for additional disease domain. Our findings demonstrate that LUQ significantly improves the correlation with models' factuality scores over previous methods across various different setups and domains. LUQ serves as a reliable indicator of model's factuality performance. Additionally, we present LUQ-ENSEMBLE, a model ensembling and selective question answering strategy, which showcases a promising avenue for enhancing the factual accuracy of LLM outputs. This research not only advances our understanding of UQ in the context of LLMs but also offers practical tools for improving the reliability and trustworthiness of AI-generated content.

Limitation

The limitations of this study include the following: (1) A primary challenge in studying uncertainty quantification for long text generation lies in the difficulty of evaluating the generated text. Unlike

classification tasks and short-answer QA, there is no straightforward metric for assessing the quality of generated text. In this study, we employ the factuality score as the primary evaluation metric, thereby leaving other text aspects, such as coherence, cohesion, and creativity, under-explored. Future work could investigate uncertainty scores using more comprehensive evaluation metrics. (2) Fact-checking for long text generation is costly, especially when involving human annotation. While we utilized FACTSCORE as an automated evaluation tool, it still requires numerous API calls to fact-check each piece of information. The development of more cost-effective evaluation metrics or datasets could allow for the expansion of the experimental scope. (3) We discovered that temperature plays a crucial role in measuring uncertainty. A common limitation among all consistency-based uncertainty quantification methods is their effectiveness only at high temperatures. In practice, if some scenarios necessitate a fixed low temperature, these consistency-based UQ methods may not perform effectively.

Ethics Statement

Our research strictly follows ethical guidelines, focusing on data privacy, bias mitigation, and societal impact. We use the publicly available FACTSCORE dataset, which ensures a balanced representation of different nationalities. Our code usage complies with original licensing agreements and is strictly for academic purposes, reflecting our commitment to ethical research standards.

References

01.ai. 2023. Building the next generation of open-source and bilingual llms. https://huggingface.co/01-ai/Yi-34B-Chat. Accessed: 2024-02-05.

Anthropic. 2023. Introducing claude 2.1. Available from Anthropic: https://www.anthropic.com/news/claude-2-1.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

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*.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

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.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. Technical report, Google.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Stephen C Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223.

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

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

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

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

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt blog post.

OpenAI. 2023. Gpt-4 technical report.

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

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

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

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, Alexander Amini, and Alejandro Perez. 2023. Uncertainty-aware language modeling for selective question answering.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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.

A Experiment Setup

For GPT-4 and GPT-3.5, we use the OpenAI API, with specific version `gpt-4-turbo-0125-preview` and `gpt-3.5-turbo-0613`. For Gemini 1.0 Pro, we call the API for developers. For Yi-34B-Chat, Tulu-2-70B (`tulu-2-dpo-70b`), and Vicuna-33B (`vicuna-33b-v1.3`), we use them off-the-shelf and only for inference. We run our uncertainty measurement experiments on A100-SXM-80GB GPUs. For our experiments, we use the following prompt:

> *Tell me a short bio of the person <entity>. Begin with their birth, significant life events, achievements, and contributions. Include their education, career milestones, any notable awards or recognitions received, and their impact on their field or society. Ensure the biography is concise, factual, and engaging, covering key aspects of their life and work.*

From the esitimation of Min et al. (2023), running FACTSCORE costs about $1 of the API cost per 100 sentences. For instance, for 100 generations, each with 5 sentences on average, it costs $5 in total.

B Baselines

LexicalSimilarity (Fomicheva et al., 2020): it computes the similarity between two phrases using metrics like ROUGE scores and BLEU. For our experiment, we utilize BERTScore (Zhang et al., 2020) to enhance performance, computing the average similarity score with other answers.

NumSemSets (Lin et al., 2023): it clusters semantically equivalent answers into the same sets. Initially, the number of semantic sets equals the total number of generated answers. Then it sequentially examines responses, making pairwise comparisons between them, and combines different answers. One of the limitation of this method is that the uncertainty score $U_{NumSemSets}$ can only take integer values. EigValLaplacian is therefore designed to overcome this problem.

EigValLaplacian (Lin et al., 2023): For a similarity matrix S, it calculates the Normalized Graph Laplacian of S using $L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, where D is a diagonal matrix and $D_{ii} = \sum_{j=1}^{m} S_{ij}$ (m is the number of responses). Consequently, the uncertainty score is defined as $U_{EigV} = \sum_{k=1}^{m} \max(0, 1 - \lambda_k)$. This value is a continuous analogue of $U_{NumSemSets}$. In extreme case if adjacency matrix S is binary these two measures will coincide.

DegMat (Lin et al., 2023): it is based on the idea that the total uncertainty of the answers might be measured as a corrected trace of the diagonal matrix D. This is because elements on the diagonal of matrix D are sums of similarities between the given answer and other answers. We thus define uncertainty estimate $U_{\text{Deg}}(x) = \text{trace}(m - D)/m^2$.

Eccentricity (Lin et al., 2023): A drawback of previously considered methods is the limited knowledge of the actual embedding space for the different answers since we only have measures of their similarities. The graph Laplacian, however, can provide us with coordinates for the responses. Denote $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^m$ as the eigenvectors of L that correspond to k smallest eigenvalues. We can efficiently construct an informative embedding $\mathbf{v}_j = [\mathbf{u}_{1,j}, \ldots, \mathbf{u}_{k,j}]$ for an answer $\mathbf{y}_j$. Then it uses the average distance from center as the uncertainty measure, defined as : $U_{Ecc} = \left\| \left[\tilde{\mathbf{v}}_1^T, \ldots, \tilde{\mathbf{v}}_m^T \right] \right\|_2$, where $\tilde{\mathbf{v}}_j = \mathbf{v}_j - \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{v}_\ell$.

SelfCheckNLI (Manakul et al., 2023): As defined in Section 2, SelfCheckNLI primarily functions as a confidence measurement tool, calculating the similarity exclusively between the primary response r_a and the other generated samples. Distinctively, it evaluates $\mathcal{P}(\text{contradict} \mid s, r')$ and focuses solely on $C(x, r_a)$.

C Number of Facts in a Response

Figure 4 shows the average atomic facts provided by various AI models for the FACTSCORE dataset. GPT-4 has the highest average number of atomic facts at 52.24, indicating it provides the most detailed factual responses. Tulu-2-70B follows with an average of 52.17, nearly matching GPT-4 in factual details. GPT-3.5 has an AF of 50.67, showing it also delivers a high level of factual details in its responses. Yi-34B-Chat and Gemini 1.0 Pro have comparatively lower averages, at 45.80 and 42.72 respectively. Vicuna-33B has the lowest AF at 36.20, indicating it offers the least amount of factual information in its responses. Generally, these models provide similar number of atomic facts in their responses.

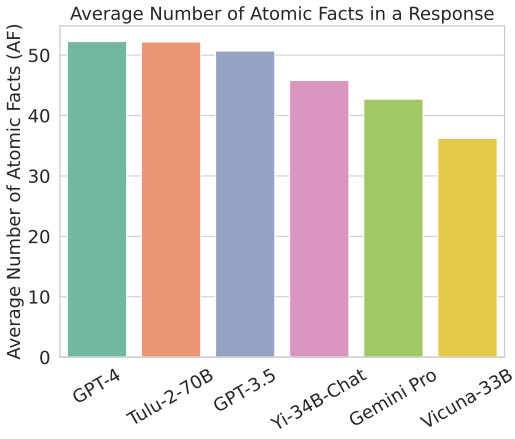

Figure 4: Average number of atomic Facts (AF) in a response for each model.

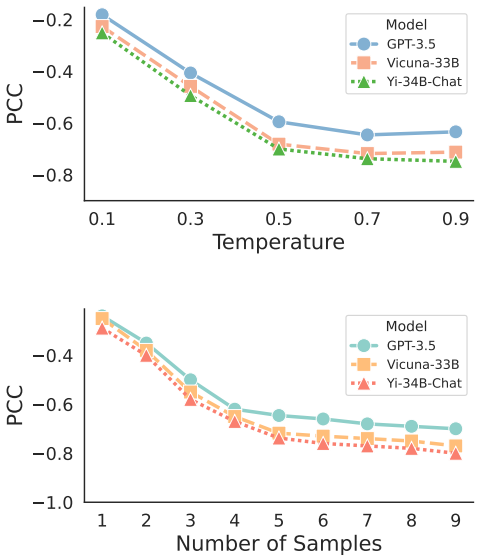

Figure 5: The effect of different temperatures (upper) and the number of samples (lower) on the PCC with LUQ.

C.1 Ablation Study

Temperature As the diversity of content generated by LLMs may be influenced by the temperature setting, we adjust the temperature to test the robustness of our methods. Due to limitations in computational resources and API budget constraints, we selected GPT-3.5, Yi-34B-Chat, and Vicuna-33B for our experiments (refer to Figure 5). Our findings indicate that a lower temperature leads to a weaker correlation score, likely because the generated responses are more uniform, providing limited information for the self-consistency test. As the temperature increases, we observe a strengthening in correlation. However, beyond a certain point, further increases in temperature lead to diminishing improvements and can even result in a weaker correlation. We hypothesize that excessively diverse responses may complicate the NLI process, as a greater number of sentences fail to be supported by other samples.

Number of Samples Previous research on short answer generation (Kuhn et al., 2022; Lin et al., 2023) has demonstrated that an increase in the number of samples correlates with enhanced performance. We investigate whether it also applies to long-text generation and find that with more samples, LUQ shows better performance and PCC scores, which corroborates with previous observations in short-text generation, as depicted in Figure5. Providing a greater number of samples enables the NLI process to predict sentence factuality with higher accuracy. However, a notable drawback of increasing the sample size is the associated rise in computational costs.

D Experiments on FACTSCORE-DIS

To demonstrate the generalization of our proposed LUQ model across various domains, we create a new dataset adopting the methodology used to construct the original FACTSCORE dataset for the disease entities. To differentiate, we refer the original dataset as FACTSCORE-BIO and the new dataset as FACTSCORE-DIS. The detailed information of FACTSCORE-DIS dataset is as follows:

Data Collection Following FACTSCORE-Bio, we use Wikipedia as our main knowledge source. We first select all the diseases names using the following SPARQL codes calling the wiki API. We then removed those diseases with empty Wikipedia pages.

Following FACTSCORE-BIO, we utilized Wikipedia as our primary knowledge source. Initially, we extracted all disease names using the following SPARQL queries to call the Wikidata API. Subsequently, we removed those diseases with empty Wikipedia pages.

```
SELECT ?item ?itemLabel WHERE {
  ?item wdt:P31 wd:Q112193867. # is an
    instance of class of diseases
  SERVICE wikibase:label { bd:
    serviceParam wikibase:language "[
    AUTO_LANGUAGE],en". }
}
```

Frequency For each entity retrieved, we adhere to the methodology described by Min et al. (2023) to assign a frequency label ranging from "Very

Rare" to “Very Frequent" based on an entity’s pageviews. It’s crucial to acknowledge that in the context of diseases, the number of diagnosed cases is commonly used as a metric. However, we opted not to use this metric because our goal is to simulate the distribution of these diseases within the training corpus of LLMs. Relying solely on diagnosed case numbers may underrepresent the prominence of a disease within the corpus. Diseases like Amyotrophic Lateral Sclerosis (ALS), despite their low incidence rate in the population, attract significant global interest and impact. As a result, LLMs may demonstrate extensive knowledge about such diseases, reflecting their visibility in the data on which they are trained, rather than their actual morbidity rates.

After determining the frequencies, we sampled 36 disease entities for each category, amassing a total of 180 data points. Subsequently, we conducted a human evaluation to validate the selected data points, replacing any that were deemed unsuitable with diseases that were more clearly defined and well-documented. Several examples from the dataset are showcased in Table 5.

Frequency	Wikidata ID	Disease Name
Very Freq		
	Q8071861	Zika fever
	Q12199	HIV/AIDS
	Q12152	myocardial infarction
	Q12206	diabetes
	Q12204	tuberculosis
Freq		
	Q154874	yellow fever
	Q188638	mood disorder
	Q159701	glaucoma
	Q1138580	Ewing sarcoma
	Q209369	Hodgkin lymphoma
Medium		
	Q5134736	cloacal exstrophy
	Q247978	anisometropia
	Q2373361	tree nut allergy
	Q778731	pyuria
	Q7900433	urethral syndrome
Rare		
	Q220322	agnosia
	Q2735907	cutis laxa
	Q500695	retinoblastoma
	Q627625	histoplasmosis
	Q1347729	Epstein syndrome
Very Rare		
	Q21505502	spina bifida
	Q1862031	pinguecula
	Q1361850	patulous eustachian tube
	Q4667534	leiomyoma
	Q595010	hypertrichosis

Table 5: Frequency Categories of Diseases

E Human Evaluation

We also engage human annotators to assess the factuality of the generated passages. Although Min et al. (2023) conducted comprehensive experiments to demonstrate the effectiveness of the FACTSCORE framework, we perform a sanity check by directly correlating the annotated passage factuality with uncertainty scores. The annotators are recruited from student volunteers with a Masters Degree in Computer Science. Annotators are compensated above the local minimum hourly wage standard. The instructions provided to the annotators are as follows:

Your task is to evaluate the veracity of each sentence in the provided passage. It is crucial to carefully assess each statement for accuracy and relevance to the main topic.
Steps to Follow:

1. **Read the Passage Thoroughly:** Begin by reading the entire passage to grasp the overall context and the main topic.

2. **Check Each Sentence:** Examine each sentence individually for accuracy and completeness. Determine if the information is factual and supported by reliable sources, and whether the sentence presents a partial truth or is fully accurate.

3. **Scoring:** Assign each sentence a score based on its accuracy, using a specified range (e.g., 1 to 3). Scores should reflect:
    - The sentence is entirely accurate and provides a complete picture. [Highest Score: 3]
    - The sentence is partially correct but may lack context or omit important details. [Mid-Range Score: 2]
    - The sentence is largely inaccurate or misleading. [Lowest Score: 1]

4. **Relevance:** Flag any sentence that does not contribute to or is off-topic as *Not Relevant*.

Guidelines:

- Use reliable sources (e.g. Wikipedia) to verify factual information, maintaining an impartial stance throughout.
- Keep the passage and your assessments confidential.

We randomly selected 50 passages from the responses generated by the Yi-34B-Chat model. We observed a Pearson correlation coefficient of 0.88 between the FACTSCORE factuality score and the human-annotated factuality score. This finding aligns with the results reported by Min et al. (2023), demonstrating that FACTSCORE is a reliable tool in our experiments. Table 6 compares the results of different UQ methods with those obtained using FACTSCORE and human annotation.

Methods	FactScore		Human	
	PCC	SCC	PCC	SCC
LexSimilarity	-67.31	-66.45	-65.63	-64.0
Eccentricity	-26.28	-25.46	-22.58	-25.12
NumSemSets	-26.42	-26.92	-24.28	-23.52
EigValLaplacian	-45.77	-43.94	-43.35	-42.71
DegMat	-38.93	-39.68	-36.81	-31.61
SelfCheckNLI	-68.52	-67.27	-66.13	-69.22
LUQ	**-72.72**	**-71.4**	**-69.02**	**-68.27**

Table 6: Pearson and Spearman correlation coefficients (expressed as percentages) between different factuality scores and various UQ methods on the **FactScore-Bio** dataset using Yi-34B-Chat.