IUQ: Interrogative Uncertainty Quantification for Long-Form Large Language Model Generation

Anonymous ACL submission

Abstract

Despite the rapid advancement of Large Language Models (LLMs), uncertainty quantification in LLM generation is a persistent challenge. While recent methods have achieved remarkable accuracy by limiting LLMs to generate short or constrained answer sets, the most common usage of LLMs is for long and free-form generation, where the underlying semantics are multifaceted and linguistic structure is complex. One major complication emerged from this use case is the tendency of LLMs to produce semantically coherent yet factually incorrect responses. To tackle this challenge, this paper introduces **Interrogative Uncertainty Quantification (IUQ)**, a novel framework that leverages inter-sample consistency and intra-sample faithfulness to quantify the uncertainty in long-form LLM outputs. By utilizing an interrogate-respond paradigm, our method provides reliable measures of claim-level uncertainty and the model's faithfulness. Experimental results across diverse model families and model sizes demonstrate that IUQ outperforms baselines by at least 1.7% on average over long-form generation datasets.

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, LLMs remain susceptible to hallucination, as they 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 answer multiple-choice questions (Kuhn et al., 2023; Lin et al., 2024; Duan et al., 2024; Chen et al., 2024a). These approaches leverage token-probabilities or semantic entailment between responses to construct uncertainty estimates. However, in the scenario of free-form generation, where the response lengthens and exhibits structure and logic, it can be difficult to evaluate the entailment relationships between long answers, and aggregating token-probabilities becomes less indicative of uncertainty.

Current work on long-form UQ involves evaluating the semantic consistency between LLM responses. The long-form response is decomposed

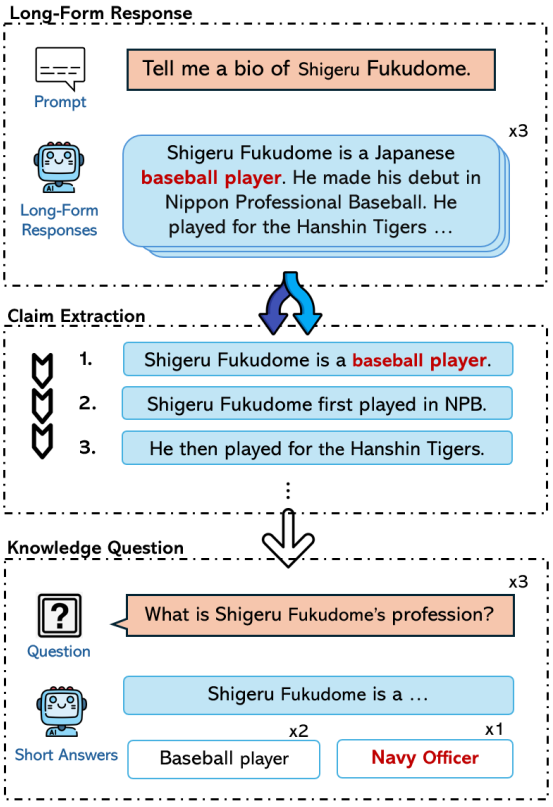


Figure 1: An example of LLM generation on biography. The model incorrectly identifies the individual's profession and fabricates a complete biography to maintain logical consistency. When multiple outputs are sampled, they exhibit a high level of similarity even though the biographies are based on false information. The model is then shown to be uncertain about the subject in a separate session asking specifically about the individual's profession.

into sentences or claims, which are then compared against additional sampled responses to obtain uncertainty estimates (Manakul et al., 2023; Zhang et al., 2024; Jiang et al., 2024b; Wei et al., 2024). However, the following observation requires a closer look at the fine-grained contextual dependence in LLM generations: long-form output may differ across samples but is rarely self-contradictory due to next-token conditioning on the preceding context. Therefore, a critical challenge occurs when models fabricate information for the sake of logical consistency. As illustrated in Fig. 1, when the LLM is prompted to provide information on a historical figure, it mistakenly identifies the individual's profession but continues to generate a plausible story to maintain logical consistency. On the other hand, when asked specifically about the individual's profession, the model returns inconsistent answers, showing uncertainty about the subject. The tendency for LLM to fabricate information is not captured by current UQ methods, which only evaluate consistency across sampled generations without testing the models' real knowledge on the subject.

Recent studies reveal LLMs hallucinate on facts that are present in the training data (Jiang et al., 2024a). When the topic is underrepresented, LLMs may even be overconfident about false knowledge (Kandpal et al., 2023; Mallen et al., 2023; Ren et al., 2025). These findings corroborates our observation that LLMs are not always faithful, and it has become increasingly difficult to identify incorrect information, as LLMs are more capable of formulating plausible responses (Hu et al., 2024; Ji et al., 2024).

To tackle this challenge, we first differentiate the claims in a generative context from the model's knowledge. Specifically, we incorporate an interrogator LLM to construct tailored short questions for each factual claim made in the long-form output. As a result, the original long-form response is decomposed into atomic knowledge decoupled from the generative context. To demonstrate correct knowledge, the model needs to answer consistently to each independent question and not provide answers that contradict the original claims. Naturally, the amount by which the model returns contradictory answers constitutes a measure of the tendency to fabricate information.

We then propose a novel UQ framework: Interrogative Uncertainty Quantification (IUQ) to facilitate fine-grained probing of long-form LLM responses. Different from other UQ methods that only implement inter-sample consistency, IUQ also enforces intra-sample claims consistency through independent question-answering. The strategy is analogous to an interrogate-respond scenario in which the responder is continuously questioned by the interrogator to screen untruthfulness. Furthermore, since the short questions extracted from claims are independent of other sampled generations, IUQ presents a confidence landscape for each generation by viewing the quantified uncertainty of claims as data points in a time-series.

We evaluate IUQ on various model families of diverse sizes: GPT4o (OpenAI et al., 2024), Qwen2 (Yang et al., 2024), Gemma-3 (Team et al., 2025), Mistral (Jiang et al., 2023), LLaMA-3.1, and LLaMA-3.3, with model sizes range from 24B up to 72B. We use two datasets tailored for long-form generation: FActScore (Min et al., 2023), which contains items of biography, and LongFact (Wei et al., 2024), which contains prompt sets on topics of art, science, and so on. Extensive experiments have shown IUQ's superior performance. Our contribution is the following:

- We propose the Interrogative Uncertainty Quantification (IUQ) workflow that evaluates long-form responses through fine-grained probing of claim-level knowledge. Extensive experiments have demonstrated the effectiveness of IUQ over diverse model families.
- We highlight an under-explored phenomenon of long-form LLM generation where the models fabricate factual information to maintain logical consistency, and present a quantitative analysis of this tendency at the claim-level.

2 Related Work

Uncertainty Quantification Existing UQ methods 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., 2024a), whereas the black-box methods rely 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

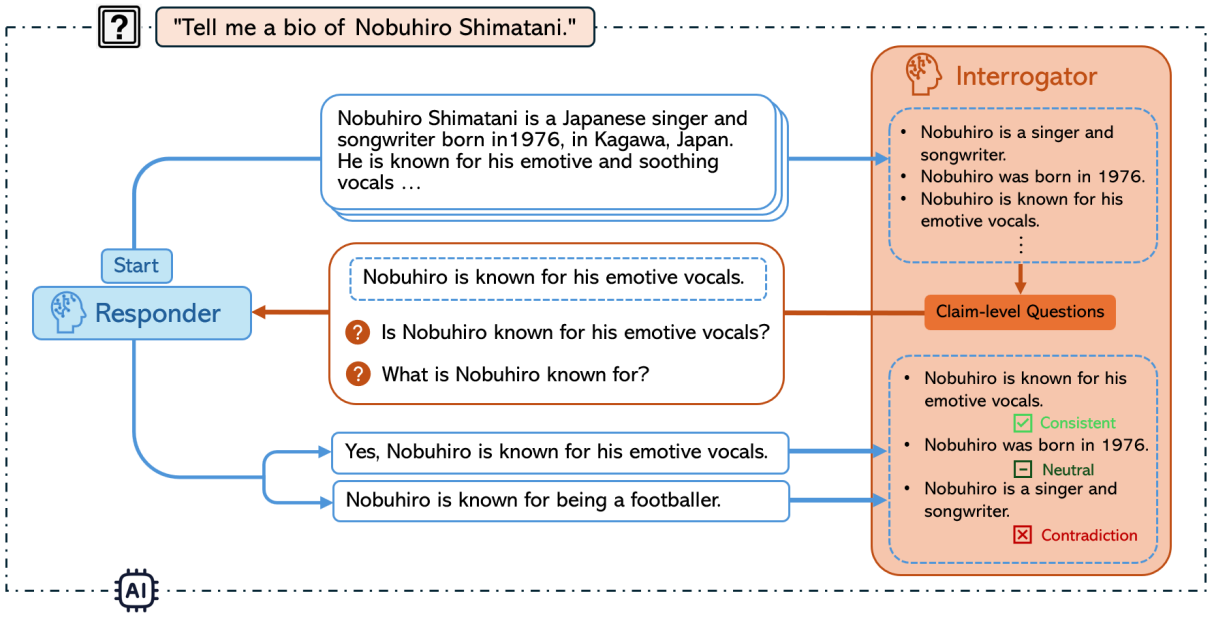


Figure 2: The framework of Interrogative Uncertainty Quantification (IUQ): Responses are sampled from LLMs and decomposed into atomic claims. The model then answers questions that address the information contained in the claim in a separate session without context. The model is unfaithful about a claim if the claim contradicts the corresponding answers, which represent the model's knowledge on the subject.

weights are obtained through Bayesian variational inference. Xiong et al. (2024) explores various strategies in prompting, sampling, and aggregating phases to acquire confidence scores from the model. Gao et al. (2024) perturbs the prompts and measures the semantic variation in responses. IUQ is distinct from these methods that it applies to long-form LLM outputs and decouples the contextual dependence to examine the models' real knowledge.

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 picking the most consistent answer. Shinn et al. (2023) robustly induces better decision-making in various agentic tasks through linguistic feedback. For quantifying uncertainty, a general workflow for consistency estimation is to perform inter-sample consistency checks or let the models output 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) utilize LLM to infer the supportiveness of responses to each claim. However, none of them address the intra-sample contextual dependence, which can lead to trusting fabricated yet consistent sampled outputs.

3 IUQ: Interrogative Uncertainty Quantification

IUQ focuses on fine-grained factuality and isolates atomic knowledge from the generative context. Structurally, IUQ is composed of a responder and an interrogator, with the interrogator continuously questioning the responder for the information it has generated, as shown in Fig. 2. In practice, we use the same language model for both the responder and the interrogator to prevent systematic bias. Please refer to Appendix C for the prompts we used in IUQ.

3.1 Response Generation

Given a model M and a prompt x, we sample N diverse responses from M with temperature $T = t$. These responses comprise 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. The responses that refuse to answer are excluded (e.g. responses of "I don't know", "I cannot provide information").

3.2 Claim-Level Question-Answering

Unlike short-form outputs, the long-form generation of LLM is phrased in natural language consisting of syntax, factual information, and colloquial phrases. A common method to extract information from long text is to incorporate an LLM to decompose the generated text into the smallest possible semantic claims (Min et al., 2023; Song et al., 2024; Jiang et al., 2024b). We follow the same practice and use the model M to decompose the response R to obtain a set of atomic claims $\mathcal{C}^R$, where

$$\mathcal{C}^R = M_{T=0}(R, x) = \{c_1, c_2, \ldots, c_k\}, \quad (1)$$

and k is the number of claims returned by model M.

As discussed in section 1, we aim to examine the model's knowledge by decoupling the claims from the generative context. This is achieved by utilizing LLM to extract the information contained in a claim and phrasing it as a question. We obtain the set of questions sampled for claim $c \in \mathcal{C}^R$ as

$$\mathcal{Q}_c = M_{T=t}(c, x) = \{q_1, q_2, \ldots, q_{n_q}\}, \quad (2)$$

where n_q denotes the number of questions generated for claim c. We generate $\mathcal{Q}_c$ in a single inference request and require $n_q \leq 3$ to prevent repetitive questions and limit computation costs.

Ideally, only one question should be derived for each claim due to its semantic atomicity; however, since the decomposition of the original response could be non-exhaustive (e.g. model M could return a claim "Nobuhiro was born in 1976, in Osaka, Japan.", which is still divisible), sampling multiple questions complements the claim extraction process to support fine-grained analysis.

We then obtain the set of answers $\mathcal{A}_q$ for each question $q \in \mathcal{Q}_c$ as

$$\mathcal{A}_q = \{a_i\}_{i=1}^{n_a}, \quad \text{where } a_i \sim M_{T=t}(q, x), \quad (3)$$

and n_a is a hyperparameter specifying the number of answers to generate for each question.

3.3 Claim-Level Faithfulness

Since the answer set obtained in Eq. 3 is independent of the generative context R, it serves as a truthful representation of the model's knowledge and a testing criterion for the information claimed in R. Specifically, we acquire the semantic relationship between Eq. 3 and Eq. 1 to check if the model has fabricated information. One common approach is to incorporate the Natural Language Inference model to infer the entailment relationship (Kuhn et al., 2023; Lin et al., 2024). However, it requires an exhaustive check of pairwise combinations between elements in the answer set and the claim set. Instead, we directly prompt the model M to check if the answer set contradicts the previous claims and output the percentage of contradiction.

Given an atomic claim c_i extracted from R, we denote $C_{i\leq}$ as all claims that precede and include c_i. We are then interested in knowing whether the model truly possesses the knowledge of c_i or fabricates c_i from the generative context. This is achieved by using the model M to provide an estimate of the percentage of contradiction between $\mathcal{A}_q$ and $C_{i\leq}$, denoted as $X(\mathcal{A}_q, C_{i\leq})$. We define the faithfulness for claim $c_i \in R$ as

$$F(c_i) = 1 - \frac{1}{|\mathcal{Q}_{c_i}|} \sum_{q \in \mathcal{Q}_{c_i}} X(\mathcal{A}_q, C_{i\leq}), \quad (4)$$

where $|\mathcal{Q}_{c_i}|$ is the number of questions in $\mathcal{Q}_{c_i}$. If the claim c_i is a faithful representation of the model's knowledge, $X(\mathcal{A}_q, C_{i\leq})$ will be zero for all questions, and thus $F(c_i) = 1$. The average faithfulness for claims in R then constitutes the faithfulness of the response R

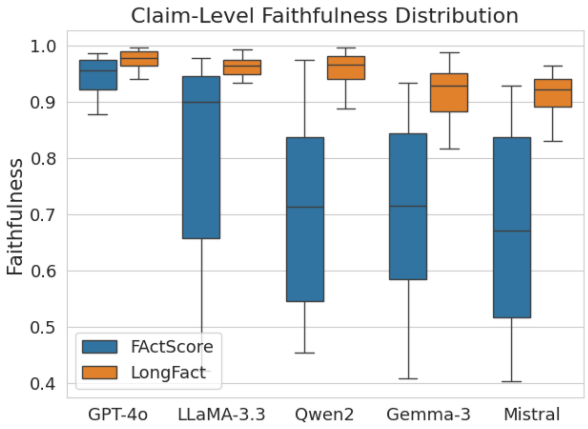
$$F(R) = \frac{1}{|\mathcal{C}^R|} \sum_{c_i \in \mathcal{C}^R} F(c_i). \quad (5)$$

Here, $F(R)$ indicates the tendency of the model to fabricate information, as lower $F(R)$ suggests the possibility that the model's response contradicts its knowledge.

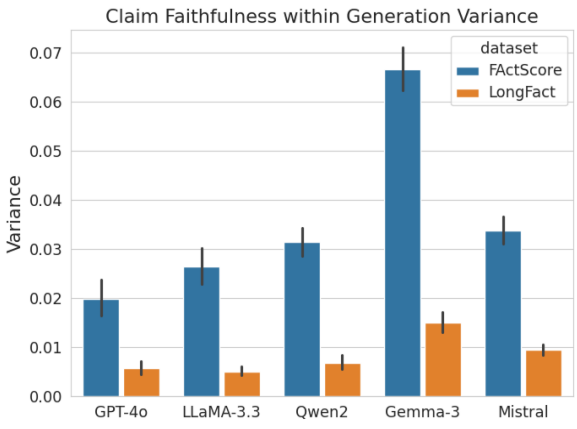
3.4 Claim-Level Uncertainty

We build uncertainty estimation on inter-sample consistency while accounting for the influence of claim faithfulness defined in Eq. 4. Following (Jiang et al., 2024b) and (Zhang et al., 2024), the inter-sample consistency is evaluated at the claim-level by aggregating the number of sampled responses that entail claim $c_i \in \mathcal{C}^R$. For each claim c_i, the consistency score is computed as $S(c_i) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}[R^k \Rightarrow c_i]$, where N is the number of sampled responses. However, $S(c_i)$ does not take into account the impact of conditioning on unfaithful information when generating new tokens.

Therefore, we propose to utilize the sequential property of $c_i \in R$ to model the influence of unfaithfulness $1 - F(c_i)$. Specifically, we convolve



(a) Distribution of all claim faithfulness by models.



(b) Variance of claim faithfulness within generated responses.

Figure 3: Statistics of the claim-level faithfulness over selected models. (a) The faithfulness scores of all claims over FActScore and LongFact. Higher scores indicate less contradiction between claims and model's knowledge. (b) The average variance of claim faithfulness within each sampled response.

the sequence $[1-F(c_1), 1-F(c_2), \ldots, 1-F(c_k)]$ with a influence kernel E to propagate the impact of unfaithful claims to subsequent claims. We then define the unfaithfulness weighting for claim c_i as

$$W(c_i) = \sum_{j=1}^{i} (1 - F(c_j)) \cdot E(i-j). \quad (6)$$

We choose the kernel E as the exponential decay function $E(i) = e^{-\lambda i}$ for claim indices $i = 0, 1, \ldots, k$. Different choices of kernels are evaluted in Table. 4. The uncertainty score for claim c_i is then S scaled by the influence of claim unfaithfulness

$$U(c_i) = S(c_i) \cdot W(c_i). \quad (7)$$

The uncertainty score U encompasses inter-sample consistency while accounting for the influence of preceding unfaithful claims. The weighting W can also be applied to other consistency-based methods to consider the intra-sample claims dependence.

3.5 Answer-Level Uncertainty

One major challenge for long-form UQ comes from the large number of generated tokens. Empirically, token-probability based approaches become less effective when the length of generation increases. Leveraging the convenience of short-answers, we reformulate token-based approaches by operating on answer sets $\mathcal{A}_q$, where the answers are treated independently as the model's response.

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:

$$H(R|x) = -\sum_i p(z_i|x) \log p(z_i|z_{<i}, x), \quad (8)$$

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

As a result, we propose an indirect approach based on token-probability of the answers in the set $\mathcal{A}_q$ without the need to tackle the original long-form response. Since the context is bound to claim c_i, their token-probabilities are indicative of the LLM's uncertainty over the claim c_i. We define the uncertainty estimate built on entropy H as

$$U_A(c_i) = \frac{1}{|\mathcal{Q}_{c_i}|} \sum_{q \in \mathcal{Q}_{c_i}} \frac{1}{|\mathcal{A}_q|} \sum_{a \in \mathcal{A}_Q} H(a|c_i). \quad (9)$$

The uncertainty quantification result of U_A is shown in Table. 2, complementing the semantic-based approach of IUQ. Methods including perplexity and maximum token entropy can be employed for additional comparison but are excluded for clarity.

4 LLM Faithfulness

In this section, we present the analysis of how faithful the model is in generating long-form responses and quantify the model's tendency to fabricate information.

We first collect all claims tested across the datasets and their corresponding faithfulness defined in Eq. 4. The analysis is performed based

on: (a) the overall distribution of claim faithfulness across all data samples, and (b) the variance of claim faithfulness within each sampled response. We make the following observation from Fig. 3: (i) The average claim faithfulness exhibits a clear distinction among different datasets, as the models are often less faithful with the topics in FActScore than in LongFact. A possible explanation is that FActScore mainly contains biographies of lesser-known individuals, while LongFact addresses popular topics in various fields. (ii) This distinction is also evident in the variance of claim faithfulness within generation, showing that the models are more likely to mix faithful claims with fabricated information.

Given the definition of response faithfulness in Eq. 5 , the model's faithfulness on the prompt x and the underlying topic is then $F(\mathcal{R}) = \frac{1}{|\mathcal{R}|}\sum_{R\in\mathcal{R}} F(R)$, where $|\mathcal{R}|$ is the number of sampled claims. We then average $F(\mathcal{R})$ over all topics in the dataset to define the faithfulness of a model as

$$F(M) = \frac{1}{|D|}\sum_{\mathcal{R}\in D} F(\mathcal{R}). \tag{10}$$

Computing $F(M)$ for selected models, we present the quantified model faithfulness in Table. 1.

Dataset	GPT	LLaMA3.3	Qwen	Gemma	Mistral
FActScore	0.927	0.816	0.700	0.697	0.679
LongFact	0.974	0.959	0.954	0.919	0.911

Table 1: Model faithfulness on FActScore and LongFact.

$F(M)$ serves as a measure of the model's tendency to fabricate information in long-form responses.

5 Experiments

In this section, we present the setup of the experiments and ablation studies to demonstrate the effectiveness of IUQ.

5.1 Baselines

We select both white-box and black-box uncertainty quantification methods as baselines, using implementations from (Fadeeva et al., 2023) for white-box methods. Specifically, we include:

- **Max Token Entropy** (Fomicheva et al., 2020): This method quantifies uncertainty by calculating the Shannon entropy of the generated tokens. For a given output sequence, it identifies the maximum entropy value across the aligned tokens of claims in the original response, where high entropy at any step indicates a lack of confidence in the token selection.

- **Perplexity (PPL)** (Fomicheva et al., 2020): Perplexity represents the geometric mean of the inverse probability of the tokens. For claim-level analysis, the extracted claims are aligned with the original response to use the corresponding tokens.

- **Claim-Conditioned Probability (CCP)** (Fadeeva et al., 2024b): This method improves upon the entropy-based method by isolating factual uncertainty from linguistic variation. It identifies the semantically important tokens and takes into account the probabilities of their alternatives. This strategy effectively leverages information encapsulated in the output without the need to perform additional sampling.

- **Frequency Scoring** (Mohri and Hashimoto, 2024): This approach samples multiple alternative responses from the model to define the associated uncertainty sets, where each set contains statements that entail the model's output. The author show how conformal prediction defines a back-off algorithm for ensuring the correctness of LM outputs, and correspondingly define a uncertainty metric at claim-level.

- **Claim Entailment** (S): This approach is adopted in (Zhang et al., 2024) and (Jiang et al., 2024b) to evaluate the uncertainty of a sequence by aggregating the number of sampled generations that entail the sequence. For claim-level analysis, the score S for claim c_i is computed as $S(c_i) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}[R^k \Rightarrow c_i]$, where N is the total number of sampled responses.

- **Closeness Centrality** (C_C) (Jiang et al., 2024b): Building on the Claim Entailment S, Closeness Centrality is a graph-based method that exploits the connectivity of nodes to estimate the likelihood for the claim to hold true. A bipartite graph is constructed by treating each claim as a node and drawing an edge between the node and sampled responses according to their entailment relationship. While multiple graph-based methods are explored (betweenness, eigenvalue, PageRank), we only compare Closeness Centrality (C_C), which is the best-performing metric.

	Metric	GPT-4o	LLaMA-3.1	LLaMA-3.3	Qwen2	Gemma-3	Mistral	Avg.
FActScore	Max Token Ent.	-	0.596	0.672	0.637	0.625	0.659	0.638
	PPL	-	0.577	0.661	0.622	0.593	0.647	0.620
	CCP	-	0.623	0.663	0.683	0.627	0.659	0.651
	Freq. Scoring	-	0.751	0.724	0.763	0.579	0.747	0.713
	U_A	0.617	0.634	0.633	0.838	0.706	0.799	0.705
	S	0.732	0.819	0.847	0.901	0.820	0.880	0.833
	C_C	**0.749** (-0.1%)	0.822	0.843	0.929	0.840	0.862	0.841
	IUQ (Ours)	0.748	**0.847** (+2.5%)	**0.875** (+2.8%)	**0.932** (+0.3%)	**0.867** (+2.7%)	**0.913** (+3.3%)	**0.864** (+2.3%)
LongFact	Max Token Ent.	-	0.552	0.521	0.558	0.528	0.554	0.543
	PPL	-	0.569	0.518	0.577	0.524	0.572	0.552
	CCP	-	0.559	0.520	0.508	0.537	0.539	0.533
	Freq. Scoring	-	0.643	0.666	0.699	0.559	0.696	0.653
	U_A	0.592	0.573	0.591	0.659	0.557	0.625	0.600
	S	0.705	0.736	0.714	0.791	0.656	0.733	0.723
	C_C	0.722	0.724	0.702	0.782	0.639	0.712	0.714
	IUQ (Ours)	**0.733** (+1.1%)	**0.749** (+1.3%)	**0.722** (+0.8%)	**0.806** (+1.5%)	**0.689** (+3.3%)	**0.743** (+1.0%)	**0.740** (+1.7%)

Table 2: AUROCs of the uncertainty quantification metrics across models of diverse sizes. Bold-text indicates the highest scores, and italic-text indicates the second highest scores. The White-box uncertainty quantification results for GPT-4o is unavailable due to its closed-source nature.

5.2 Datasets and Annotation

Dataset	Responses	Claims	Questions	Answers
FActScore	235	4759	10433	31299
LongFact	250	4276	9954	29862
Total	**485**	**9035**	**20387**	**61161**

Table 3: Statistics of the total numbers of generated items by GPT-4o on FActScore and LongFact.

We evaluate IUQ on FActScore (Min et al., 2023) and LongFact(Wei et al., 2024). We select entities from each dataset, using the provided prompt as input and the reference text to check for claim-level correctness. A statistics of the data composition is shown in Table. 3. The processing of each dataset is as follows:

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.

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 using search engine to fetch and evaluate internet-based sources. To maintain consistency and reproducibility, we manually select 50 entities of diverse topics in LongFact that include dedicated Wikipedia articles, and employ the same method we used for FActScore to evaluate the factuality of claims.

5.3 Models and Parameters

We conduct experiments over models across various model families, including GPT4o (OpenAI et al., 2024), LLaMA-3.1-70B-Instruct, LLaMA-3.3-70B-Instruct (Touvron et al., 2023), Qwen2-VL-72B (Yang et al., 2024), Gemma-3-27b-it (Team et al., 2025), and Mistral-Small-24B-Instruct (Jiang et al., 2023). For each data entity, we sample 5 long-form responses using temperate $t = 1.0$, and use temperate $t = 0$ to evaluate the correctness of claims.

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.

Method	FActScore					LongFact				
	GPT-4o	LLaMA-3.3	Qwen2	Gemma-3	Avg.	GPT-4o	LLaMA-3.3	Qwen2	Gemma-3	Avg.
Lin-E	0.732	0.858	0.917	0.834	0.835	0.725	0.714	0.801	0.682	0.731
Acc-E	0.713	0.841	0.889	0.804	0.812	0.723	0.710	0.800	0.675	0.727
No-E	**0.748**	0.871	0.931	0.847	0.849	0.724	**0.722**	**0.806**	0.678	0.733
Exp-E (IUQ)	**0.748**	**0.875**	**0.932**	**0.867**	**0.856**	**0.733**	**0.722**	**0.806**	**0.689**	**0.738**

Table 4: Ablation study on the impact of claim consistency score with different error propagation (E) function. The presented values are AUROCs of the uncertainty quantification metric U_S.

5.5 Ablation Study

In this section, we present an experimental study to show the effectiveness of our claim-consistency paradigm (Section 3.3). Firstly, we illustrate that IUQ captures 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 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. 4 and Fig. 4. Additional experiments are reported in Appendix B.

Effectiveness of Claim Consistency Score The claim faithfulness 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.

The result illustrates that although S is not a particualrly strong baseline, IUQ shows superior performance over all tested models. This observation consolidates our motivation that LLM has limitations in identifying its own lack of knowledge. Without sampling multiple responses and performing fine-grained analysis, it is risky to trust LLM responses, especially in long-form generation.

Effectiveness of Influence Kernel The influence kernel E serves to propagate the impact of an inconsistent claim to subsequent claims. In this section, we investigate the influence of different kernels as propagation functions on the uncertainty estimation performance. The results are shown in Table. 4 and the notations used are explained as follows: (1) No-E: No error is propagated to subsequent claims, and we build the uncertainty estimate solely on the claim consistency score. (2) Lin-E: Linear error propagation, where the unfaithfulness scores 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-E: accumulative error propagation, where the cumulative sums of the claim-level unfaithfulness are used as the weighting to the entailment score S.

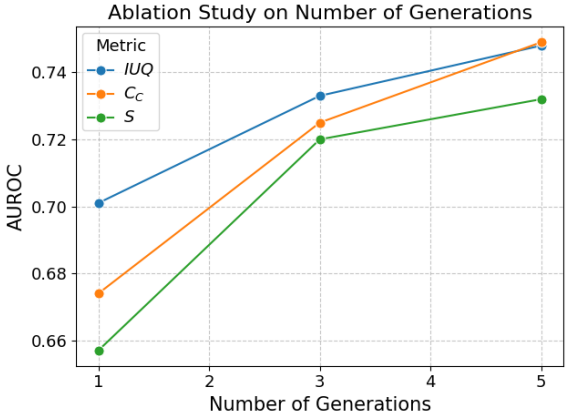


Figure 4: AUROCs of uncertainty metric U_S and baselines on different numbers of sampled responses.

Influence of Number of Generations We show the influence of the number of sampled responses on our uncertainty metric IUQ and black-box baseline methods in Fig. 4. The experiments are performed using GPT-4o. The result demonstrates that the number of sampled responses has a non-negligible effect in quantifying the uncertainty, as more samples leads to more accurate claim-entailment scrores S.

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 possible hallucination introduced in the workflow, and there is no guarantee that such hallucination will be detected. This problem is partially addressed by adapting the source code to incorporate the model provider's support of structured output, which is limited to a few latest models. Additional measures we take are to manually parse the model's output and perform heuristic sanity checks to ensure model responses are at least minimally sensible.

8 Potential Risks

This work does not involve personally identifiable data or sensitive information. All experiments are conducted using publicly available data and therefore raise no direct ethical or privacy concerns. We have carefully adhered to the ACL Guidelines of Ethics throughout the research and writing process.

9 AI Assistance Statement

The use of AI tools were solely to assist with the linguistic polishing of this manuscript, such as improving grammar, clarity, and readability. All conceptual contributions, technical methods, experimental designs, and analyses were developed entirely by the authors without the use of LLMs.

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*.

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. 2024a. Fact-checking the output of large language models via token-level uncertainty quantification. *Preprint*, arXiv:2403.04696.

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. 2024b. Fact-checking the output of large language models via token-level uncertainty quantification. *Preprint*, arXiv:2403.04696.

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. 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.

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.

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*.

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.

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.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. *Preprint*, arXiv:2402.10978.

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.

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. 5.

B Additional Experiments

Claim Consistency Landscapes The claim consistency score computed in Eq. 4 encapsulates the faithfulness of the claim and consistency of inter-sample generations. 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 5: Example LongFact prompts and corresponding Wikipedia entries.

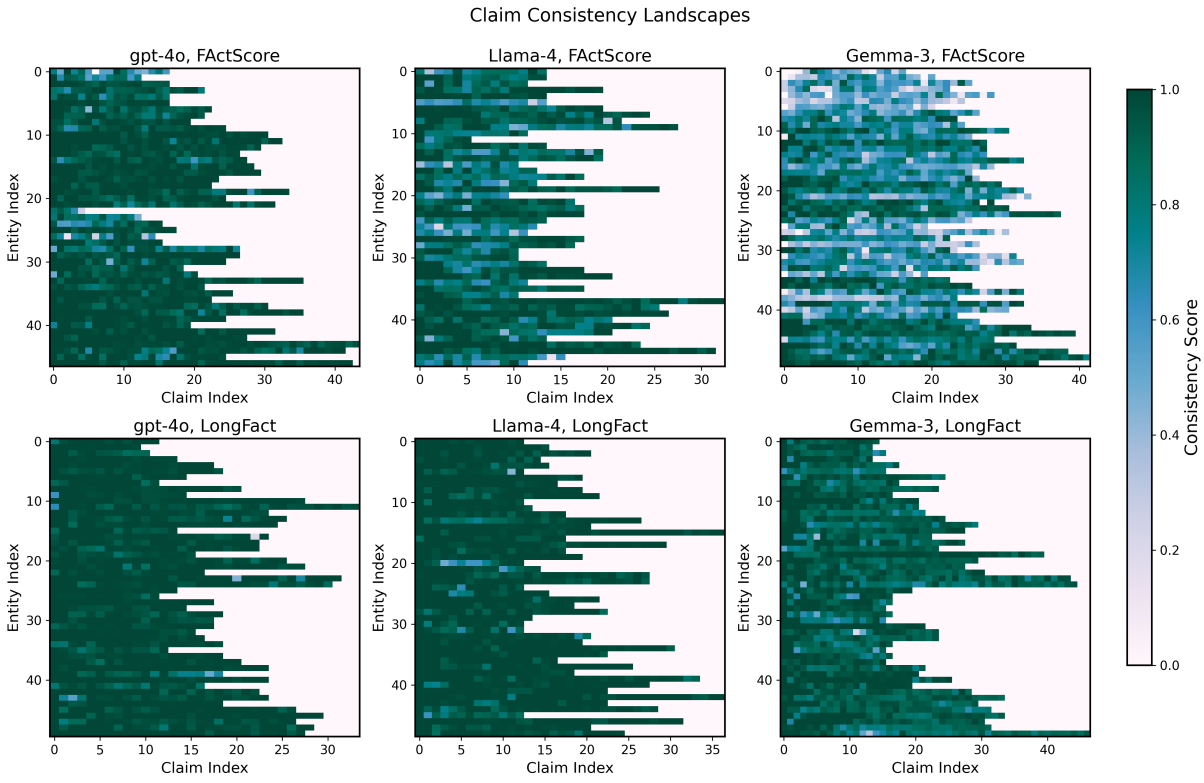


Figure 5: Claim-consistency scores within individual generations. The x-axis is the index of the claim made in LLM's response, and y-axis is the index of the topic in datasets. Results for FActScore and LongFact are shown with selected models.

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..