IUQ: Interrogative Uncertainty Quantification for Long-form Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have seen remarkable development but are still prone to hallucination. Developing robust and comprehensive Uncertainty Quantification (UQ) approaches for long-form text generation remains a major challenge. In this paper, we present Interrogative Uncertainty Quantification (IUQ), a novel self-consistency based UQ approach that leverages the language model's tendency to generate semantically coherent yet factually incorrect responses. IUQ builds its estimation on both the trustworthiness of individual facts and their contextual consistency within the model generation. By prompting the language model to go through an interrogate-respond process, IUQ can reliably measure generation-level uncertainties in addition to the model's overall tendency to hallucinate. We evaluate our method with the latest models over diverse model families, and observe a consistent gain in classification metrics.

1 Introduction

Large Language Models (LLMs) have shown remarkable improvement across a diverse range of Natural Language Processing tasks (Brown et al., 2020; Chowdhery et al., 2022; Kamalloo et al., 2023). However, the hallucination problem is still evident, in which the LLMs generate plausible answers that are factually incorrect (Zhang et al., 2023; Huang et al., 2025).

Recent Uncertainty Quantification (UQ) methods effectively measure hallucination within a confined answer space, where the models are prompted to generate short responses or given questions that have definite answers (Kuhn et al., 2023; Lin et al., 2024; Duan et al., 2024; Chen et al., 2024a). These approaches utilize token-probabilities or semantic entailment of the responses to construct uncertainty estimates. However, long-form answers typically include more information, exhibit structure and logic, and contain filler phrases to promote fluency. Therefore, it can be difficult to evaluate entailment relationships between long-answers, and, for the same reason, token-probabilities are much less indicative of hallucination.

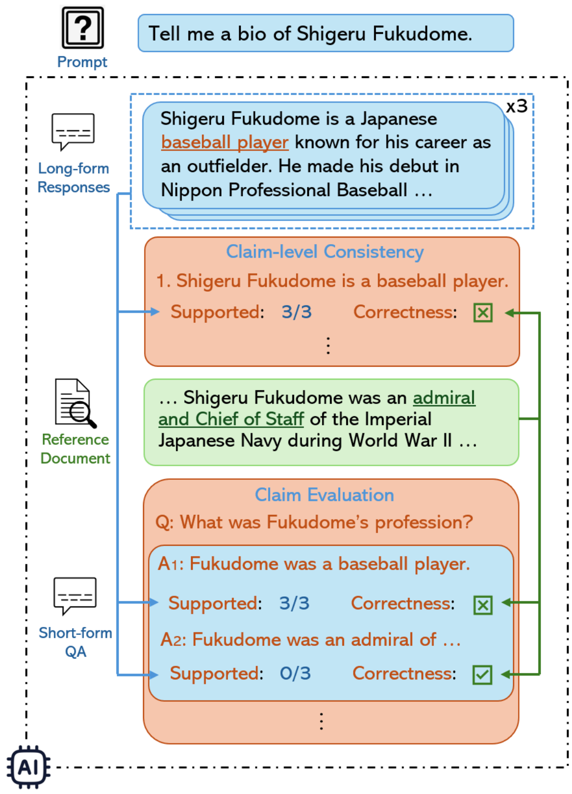


Figure 1: An example of LLM generating multiple consistent but factually incorrect responses. The model hallucinates on a claim it made at the beginning of a long-form response. Even though the LLM has the correct knowledge on the topic, as shown by a separate short-form QA, it continues with its false claim in the long-form responses to fabricate a coherent narrative.

Current efforts on long-form UQ leverage the consistency between LLM generations to evaluate factual correctness. By decomposing the long-form response into sentences or claims, each of them can be compared against additional samples of generation to obtain an uncertainty estimate (Manakul et al., 2023; Zhang et al., 2024; Jiang et al., 2024b;

Wei et al., 2024). However, a concerning scenario arises when LLMs produce consistent yet incorrect responses across multiple queries. As illustrated in Fig. 1, when the LLM is prompted to generate a human biography, it hallucinates on the stated facts at the very beginning of its responses. Although the LLM possesses the right knowledge, as shown by performing a separate QA, it still choose to continue its narrative to maintain coherency. This problem is specific to long-form generations. Without verifying the factual correctness using outside sources, existing UQ methods may misleadingly indicate that the model has low uncertainty over the topic, because each claim does not contradict with any of the sampled responses.

This phenomenon coincides with recent studies that reveal LLMs can exhibit overconfidence over false knowledge (Ren et al., 2025), possibly due to the long-tail distribution of the training data (Mallen et al., 2023; Kandpal et al., 2023). Therefore, it has become increasingly difficult to discern the incorrect information when LLM formulates a plausible response with human-like fluency (Jiang et al., 2024a; Hu et al., 2024; Ji et al., 2024).

Inspired by this observation, we propose a novel UQ framework, called Interrogative-Uncertainty-Quantification (IUQ), to facilitate in-depth probing of the LLM's tendency to hallucinate. IUQ sequentially examines the claims extracted from the response, raising perturbed questions for each to encourage diverse answers from the LLM on specific details. The answers are then checked against all previous claims to identify any conflict. This strategy enforces a stricter constraint on the LLM that the tendency to fabricate a false narrative will be detected. However, an extreme case where IUQ does not work is when the model is trained on false knowledge source. This process is akin to an interrogate-respond scenario where the responder is being questioned continuously to identify any disguise and untruthfulness. Empirically, we found that even minimally rephrased questions can induce semantically diverse answers.

Furthermore, since the questions generated from claims are independent of other generations, IUQ can present a confidence landscape for each generation by simply treating the uncertainty of claims as data points in a time-series. Based on such analysis, we also provide an experimental study on the LLMs' tendency to diverge over their responses to a given topic.

We evaluate IUQ on various model families with their latest models: GPT4o (OpenAI et al., 2024), Qwen2 (Yang et al., 2024), Gemma-3 (Team et al., 2025), Mistral (Jiang et al., 2023), LLaMA-3.1, LLaMA-3.3 and LLaMA-4 (Touvron et al., 2023), with model size up to 72B. We use two widely used datasets tailored for long-form generations: FActScore (Min et al., 2023), which contains entities of human biography, and LongFact (Wei et al., 2024), which contains a prompts set spanning diverse topics. Extensive experiments have shown IUQ's superior performance. Our contribution is the following:

- We highlight the difficulty in accessing long-form generation, as language models often invent or fabricate facts in order to maintain a coherent narrative. This tendency to prioritize coherence poses a significant challenge for uncertainty quantification.

- We propose an Interrogative Uncertainty Quantification (IUQ) paradigm that evaluates a model's long-form responses by probing its knowledge on the topic through fine-grained and diversely-sampled questioning. Extensive experiments have demonstrated the effectiveness of IUQ over diverse topics.

2 Related Work

Uncertainty Quantification Existing approaches of UQ can be roughly categorized into white-box and black-box methods. White-box methods assume the model architecture is partially or completely visible (Kuhn et al., 2023; Nikitin et al., 2024; Duan et al., 2024, Fadeeva et al., 2024), whereas the black-box methods rely only on the input prompts and LLM responses to measure uncertainties (Lin et al., 2024; Xiong et al., 2024; Gao et al., 2024). Our work follows the line of black-box methods. Among them, Tonolini et al. (2024) utilizes a weighted ensemble of semantically equivalent prompts to compute output uncertainty, where the weights are obtained through Bayesian variational inference. Xiong et al. (2024) explores various strategies in prompting, sampling, and aggregating phases to acquire a confidence score from the model. Gao et al. (2024) perturbs the prompts and investigates the variation in responses to measure uncertainty. These mentioned black-box approaches are similar to ours in that we also incorporate perturbation to elicit a greater variety of model responses. The distinction is our

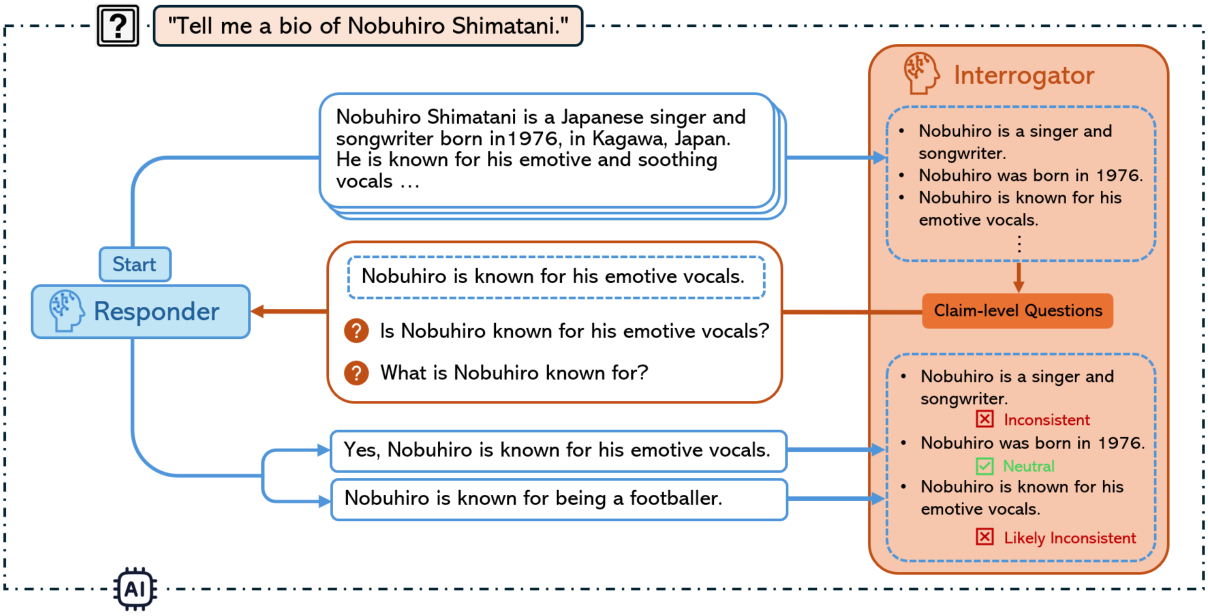


Figure 2: The framework of Interrogative Uncertainty Quantification (IUQ): Responses are sampled from LLMs and decomposed into atomic claims. LLMs then propose several questions for each claim to be answered by itself. The answers are evaluated against the original claims to check for consistency.

method applies to long-form generation, and perturbation is applied at claim-level, letting LLM format its own questions without additional design.

Self-Consistency in LLMs Self-consistency based approaches are proven to be effective in diverse domains associated with LLMs (Pan et al., 2024). Wang et al. (2023) have shown significant improvement in Chain-of-thought prompting by sampling multiple paths and pick the most consistent answer. Shinn et al. (2023) robustly induces better decision-making in various agentic tasks through linguistic feedback. On quantifying uncertainty, the general idea of self-consistency is to perform inter-sample consistency checks, or let LLMs generate verbal-confidence (Manakul et al., 2023; Chen et al., 2024b; Rivera et al., 2024; Jiang et al., 2024b). Kuhn et al. (2023) and Lin et al. (2024) utilize Natural Language Inference models and pairwise entailment to compute uncertainty estimates over a set of sampled responses. Zhang et al. (2024) and Jiang et al. (2024b) let LLM infer the supportiveness of its responses to each claim it has made. Our work is inspired by the similar idea, but we enforce self-consistency both on factual information and contextual coherence.

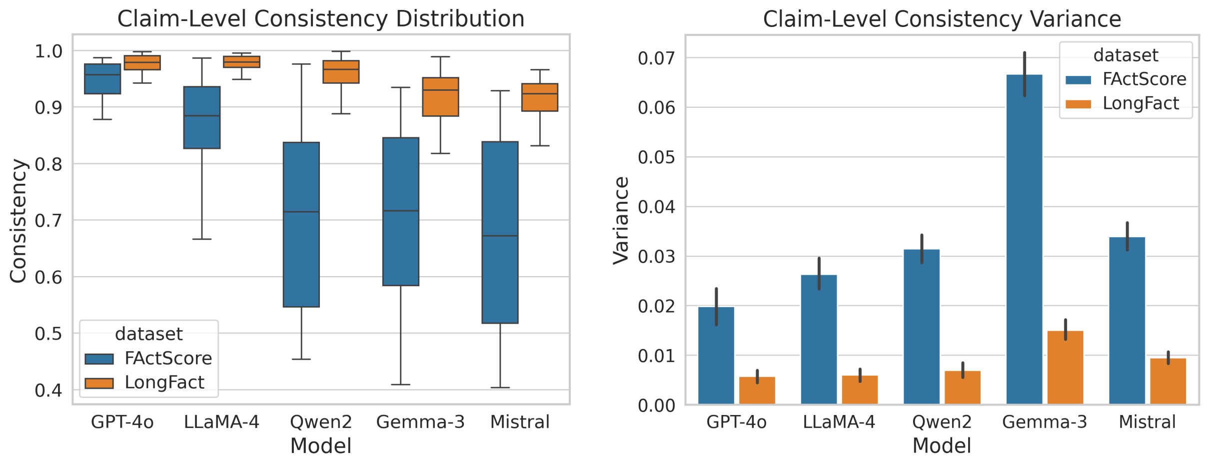
3 IUQ: Interrogative Uncertainty Quantification

IUQ focuses on the fine-grained uncertainty quantification for LLMs, and incorporates prompts-perturbation at claim-level to elicit diverse responses. To perform well, model must answer consistently when generating long-form responses, and when asked separately with the specific details of its generations.

Structurally, IUQ is composed of a responder and an interrogator, with the interrogator continually questioning the responder for the information it has generated, as shown in Fig. 2. In practice, both the responder and interrogator are the same language model. Please refer to Appendix C for the prompts we used in IUQ.

3.1 Response Generation

Given prompt x and a model M, we draw N samples from M with predefined temperature $T = t$. These responses compose a set $\mathcal{R}$ such that $\mathcal{R} = \{R_1, \ldots, R_N\}$, where $R_i = M_{T=t}(x)$ for $i \in \{1, \ldots, N\}$. The generated responses are free-form texts that have variable lengths. To ensure meaningful analysis, we exclude the generations that evidently refuse to respond (e.g. responses of "I don't know", "I cannot provide information"). This is a nontrivial process in traditional natural language processing, so an additional query with LLM will be made to check if its original response is sensible. A data entity will be skipped if at least one response refuses to answer.



(a) Distribution of all claim consistency scores by models.

(b) Variance of claim consistency within generated responses.

Figure 3: Statistics of the claim-level consistency over selected models. (a) The consistency scores of all claims extracted from model responses are collected to view their distribution over datasets. Notably, for FActScore, which contains less-known entities, models exhibit different degrees of inconsistency. (b) The variance shown in the graph is computed over claims within individual model responses, over all data entities. Low variance indicates that a model rarely makes self-contradictory claims.

3.2 Response Decomposition

The output from LLMs typically consists of a few paragraphs of text, which may include redundant information and colloquial language to maintain coherence. Therefore, a common practice is to rely on the LLM to decompose the generated text into a set of claims, with each claim representing the smallest unit that states a fact (Min et al., 2023; Song et al., 2024). However, how to maintain a balance between verbosity (e.g., obvious claims like "He is a man") and ineffectiveness (e.g., failing to decompose and instead returning whole sentences) remains underexplored. Empirically, the best practice is to prompt the LLM with the full generated text and directly extract a list of claims (Jiang et al., 2024b).

As a result, for each response $R \in \mathcal{R}$, we ask the LLM to decompose R into a sequence of claims $\mathcal{C}^R$, making LLM aware of the context by joining prompt x:

$$\mathcal{C}^{R^i} = M_{T=0}(R_i, x) = (C_1^{R^i}, C_2^{R^i}, \dots, C_k^{R^i}), \tag{1}$$

where k is the number of claims returned by the LLM.

3.3 Claim-Level Question-Answering

For each claim, a set of questions are generated in a multi-pass manner, using the same hyper-parameter when sampling the long-form generations. We prompts LLM with a restriction that each question must have its answer contained in the claim to prevent unpredictable behavior. For claim C, the set of generated questions is defined as

$$\mathcal{Q}_C = \{M_{T=t}^{(1)}(C, x), M_{T=t}^{(2)}(C, x), \dots\}, \tag{2}$$

where the number of questions $|\mathcal{Q}_C|$ is a predefined parameter.

IUQ enforces an exact-match filtering rule for the generated questions to preserve as much diversity as possible in the questions set. The filtered question set is defined as

$$\hat{\mathcal{Q}}_C = \{Q \in \mathcal{Q}_C \mid for\ all\ Q_i \neq Q_j\}.$$

For a specific claim, we find the model typically generates multiple paraphrased questions when the claim is 'atomic' enough (e.g. "Nobuhiro was born in 1976"). On the other hand, when the claim contains more than one piece of information (e.g. "Nobuhiro was born in 1976, in Osaka, Japan"), the generated questions tend to be diverse (e.g. "When was Nobuhiro born?", "Where was Nobuhiro born?"), thus complementing the claim to achieve finer-grained analysis .

IUQ then queries the LLM with the generated questions, passing the original prompt as context. We sample several answers for each question in $\hat{\mathcal{Q}}_C$. The set of answers for a claim C is defined as

$$\mathcal{A}_C = \{M_{T=t}(Q, x) \mid Q \in \hat{\mathcal{Q}}_C\}. \tag{3}$$

We empirically observe that, when asked for detailed information using claim-level questions $\hat{\mathcal{Q}}_C$, LLMs can produce more accurate answers.

However, as shown in Fig. 1, LLM can still hallucinate when it possesses the correct knowledge in its parametrized memory.

3.4 Claim-level Consistency

IUQ builds its uncertainty estimation on factual and contextual consistency, which is quantified by performing consistency checks between answers $\mathcal{A}_{C_i}$ and all previous claims $C_{i\leq}$, including C_i. Denoting the consistency score for claim C_i as $S_C(C_i)$, one way to compute $S_C(C_i)$ is through exhaustive check between every pair of claims and answers, using either Natural Language Inference model as in (Lin et al., 2024), or the LLM M. However, the cost of exhaustive checks significantly outweighs the performance gain, so we ask the model M to return a numerical value representing the degree of consistency between each answer $A \in \mathcal{A}_{C_i}$ and claims $C_{i\leq}$. The consistency score is then:

$$S_C(C_i) = \frac{1}{|\mathcal{A}_{C_i}|} \sum_{A \in \mathcal{A}_{C_i}} M_{T=0}(A, C_{i\leq}, x), \quad (4)$$

where $M_{T=0}(A, C_{i\leq}, x) \in [0, 1]$. We present the statistics of consistency scores over all data instances in our experiment, for selected models, in Fig. 3.

When the context and reasoning-chain grow longer over time, LLMs performance can fail catastrophically (Chen et al., 2023; Kotha et al., 2024). Similarly, hallucination accumulates and lead to further inconsistencies. IUQ propagates the impact of inconsistency in claim C_i to subsequent claims by superimposing an exponentially decaying function. Defining the inconsistency over the sequence of claims $\mathcal{C}$ as

$$1-S_C(\mathcal{C}) = (1-S_C(C_1), \ldots, 1-S_C(C_k)). \quad (5)$$

The inconsistency impact is then defined as the convolution between the claim-level inconsistency and the exponential decay function $f(k)$:

$$I(\mathcal{C}) = f(k) * (1 - S_C(\mathcal{C})) \quad (6)$$

With a predefined constant λ, we use the exponential decay, defining

$$f(k) = e^{-\lambda i} \; for \; i = 0, 1, \ldots, k. \quad (7)$$

4 Uncertainty Estimation with Claim-level Consistency

In this section we present several metrics to evaluate claim-level uncertainty. First we show the sampled responses can be used with Eq. 6 to produce an uncertainty estimate adjusted for inconsistency in claims. We also present a metric that utilize consistency between answers in set $\mathcal{A}_C$. Additionally, IUQ allows token-probability based methods to be explored in long-form generations, by directly operating on the short-form answer in set $\mathcal{A}_C$. We present two fundamental metrics, perplexity (PPL) (Jelinek et al., 2005) and predictive entropy (PE) (Kadavath et al., 2022).

4.1 Response-Claim Entailment

In long-form generation UQ, existing method utilize LLM to infer whether the response entails a claim or sentence (Jiang et al., 2024b; Zhang et al., 2024; Wei et al., 2024). We use this "entailment score" combined with the inconsistency impact $I(C)$ to produce a fine-grained and context-aware metric for uncertainty estimation.

Following Jiang et al. (2024b) and Zhang et al. (2024), we define the response entailment score (S_R) for claim C as the ratio between number of entailment and the total number of responses

$$S_R(C) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[R_i \Rightarrow C], \quad (8)$$

where the entailment relation ($\Rightarrow$) is inferred by the model M by asking whether the response R support the claim C. The uncertainty estimation based on $S_R(C)$ is then

$$U_R(C) = 1 - S_R(C). \quad (9)$$

For the motivation discussed in section 1, we utilize the claim-level inconsistency impact $I(\mathcal{C})$ as a measure of trustworthiness for a single response. Therefore, we combine $I(\mathcal{C})$ with the inter-generation consistency $S_R(C)$ to obtain a new uncertainty metric

$$U_S(C) = S_R(C) \cdot I(C), \quad (10)$$

where $I(C)$ is an element in the sequence $I(\mathcal{C})$.

4.2 Answer-Claim Entailment

Similar to response-claim entailment, the consistency of answers in $\mathcal{A}_C$ indicates the LLM's confidence on its knowledge. Therefore, we investigate whether the entailment of the short-form answers are good indicators of hallucination, given

	Metric	GPT-4o	LLaMA-3.1	LLaMA-3.3	LLaMA-4	Qwen2	Gemma-3	Mistral
FActScore	U_V	0.649	0.640	0.595	0.620	0.768	0.768	0.659
	U_R	0.732	0.819	$\underline{0.847}$	0.809	0.901	0.820	$\underline{0.880}$
	U_{RV}	**0.750**	$\underline{0.826}$	0.846	$\underline{0.810}$	0.915	$\underline{0.843}$	0.860
	U_{GC}	$\underline{0.749}$	0.822	0.843	$\underline{0.810}$	$\underline{0.929}$	0.840	0.862
	U_E	0.591	0.648	0.593	0.591	0.581	0.629	0.587
	U_P	0.620	0.701	0.641	0.649	0.675	0.677	0.736
	U_A	0.617	0.634	0.633	0.684	0.838	0.706	0.799
	U_S	0.748	**0.847**	**0.875**	**0.833**	**0.932**	**0.867**	**0.913**
LongFact	U_V	0.599	0.611	0.567	0.680	0.632	0.574	0.576
	U_R	0.705	0.736	$\underline{0.714}$	0.759	0.791	0.656	$\underline{0.733}$
	U_{RV}	0.721	$\underline{0.748}$	**0.728**	$\underline{0.762}$	$\underline{0.792}$	$\underline{0.660}$	0.709
	U_{GC}	$\underline{0.722}$	0.724	0.702	0.755	0.782	0.639	0.712
	U_E	0.582	0.618	0.591	0.615	0.560	0.620	0.609
	U_P	0.597	0.638	0.578	0.608	0.591	0.596	0.617
	U_A	0.592	0.573	0.591	0.601	0.659	0.557	0.625
	U_S	**0.733**	**0.749**	0.722	**0.780**	**0.806**	**0.689**	**0.743**

Table 1: AUROCs of the uncertainty quantification metrics proposed by IUQ and other baseline methods across various instruction-tuned LLMs. Bold-text indicates the best result, and underline indicates second best result. The experimental setup is detailed in Section 5.1 and he baseline methods are described in Section 5.3. AUPRCs of the same experiments are reported in Appendix B.

that these answers collectively represent the information in LLM's long-form responses.

For each question Q in $\hat{\mathcal{Q}}_C$, denoting the set of answers to Q as $\mathcal{A}_Q$, we define the uncertainty estimate for claim C based on answer-consistency as

$$U_A = 1 - \frac{1}{|\hat{\mathcal{Q}}_C|} \sum_{Q \in \hat{\mathcal{Q}}_C} M_{T=0}(\mathcal{A}_Q, x), \quad (11)$$

where $M_{T=0}(\mathcal{A}_Q, x) \in [0, 1]$ is the consistency estimate given by LLM.

4.3 Answer Token-probability

By characterizing the language generation as a classification problem, the uncertainty of an response can be measured by the entropy of the prediction (Wellmann and Regenauer-Lieb, 2012; Kuhn et al., 2023). In general, the predictive entropy (PE) for input x is the conditional entropy (H) of the output R:

$$H(R|x) = -\sum_i p(z_i|x) \log p(z_i|z_{<i}, x), \quad (12)$$

where z_i is the i-th token generated by the LLM and $z_{i<}$ is all the tokens before z_i.

Token-probability based approaches are commonly adopted in short-form UQ. However, they are not employed in existing approaches of long-form UQ, as the LLM response contain noisy tokens but meaningful ones are sparse.

On the other hand, we propose an indirect approach, using token-probability of the answers in the set $\mathcal{A}_Q$. Since the context is bound to claim C and the answers for $Q \in \hat{\mathcal{Q}}_C$ is much shorter than the long-form response R, their token-probabilities are indicative of the LLM's uncertainty over the claim C. We define the uncertainty estimate built on entropy H as

$$U_E = \frac{1}{|\hat{\mathcal{Q}}_C|} \sum_{Q \in \hat{\mathcal{Q}}_C} \frac{1}{|\mathcal{A}_Q|} \sum_{A \in \mathcal{A}_Q} H(A|C). \quad (13)$$

We also utilize perplexity (PPL) (Jelinek et al., 2005) to measure uncertainty of the answers in $\mathcal{A}_C$, which is defined as

$$PPL(A) = exp(-\frac{1}{t} \sum_i^t \log p(z_i|z_{<i})), \quad (14)$$

where t is the number of tokens in answer A. Similarly, the uncertainty estimate of claim C using answer-perplexity $PPL(A)$ is then defined as

$$U_P = \frac{1}{|\hat{\mathcal{Q}}_C|} \sum_{Q \in \hat{\mathcal{Q}}_C} \frac{1}{|\mathcal{A}_Q|} \sum_{A \in \mathcal{A}_Q} PPL(A). \quad (15)$$

5 Experiments

5.1 Datasets and Annotation

We evaluate our proposed uncertainty estimation methods on FActScore (Min et al., 2023) and LongFact(Wei et al., 2024). For each dataset, we select 50 entities, which are decomposed into claims as described in Section 3.2. Based on the content of the claim, we let LLM generate 3 context-related questions, and for each question sample 3 answers. The statistics of data generated by GPT-4o on FActScore and LongFact are shown in Table. 2.

FActScore			
Responses	**Claims**	**Questions**	**Answers**
235	4759	10433	31299
LongFact			
Responses	**Claims**	**Questions**	**Answers**
250	4276	9954	29862

Table 2: Statistics of the total numbers of generated items by GPT-4o on the FActScore and LongFact datasets.

FActScore (Min et al., 2023) contains entities of human biography, where each of them has a dedicated Wikipedia article. We randomly select 50 entities. To evaluate the factuality of claims, IUQ employs a similar method in Min et al. (2023), labeling each fact as "correct" or "incorrect" based on the corresponding Wikipedia article. The factuality evaluation is independent of the uncertainty estimation process, and is performed using GPT-4o due to its low error rate.

LongFact (Wei et al., 2024) is a prompt set comprising thousands of questions spanning 38 topics. We choose LongFact to test our uncertainty metrics since it complement FActScore on the domains of topics. While FActScore verifies the correctness of atomic claims through reference passages from Wikipedia, the approach proposed in (Wei et al., 2024) does so by performing web-search. To maintain consistency and reproducibility, we manually select 50 entities of diverse topics in LongFact that have dedicated Wikipedia articles, and employ the same method we used for FActScore to evaluate the factuality of claims.

5.2 Models and Parameters

We conduct experiments over the latest models across various model families, including GPT4o (OpenAI et al., 2024), LLaMA-3.3 and LLaMA-4 (Touvron et al., 2023), Qwen2 (Yang et al., 2024), Gemma-3 (Team et al., 2025), and Mistral (Jiang et al., 2023), with model size up to 72B. We set the temperate $t = 1.0$ to sample 5 long-form responses for each entity in dataset, and use greedy search (temperature $t = 0$) to evaluate the correctness of the claims.

5.3 Baselines

Following prior works (Tian et al., 2023; Jiang et al., 2024b), we employ the LLM's verbal confidence on claims as an uncertainty metric. This metric directly prompts the LLM with the claim C to rate its confidence on the claim from 0 to 1. The confidence rating is then compared directly with the ground-truth label. We denote this metric as U_V and the result is shown in Table. 1. Additionally, similar to Eq. 10, we utilize the verbal confidence as a weight to the response entailment score defined in Eq. 8 to obtain a new metric U_{RV}. The results of these metric are shown in Table. 1.

We also adopt the graph-based uncertainty metric defined in Jiang et al. (2024b). In this work, a bipartite graph is built from the entailment relation in Eq. 8, where each claim is a node and each entailment relation between claim and generation implies an edge. We directly apply the procedure in Jiang et al. (2024b) to compute the "closeness" of a node as one uncertainty metric. We denote this metric as U_{GC} and show the result in Table. 1.

5.4 Evaluation Metrics

Following prior works (Manakul et al., 2023; (Kuhn et al., 2023); Jiang et al., 2024b), we formulate the evaluation process as a classification problem, where the predicted probability of claims being correct is given by our uncertainty metrics, and the procedure to obtain ground-truth labels is detailed in Appendix A. We adopt the area under the receiver operator characteristic curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) to classify the performance of the uncertainty metrics.

5.5 Ablation Study

In this section, we present an experimental study to show the effectiveness of our claim-consistency paradigm (Section 3.4). Firstly, we illustrate that claim consistency scores S_C capture the model's self-contradictory behavior in its response, by comparing the performance of baselines and IUQ metrics. Secondly, by evaluating the influence of using

Method	FActScore				LongFact			
	GPT-4o	LLaMA-4	Qwen2	Gemma-3	GPT-4o	LLaMA-4	Qwen2	Gemma-3
No-ErrP	**0.748**	0.831	0.931	0.847	0.724	0.771	0.807	0.678
Lin-ErrP	0.732	0.809	0.917	0.834	0.725	0.763	0.801	0.682
Acc-ErrP	0.713	0.800	0.889	0.804	0.723	0.754	0.800	0.675
Exp-ErrP(U_S)	**0.748**	**0.833**	**0.932**	**0.867**	**0.733**	**0.780**	**0.806**	**0.689**

Table 3: Ablation study on the impact of claim consistency score with different error propagation (ErrP) function. The presented values are AUROCs of the uncertainty quantification metric U_S.

different error propagation functions, we show that the exponential-decay weighting is the most effective approach to estimate uncertainty in long-form generations. Lastly, we evaluate the sensitivity of our uncertainty metrics on the number of generated responses. We present ablation results on selected models in Table. 3 and Fig. 4. Additional experiments are reported in Appendix B.

Effectiveness of Claim Consistency Score The claim consistency score (Eq. 4) captures the fabricated information in long-form responses by enforcing a consistency check between claims and context. To demonstrate its effectiveness, we compare its performance with verbal-confidence, which is the confidence score elicited from the model. We also use verbal-confidence to weigh the response entailment score (Eq. 8) to compare with U_S, which is weighted by the claim-consistency score. These two uncertainty metrics are denoted as U_V and U_{RV} and the experiments results are shown in Table. 1.

The result illustrates that although U_V is not a strong baseline, U_{RV} shows surprisingly good performance over all tested models. This finding consolidates our motivation that LLM has limitations in identifying its own weaknesses. Without sampling multiple responses and performing fine-grained analysis, it is risky to trust LLM responses, especially in long-form generation.

Additionally, we present the statistics of claim consistency scores and consistency variance within generation in Fig. 3.

Effectiveness of Inconsistency Propagation The inconsistency impact (Eq. 6-Eq. 7) serves to propagate the impact of an inconsistent claim to subsequent claims. In this section, we investigate the influence of different propagation functions Eq. 7 on the uncertainty estimation performance. The results are shown in Table. 3 and the notations used are explained as follows: (1) No-ErrP: No error is propagated to subsequent claims, and we build the uncertainty estimate solely on the claim consistency score. (2) Lin-ErrP: Linear error propagation, where the inconsistency is superimposed with a linear function $f(k) = mi + b$ for $i = k, k-1, \ldots, 1$, where $m > 0$ and b is a constant. (3) Acc-ErrP: accumulative error propagation, where $I(\mathcal{C})$ (Eq. 6) is defined as the cumulative sums of the claim-level inconsistency Eq. 5.

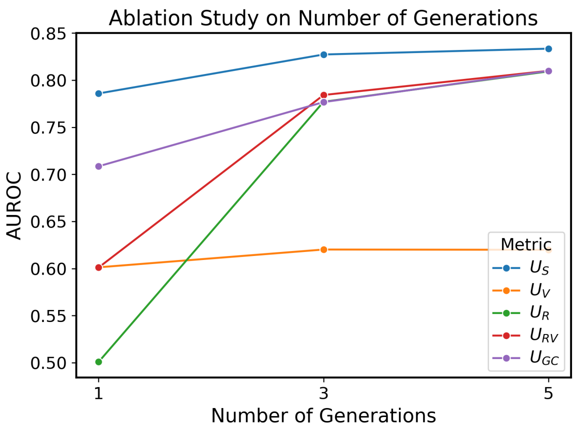


Figure 4: AUROCs of uncertainty metric U_S and baselines on different numbers of sampled responses.

Influence of Number of Generations We show the influence of the number of sampled responses on our uncertainty metric U_S and baseline methods in Fig. 4. Except for U_V which is built on verbalized confidence, all other metrics utilize the response-entailment score (Eq. 8). Consequently, more sampled responses will lead to more accurate classification.

6 Conclusion

In this work, we identify the problem of language models favoring coherence over factual correctness in long-form generation. We propose Interrogative Uncertainty Quantification (IUQ), a fine-grained approach that builds on claim-level contextual consistency to estimate the uncertainty in long-form responses. Empirical results demonstrate the effectiveness of IUQ over diverse model families.

7 Limitations

Our method relies on LLMs' reasoning and question-answering ability to perform most parts of our pipeline. A major issue is the additional hallucination introduced in the processing, and there is no guarantee that such hallucination will be detected. This problem is partially addressed by adapting the source code to incorporate model API's support for structured output, but is still limited to a few powerful models. Additional measures we take are to manually parse the model's output and perform basic sanity checks to ensure model responses are at least minimally sensible.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *Preprint*, arXiv:2307.09009.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

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

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *Preprint*, arXiv:2403.04696.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346. Association for Computational Linguistics.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

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

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024a. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053. Association for Computational Linguistics.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. 2024b.

Graph-based uncertainty metrics for long-form language model generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3697–3715. Association for Computational Linguistics.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 114–126. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12229–12272. Association for Computational Linguistics.

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

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

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

J. Florian Wellmann and Klaus Regenauer-Lieb. 2012. Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models. *Tectonophysics*, 526-529:207–216. Modelling in Geosciences.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Appendix

A Correctness Evaluation

FActScore We evalute the factual correctness of claims extracted from long-form responses using an adapted approach in Min et al. (2023). For each topic, first, the reference article is fetched from Wikipedia and broken into chunks of passages. The passages and claims are vectorized using sentence-transformer gtr-t5-large Ni et al. (2021). Based on the relevance of the claim and the reference passage, the passages are returned based on similarity. The correctness of claims are evaluated by GPT-4o and labeled as either "correct" or "incorrect".

LongFact LongFact is a dataset that contains 2,280 prompts that solicit long-form responses across 38 selected topics, including arts, chemistry, historical events and etc. Wei et al. (2024) propose to use Google Search API to exhaustively verify the factuality for each fact presented in the long-form response. However, to maintain consistency and reproducibility, we manually selected 50 prompts from LongFact that have dedicated Wikipedia entries, and use the same method for FActScore to evaluate factual correctness. Example prompts and Wikipedia entities for LongFact are shown in Table. 4.

B Additional Experiments

AUPRCs We show the experiment results of Table. 1 using Area Under the Precision-Recall Curve (AUPRC) in Table. 5. AUPRC measures how well a model separates the positive class from the negative class, focusing on the performance for the positive class. On the other hand, AUROC looks at the trade-off between true positive rate and false positive rate, and considers both classes equally.

Claim Consistency Landscapes The claim consistency score computed in Eq. 4 encapsulates the self-consistency of the claim and consistency in the context of the generated response. Since we can assign a score for every claim within a response, the scores themselves imply the LLM's hallucination degree across individual responses. Therefore, we can treat the claim consistency scores as time-series and visualize them in Fig. 5.

To accommodate for multiple samples of responses, each of different lengths and thus different numbers of claims, we interpolate the claim-consistency scores of shorter responses linearly to construct sets with equal number of elements. The sequence of claim-consistency scores representing a single topic is then the average of the interpolated sequences.

For generations across data instances, interpolation is not ideal due to LLM's varied knowledge on different topics. Therefore, we simply pad the responses across data instances with trailing zeros.

C Prompts

We follow the structure of Fig. 2 to list the prompts used in IUQ (Table. 6 - Table. 7). Generally, they include the prompts used on generating long-form responses, performing claim-level question answering, and evaluating consistency.

LongFact Prompt	**Wiki-entry**
Can you describe the occurrences during the Watts Riots?	Watts riots
Can you provide an overview of the International Monetary Fund?	International Monetary Fund
Could you explain what the Kepler Space Telescope is?	Kepler space telescope

Table 4: Example LongFact prompts and corresponding Wikipedia entries.

	Metric	GPT-4o	LLaMA-3.1	LLaMA-3.3	LLaMA-4	Qwen2	Gemma-3	Mistral
FActScore	U_V	0.844	0.734	0.634	0.704	0.518	0.583	0.498
	U_R	0.884	0.868	0.847	0.841	0.774	0.646	0.808
	U_{RV}	**0.897**	0.878	0.848	<u>0.850</u>	0.804	0.719	<u>0.814</u>
	U_{GC}	<u>0.895</u>	<u>0.889</u>	<u>0.850</u>	<u>0.850</u>	<u>0.849</u>	<u>0.723</u>	0.810
	U_E	0.756	0.776	0.657	0.717	0.438	0.478	0.542
	U_P	0.729	0.805	0.687	0.743	0.490	0.494	0.640
	U_A	0.837	0.757	0.659	0.736	0.648	0.498	0.695
	U_S	**0.897**	**0.908**	**0.896**	**0.870**	**0.857**	**0.767**	**0.863**
LongFact	U_V	0.901	0.838	0.837	0.891	0.881	0.854	0.893
	U_R	0.929	0.893	0.891	0.920	0.937	0.881	0.934
	U_{RV}	0.934	0.902	0.899	<u>0.932</u>	<u>0.940</u>	0.884	0.933
	U_{GC}	<u>0.943</u>	<u>0.907</u>	<u>0.900</u>	0.930	0.936	<u>0.890</u>	<u>0.937</u>
	U_E	0.858	0.860	0.857	0.885	0.872	0.889	0.915
	U_P	0.850	0.865	0.851	0.883	0.881	0.877	0.913
	U_A	0.904	0.844	0.851	0.875	0.893	0.850	0.915
	U_S	**0.944**	**0.912**	**0.901**	**0.937**	**0.950**	**0.909**	**0.943**

Table 5: AUPRCs of the uncertainty quantification metrics proposed by IUQ and other baseline methods across various instruction-tuned LLMs. Bold-text indicates the best result, and underline indicates second best result.

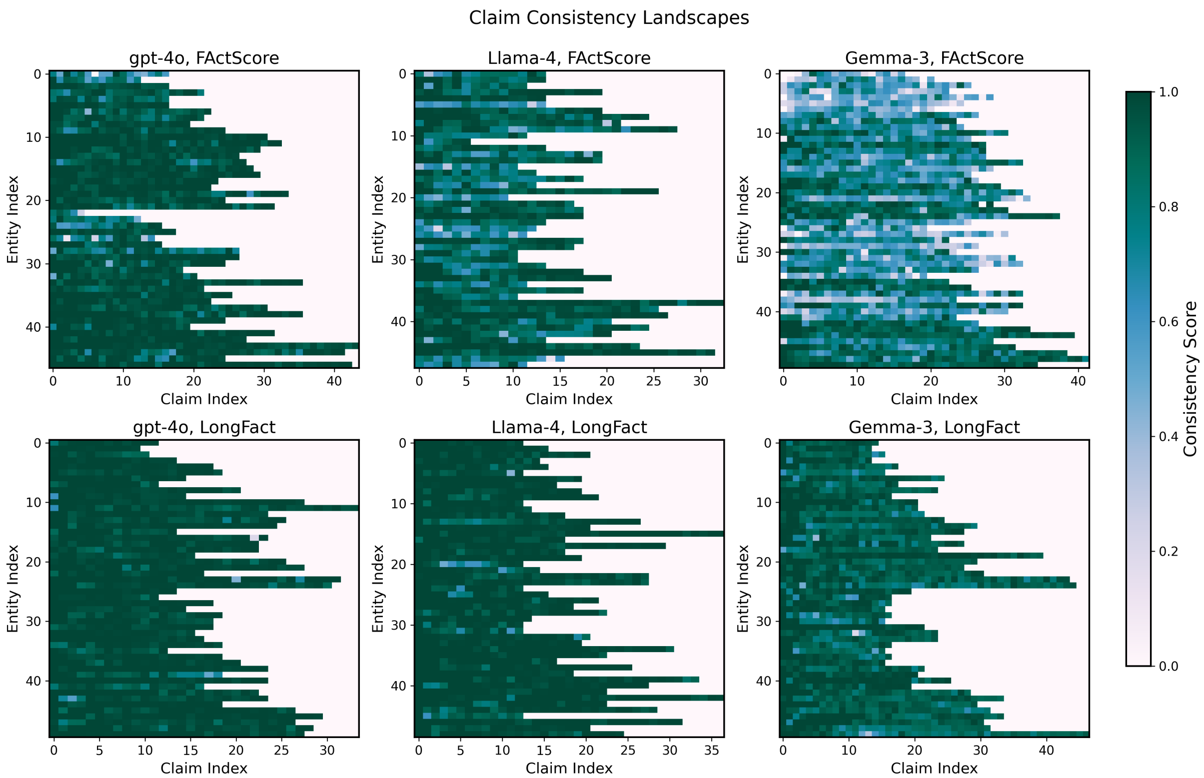


Figure 5: Claim-consistency scores within individual generations. The x-axis is the index of the claim made in LLM's response, and y-axis is the index of the topic in datasets. Results for FActScore and LongFact are shown with selected models.

Prompt	Role
"Answer the following question in plain text, without any additional formatting: {prompt}"	Generate response
"Given context and a paragraph of text, deconstruct the text into the smallest possible standalone and self-contained facts without semantic repetition. Each fact should come from the text and must be related to the context. <Context>{context}</Context> <Text>{text}</Text> Return ONLY a list of facts, with no additional text."	Decompose response
"Given context and a claim, generate one specific, clear question that has its answer contained in the claim. The generated question must be self-contained and related to the context. Return only the question, with no additional text. Context: {context} Claim: {claim}"	Claim-level questions
"Answer the following question based on the given context. Format your answer in one sentence: Context: {context} Question: {question} Answer: "	Question answering
"You will be given a statement and a context. Please estimate how much of the context contradicts the statement? Your final answer should be a percentage number between 0 and 100, representing the percentage of the context that contradicts the statement. <Statement> {statement} </Statement> <Context> {context} </Context> Return your answer as a percentage number ONLY, with no additional text."	Claim-level consistency

Table 6: Prompts used in IUQ.

Prompt	Role
"Is the following claim supported by the reference passage? Choose your answer from <supported/not supported>. <Claim>{claim}</Claim> <Reference>{reference}</Reference>"	Evaluate correctness

Table 7: Prompts used in IUQ cont..