Quantifying Uncertainty in Natural Language Explanations of Large Language Models

Sree Harsha Tanneru
Harvard University

Chirag Agarwal
Harvard University

Himabindu Lakkaraju
Harvard University

Abstract

Large Language Models (LLMs) are increasingly used as powerful tools for several high-stakes natural language processing (NLP) applications. Recent works on prompting language models claim to elicit intermediate reasoning steps and key tokens that serve as proxy explanations for LLM predictions. However, there is no certainty whether these explanations are reliable and reflect the LLM's behavior. In this work, we make one of the first attempts at quantifying the uncertainty in explanations of LLMs. To this end, we propose two novel metrics — *Verbalized Uncertainty* and *Probing Uncertainty* — to quantify the uncertainty of generated explanations. While verbalized uncertainty involves prompting the LLM to express its confidence in its explanations, probing uncertainty leverages perturbations as means to quantify the uncertainty. Our empirical analysis of benchmark datasets reveals that verbalized uncertainty is not a reliable estimate of explanation confidence. Further, we show that the probing uncertainty estimates are correlated with the faithfulness of an explanation, with lower uncertainty corresponding to explanations with higher faithfulness. Our study provides insights into the challenges and opportunities of quantifying uncertainty in LLM explanations, contributing to the broader discussion of the trustworthiness of foundation models.

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Large Language Models (LLMs), such as GPT4 (OpenAI, 2023), Bard (Manyika, 2023), Llama-2 (Touvron, 2023), and Claude-2 (Anthropic, 2023), have garnered significant attention and are employed across a wide range of applications, including chat-bots, computational biology, creative work, and law (Kaddour et al., 2023) due to their impressive natural language understanding and generation capabilities. However, state-of-the-art LLMs are complex models with billions of parameters, where their inner working mechanisms are not fully understood yet, making them less trustworthy amongst relevant stakeholders. This lack of transparency causes hindrance to deploying LLMs in high-stakes decision-making applications, where the consequences of incorrect decisions are severe and could result in the generation of harmful content, misdiagnosis (Zhang et al., 2023), and hallucinations (Ji et al., 2023; Weidinger et al., 2021). The lack of user trust demands the development of robust explanation techniques to gain insights into how these powerful LLMs work.

Previous works for explaining language models can be broadly categorized into perturbation-based methods (Li et al., 2016a,b), gradient-based methods (Kindermans et al., 2017; Sundararajan et al., 2017)), attention-based methods (DeRose et al., 2020; Vig, 2019), example-based methods (Jin et al., 2020; Treviso et al., 2023; Wang et al., 2022; Wu et al., 2021), and Natural Language Explanations (NLEs) (Wei et al., 2023). While most of the above methods require white-box access to models (*e.g.,* model gradients and prediction logits), NLEs can be generated by LLMs in a self-explanatory manner, thereby enabling us to understand the behavior of these models even when the models are closed-source. For instance, Chain-of-Thought (CoT) (Wei et al., 2023) explanations, a popular class of NLEs generated by LLMs show the step-by-step reasoning process leading to the outputs generated by these models. While CoTs and other natural language explanations generated by LLMs often seem quite plausible and believable Turpin

et al. (2023), recent works have demonstrated that these natural language explanations may not always faithfully capture the underlying behavior of these models Turpin et al. (2023). However, there is little to no work that focuses on deciphering if and to what extent the generated NLEs are trustworthy. One way to address this problem is to quantify the uncertainty in the NLEs generated by LLMs. However, this critical direction remains unexplored.

Prior works on uncertainty estimation in the context of LLMs have only focused on providing uncertainty estimates (*i.e.,* confidence) corresponding to the responses (answers) generated by LLMs (Xiong et al., 2023). While uncertainty in LLM predictions has been studied using external calibrators (Jiang et al., 2021), model fine-tuning (Lin et al., 2022), and non-logit-based approaches (Xiong et al., 2023), there is little to no work on estimating the uncertainty of LLM explanations. Understanding the uncertainty in natural language explanations generated by LLMs is paramount to ensuring that these explanations are trustworthy and are not just plausible hallucinations.

Present work. In this work, we make an attempt at quantifying the uncertainty in natural language explanations generated by LLMs. In particular, we propose two novel approaches – *Verbalized uncertainty* and *Probing uncertainty* metrics – to quantify the confidence of NLEs generated by large language models and compare their reliability. While verbalized uncertainty metrics focus on prompting a language model to express its uncertainty in the generated explanations, probing uncertainty metrics leverage different kinds of input perturbations (e.g., replacing words with synonyms, paraphrasing inputs) and measure the consistency of the resulting explanations. Using our proposed metrics, we provide the first definition of uncertainty estimation of language model explanations. In addition, our work also demonstrates key connections between *uncertainty* and *faithfulness* of natural language explanations generated by LLMs.

We evaluate the effectiveness of our proposed metrics on three math word problems and two commonsense reasoning benchmark datasets and conduct experiments using different GPT variants. Our empirical results across these datasets and LLMs reveal the following key findings. 1) Verbalized uncertainty is not a reliable estimate of explanation confidence and LLMs often exhibit very high verbalized confidence in the explanations they generate. 2) Probing uncertainty is correlated with the predictive performance of the LLM, where correct answers from a model tend to generate more confident/less uncertain explanations. 3) A clear connection exists between the uncertainty and faithfulness of an explanation, where less uncertain explanations tend to be more faithful to the model predictions. While our study primarily focuses on self-explanations generated by LLMs, the approaches and metrics outlined can be easily extended to Natural Language Explanations (NLEs) generated by surrogate models too.

2 RELATED WORKS

Our work lies at the intersection of large language models, explainability, and uncertainty estimation, which we discuss below.

Large Language Models. In the field of natural language processing (NLP), Large Language Models (LLMs) have proven their efficacy in various tasks, including sentiment analysis, text summarization, and machine translation. Moreover, they have become the backbone of modern conversational agents and virtual assistants, powering chatbots like GPT-3 Brown et al. (2020), GPT-4 Bubeck et al. (2023), Llama-2 Touvron et al. (2023), and Claude Anthropic (2023) that offer human-like interactions. Kaddour et al. (2023) provide a survey of the applications of LLMs in Chatbots, Computational Biology, Creative Work, Knowledge Work, Law, Medicine, Reasoning, Robotics, Social Sciences, and Synthetic Data Generation. While the adaptability of LLMs to a wide array of applications underscores their potential to reshape industries through enhanced natural language understanding, there exists a gap in trusting these models and deploying them to real-world users.

Explainability. Explainability in machine learning has gained significant attention as the deployment of complex models has become more pervasive in critical applications. The need for understanding model decisions and gaining insights into their inner workings has led to the development of various explainability techniques. While there have been a plethora of perturbation-based, gradient-based, example-based, and surrogate-based methods, none of these explanation methods are feasible for LLMs as they require some level of model access, *e.g.,* gradients or logits. Chain-of-Thought (CoT) explanations (Wei et al., 2023), a form of natural language explanation generated by prompting LLMs, are the state-of-the-art alternative for models without open-source access and has shown to provide plausible and human-interpretable reasoning behind LLM predictions. However, several questions have been raised about the reliability of these explanations.

Uncertainty Estimation. Traditional ways to measure the confidence of predictions primarily rely on model logits, have become less suitable for LLMs and even infeasible with the development of closed-source

LLMs. While Xiong et al. (2023) proposes approaches for confidence elicitation of black-box LLMs, to the best of our knowledge, our work is the first to tackle the problem of estimating the uncertainty in NLEs.

3 PRELIMINARIES

Notations. Large language models typically have a single vocabulary $\mathcal{V}$ that represents a set of unique "tokens" (words or sub-words). Let $\mathcal{M} : Q \rightarrow A$ denote a language model mapping a sequence of n question tokens $Q = (q_1, q_2, \ldots, q_n)$ to sequence of m answer tokens $A = (a_1, a_2, \ldots, a_m)$, where q_i and a_i are text tokens in $\mathcal{V}$. In addition to the original question Q, we design specific prompts Q_e to generate natural language explanation (NLE) A_e from the language model $\mathcal{M}$.

Uncertainty. Black-box LLMs do not provide access to parameter gradients or model logits, rendering traditional explainability techniques ineffective. To this end, most language models leverage NLEs, which are explanations generated from the language model to serve as proxy explanations and are a viable alternative. While NLEs are essentially a sequence of tokens sampled from the model that serve as explanations, there is an associated uncertainty for the generated explanations. Quantifying the uncertainty of these explanations is essential to estimate the reliability of generated NLEs. For the rest of the paper, we will use the term "confidence score" to refer to the uncertainty of an explanation, as determined by the language model.

Explanation Methods. We confine our study to two explanation methods — Token Importance and Chain of Thought (CoT) explanations. While token importance explanations Li et al. (2016a); Wu et al. (2020) aims to identify input tokens (refer to tokens t in an input text T for LLMs) that most contribute to a model's predictions, CoT explanations (Wei et al., 2023) focus on revealing the sequence of operations or reasoning steps $S_i \in S$ the language model $\mathcal{M}$ takes when processing the question Q and arriving at its predictions, where $n_s = |S|$ denotes the total number of steps in a CoT explanation. For token importance explanation, we concatenate a prompt Q_e to the given question Q using the template: "*Read the question and output the words important for your final answer...*". Whereas, the prompt Q_e to generate CoT explanations uses the following template: "*Read the question, give your answer by analyzing step by step, ...*". Please refer to Figs. 15-16 in appendix for more details.

We generate an answer from the LLM $\mathcal{M}$ as follows: $\mathcal{M}(Q) = A$. We also generate an explanation A_e along with answer A using the aforementioned template question Q_e as: $\mathcal{M}(Q_e + Q) = A + A_e$.

4 QUANTIFYING UNCERTAINTY IN EXPLANATIONS

Next, we describe our metrics which aim to estimate the uncertainty in token importance and CoT explanations generated by LLMs.

Problem formulation (Uncertainty in Explanations). *Given a question-answer pair (Q, A) and prompt Q_e to generate natural language explanation A_e from the model $\mathcal{M} : (Q, Q_e) \rightarrow (A, A_e)$, we aim to develop an uncertainty function* $\textsc{Unc} : A_e \rightarrow [0, 1]$, *which maps a generated explanation A_e to a scalar score that determines the uncertainty in the generated explanation, i.e.,*

$$\text{Uncertainty} = \textsc{Unc}(A_e),$$

where $\mathcal{M}(Q_e + Q) = A + A_e$.

As mentioned before, we confine our study to two natural language explanation methods – Token Importance and CoT. We use $\text{TI}_q : \{w \mid w \in Q\}$ to denote a token importance explanation which is a subset of words in the question Q that are important for predicting the answer A and $\text{CoT}_q : \{(S_1, c_1) \rightarrow (S_2, c_2) \cdots \rightarrow (S_{n_s}, c_{n_s})\}$ to denote a CoT explanation for a prediction A from question Q. Here $S_i = (s_1, s_2, \ldots s_{n_s})$ is a text sequence denoting the natural language reasoning and $c_i \in [0, 1]$ is the verbalized confidence of CoT step S_i from the model $\mathcal{M}$.

4.1 Verbalized Uncertainty

A straightforward approach to elicit uncertainty of an NLE is to directly request the LLM $\mathcal{M}$ to output a confidence score for the explanation ranging from 0% to 100%. By directly soliciting the model's self-assessment of uncertainty, this approach seeks to extract explicit uncertainty information inherent in the model. We provide the template of the prompts for confidence elicitation for token importance and CoT explanations in Figs. 1-2. For token importance, we ask the underlying LLM to verbally assign an importance score to each word in the question Q and then provide the final answer of the question with an overall confidence in importance scores (see Fig. 1). In contrast, for CoT explanations (see Fig. 2), we ask the LLM $\mathcal{M}$ to assign verbalized confidence to each step in the CoT reasoning and the final answer.

4.2 Probing Uncertainty

Verbalized uncertainty elicits confidence in an explanation by directly requesting the underlying LLM to output a confidence score in a given range. In contrast, for estimating uncertainty using probing, we leverage the

Read the question, and assign each word an importance score between 0 and 100 of how important it is for your answer. The output format is as follows:

Word: [Word 1 here], **Importance**: [Your importance score here]
...
Word: [Word N here], **Importance**: [Your importance score here]
Final answer and overall confidence (0-100): [Your answer as a number here], [Your confidence here]

Note: The importance scores of all words should add up to 100. The overall confidence score indicates the degree of certainty you have about your importance scores. For instance, if your confidence level is 80%, it means you are 80% certain that importance scores assigned are correct. Provide the answer in aforementioned format, and nothing else.

Q: Jake has 11 fewer peaches than Steven. If Jake has 17 peaches. How many peaches does Steven have?

Answer:
Word: Jake, **Importance:** 20%
Word: Steven, **Importance:** 20%
Word: peaches, **Importance:** 60%
Final answer and overall confidence (0-100): 28, 100%

Figure 1: **Template for generating token importance and its confidence.** The prompt Q_e appended to the original question Q to elicit a token importance explanation TI. We ask the underlying LLM to verbally assign an importance score to each word in the question Q and then provide the final answer A with overall confidence.

consistency of explanations as a measure to estimate the uncertainty in explanations generated by a language model $\mathcal{M}$. More specifically, let A_e denote the natural language explanation generated by the model $\mathcal{M}$ for a given question Q and $[A_{e_1}, A_{e_2}, \ldots, A_{e_N}]$ be N explanations generated for N perturbation of the same question using its local neighborhood. Next, we describe two different perturbation strategies to generate N explanations for a given question and answer.

Sample Probing. Motivated by the local neighborhood approximation works in XAI (Ribeiro et al., 2016; Smilkov et al., 2017), we propose uncertainty metrics that leverage the consistency of a model in generating the explanation in a local neighborhood. Here, we presume that the local behavior of the underlying LLM is consistent for perturbed samples of the original question and gradually introduce perturbations in the questions by *paraphrasing* the original question Q. Given a question Q, we paraphrase the question into N different forms $\{Q_1, Q_2, \ldots, Q_N\}$, such that each paraphrased question Q_i is semantically equivalent to Q, and the true reasoning process remains the same, *i.e.,* given a question: "*Jake has 11 fewer peaches than Steven. If Jake has 17 peaches. How many peaches does Steven have?*", some of its local paraphrased counterparts used to calculate uncertainty in explanations are i)*...What is the number of peaches Steven has?* ii)*...How many peaches is Steven in possession of?* iii)*...How many peaches does Steven possess?* Next, we generate the explanations using the LLM by probing the model using the paraphrased questions Q_i. Mathematically,

$$\mathcal{M}(Q_e + Q_i) = A_i + A_{e_i} \; ; \; i = 1, 2, \ldots, N \qquad (1)$$

where Q_i is a paraphrased form of question Q, Q_e is the prompt to generate explanations, and A_{e_i} is the corresponding generated explanation.

Model Probing. In contrast to sample uncertainty, where we quantify the uncertainty in explanations using the variance in the input questions, model uncertainty addresses the uncertainty of LLM explanations due to the inherent stochasticity of the underlying language model $\mathcal{M}$.

More specifically, we use the "temperature" parameter τ present in most LLMs that control the randomness in the generated answers by using the probability distribution of each generated token. A high value of temperature indicates an even distribution among all tokens, and a lower value of temperature indicates a sharper distribution (see Fig. 3). As the temperature parameter increases, the language model becomes more creative and stochastic in the generated explanations. Note that close-sourced models like OpenAI (2023) and Anthropic (2023) expose temperature scaling as an input parameter and recommend a default value of τ=1. Intuitively, the temperature parameter affects the sampling process when generating answers from the model. For a given question Q (say "*my favorite food is ...*"), we sample N answers and their corresponding explanations, $\{A_i, A_{e_i}\} \forall i \in 1, 2, \ldots, N$

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], **Confidence**: [Your confidence here]%
Step 2: [Your reasoning here], **Confidence**: [Your confidence here]%
...
Step N: [Your reasoning here], **Confidence**: [Your confidence here]%
Final answer and overall confidence (0-100): [Your answer as a number here], [Your confidence here]%
Note: The confidence indicates the degree of certainty you have about your reasoning. For instance, if your confidence level is 80%, it means you are 80% certain that your reasoning is correct. Provide the answer in aforementioned format, and nothing else.

Q: Jake has 11 fewer peaches than Steven. If Jake has 17 peaches. How many peaches does Steven have?

Answer:
Step 1: Jake has 11 fewer peaches than Steven. Confidence: 100%
Step 2: Jake has 17 peaches. Confidence: 100%
Step 3: If Jake has 17 peaches, then Steven has 17 + 11 = 28 peaches. Confidence: 100%
Final answer and overall confidence (0-100): 28, 100%

Figure 2: **Template for generating CoT explanation and its step-wise confidence.** The prompt Q_e appended to the original question Q to elicit a CoT explanation. We ask the underlying LLM to verbally assign an importance score to each step of the CoT explanation and then provide the final answer A with overall confidence.

from the language model $\mathcal{M}$. Mathematically, we can denote this using:

$$\mathcal{M}(Q_e + Q) = A_i + A_{e_i} \; ; \; i \in \{1, 2, \ldots, N\} \qquad (2)$$

where A_i is the i^{th} answer generated by the LLM for a given temperature τ and A_{e_i} is its respective explanation.

Figure 3: The impact of the temperature τ on model stochasticity. We find that as τ increases, the stochasticity in model responses increases. τ=0 gives near-deterministic answers to a question, whereas τ=1 gives a distribution of answers.

4.2.1 Token Importance Uncertainty

Using the above sample and model perturbation strategies, N perturbed natural language explanations A_{e_i} are generated for a given question Q, answer A, original explanation A_e. Next, we describe the metrics for estimating explanation confidence from these perturbed explanations.

We define the uncertainty in token importance explanations as the mean agreement between perturbed explanations and the original explanation. Two token importance explanations are said to agree with each other if they employ the same set of important words to arrive at a prediction. To quantify this agreement, we use token rank agreement (TR) as defined in Agarwal et al. (2023). Token rank agreement measures the fraction of important tokens that have the same position in their respective rank orders. The token rank agreement (TR) metric is defined below:

$$\text{TR}(\text{TI}_i, \text{TI}_j, k) = \frac{1}{k}\Big(\bigcup_{s \in S} \{s | s \in \text{Tokens}(\text{TI}_i, k) \wedge s \in \text{Tokens}(\text{TI}_j, k) \wedge \text{R}(\text{TI}_i, s) = \text{R}(\text{TI}_j, s)\}\Big), \qquad (3)$$

where TI_i and TI_j are any two given token importance explanations, $\text{Tokens}(\text{TI}_i, k)$ is the first k tokens in explanation TI_i, k denotes the top-K tokens a user wants as explanations, and $\text{R}(\cdot)$ function gives the rank of the word s in a token importance explanation TI. The uncertainty in token importance explanation is defined as the mean token rank agreement between the perturbed explanations TI_{e_i} and the original

explanation $\text{TI}_{\text{original}}$.

$$\text{UNC}_{\text{TI}} = \frac{1}{N}\sum_{i=1}^{N} \text{TR}(\text{TI}_{e_i}, \text{TI}_{\text{original}}, k), \tag{4}$$

4.2.2 Chain of Thought Uncertainty

While the agreement between token importance explanations is intuitive, the agreement between the chain of thought explanations is non-trivial as each CoT explanation has a sequence of steps S_i in natural language. To check if the two steps in CoT explanations are equivalent, we propose using pre-trained sentence encoder models (Reimers and Gurevych, 2019). Let us consider two CoT explanations that generate N_a and N_b steps in their respective explanations, *i.e.,* ($\text{CoT}_a = (s_{a_1}, s_{a_2}, \ldots, s_{a_{N_a}})$ and $\text{CoT}_b = (s_{b_1}, s_{b_2}, \ldots, s_{b_{N_b}})$. We define CoT agreement metric (CoTA) that measures the agreement between any two given CoT explanations as:

$$\text{CoTA}(\text{CoT}_a, \text{CoT}_b) = \frac{1}{N_a + N_b}\Big(\sum_{i=1}^{N_a} \max_{j \in 1,\ldots,N_b} \text{E}(s_{a_i}, s_{b_j}) + \sum_{j=1}^{N_b} \max_{i \in 1,\ldots,N_a} \text{E}(s_{a_i}, s_{b_j})\Big), \tag{5}$$

The intuition behind the above metric is that for every step in the first CoT explanation, we check if there exists a step in second CoT explanation which agrees with it. $\text{E}(\cdot,\cdot)$ denotes the entailment function which measures the agreement between two steps. Formally, the entailment score between two explanation steps is defined as:

$$\text{E}(s_i, s_j) = \begin{cases} 1 & \text{if statements entail each other} \\ 0 & \text{if statements do not entail each other} \end{cases}$$

Whether two statements entail each other or not is measured using pre-trained models on the Natural Language Inference (NLI) task. In NLI, given a premise (P) and a hypothesis (H), the goal is to classify a given text pair into one of the three categories: "entailment", "contradiction", or "neutral". We use a pre-trained DeBERTa He et al. (2021) model fine-tuned for the NLI task to calculate the entailment score in our experiments. We chose a binary entailment score to avoid dependence on the entailment model's confidence calibration, ensuring a consistent estimate of explanation confidence across different entailment models.

Finally, the uncertainty in the CoT explanation is calculated as the mean agreement of the perturbed chain of thought explanations with the original explanation.

$$\text{UNC}_{\text{CoT}} = \frac{1}{N}\sum_{i=1}^{N} \text{CoTA}(\text{CoT}_i, \text{CoT}_{\text{original}}) \tag{6}$$

To summarize, we introduce a metric for calculating the agreement between two CoT explanations (Eq. 5). In addition, we generate N perturbed explanations for a question, and calculate the mean agreement of perturbed explanations with the original explanation to estimate explanation uncertainty (Eq. 6).

5 EXPERIMENTS

Next, we validate the effectiveness of our proposed uncertainty metric which amounts to asking: *What is the uncertainty in explanations generated by state-of-the-art LLMs with respect to different explanation methods?* More specifically, we focus on the following research questions: RQ1) Does verbalized uncertainty estimation depict overconfidence in LLMs? RQ2) Is there a relation between uncertainty and faithfulness of an explanation? RQ3) How does explanation confidence vary for correct and incorrect answers? RQ4) How do changes in the metric parameters influence explanation uncertainty?

5.1 Datasets and Experimental Setup

We first describe the datasets and large language models used to study the uncertainty in explanations and then outline the experimental setup.

Datasets. We conduct experiments using three math word problem and two commonsense reasoning benchmark datasets. i) the **GSM8K** dataset that comprises several math word problems (Cobbe et al., 2021), ii) the **SVAMP** dataset contains math word problems with varying structures (Patel et al., 2021), iii) the **ASDiv** dataset consisting of diverse math word problems (Miao et al., 2021), iv) the **StrategyQA** (Geva et al., 2021) requires a language model to deduce a multi-step reasoning strategy to answer questions and v) the **Sports Understanding** dataset, which is a specialized evaluation set from the BIG-bench (Srivastava et al., 2022) that involves determining whether a sentence relating to sports is plausible or implausible.

Large language models. We generate and evaluate the uncertainty in explanations by generating explanations using three large language models — InstructGPT, GPT-3.5, and GPT-4.

Performance metrics.

While accuracy serves as a primary performance metric for evaluating a model's predictions, the evaluation

of explanations involves metrics like faithfulness, simplicity and human comprehension. In this work, we look at the correlation between uncertainty and faithfulness of explanations. An explanation is considered faithful if it accurately represents the reasoning of the underlying model Jacovi and Goldberg (2020). Some recent works (Atanasova et al., 2023; Lanham et al., 2023; Lyu et al., 2023) have explored defining faithfulness for natural language explanations.

(i) *Faithfulness of token importance explanations:* We use the counterfactual test (Atanasova et al., 2023) for NLEs by intervening on input tokens and checking whether the explanation reflects these tokens. Specifically, we replace identified importance tokens in the explanation with synonyms and check whether the new explanation reflects these changes. Faithfulness is then quantified by the rank agreement (Eq. 3) between the new explanations and the expected explanation with intervened tokens.

(ii) *Faithfulness of chain of thought explanations:* Recent works that explored the topic of faithfulness in CoT explanations don't explicitly quantify the faithfulness of an individual explanation. Hence, we follow suit and follow Lanham et al. (2023) to measure faithfulness at a dataset level. In our experiments, we use a strategy called "Early Answering" proposed by Lanham et al. (2023) to measure the faithfulness of CoT explanations. "Early Answering" strategy involves truncating the previously collected reasoning samples and prompting the model to answer the question with the partial CoT rather than the complete one, *i.e.,* for a question Q and CoT $[s_1, s_2, \ldots s_n]$, the model is prompted to answer with $Q + s_1$, $Q + s_1 + s_2$, until, $Q + s_1 + s_2 \cdots + s_n$. After collecting answers with each truncation of the CoT, we measure how often the model comes to the same conclusion as it did with the complete CoT. If the amount of matching overall is low, this indicates that less of the reasoning is post-hoc and subsequently more faithful. The intuition behind being that if the reasoning is not post-hoc, there are fewer ways for an explanation to be unfaithful than there are for reasoning which is post-hoc (Lanham et al., 2023).

Implementation details. To run the paraphrase probing uncertainty, we formulate 10 semantically equivalent paraphrases of every question to measure uncertainty using sample probing. In the model probing uncertainty experiment, we sample five natural language explanations at a temperature of 1.0. To compute the rank agreement of token importance explanations, we use the top-3 words *i.e.,* $k = 3$. We run on a randomly sampled subset of 100 samples for each dataset. See the Appendix A.5 for more implementation details.

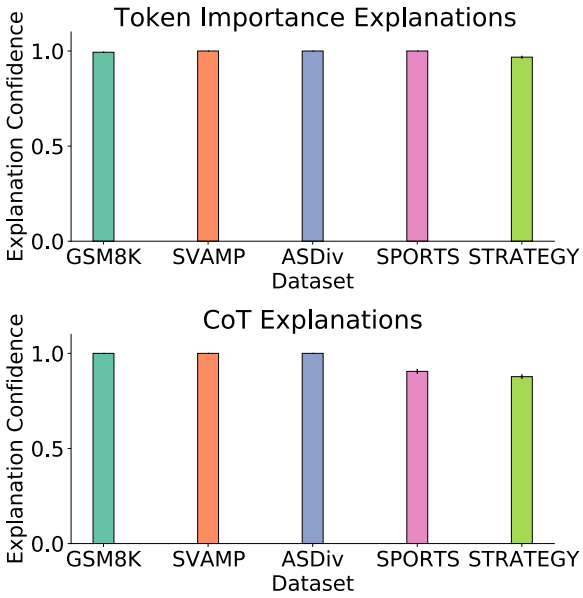

Figure 4: Verbalized explanation confidence of Token Importance and CoT explanations on three math word problems and two commonsense reasoning datasets. We observe that the verbalized explanation confidence is mostly high for explanations across all five datasets.

5.2 Results

Next, we discuss experimental results to answer questions (RQ1-RQ4) about uncertainty in explanations.

RQ1) Analyzing verbalized uncertainty. Verbalized confidence scores of both natural language explanation methods are almost always 100%. It raises questions about whether these uncertainty estimates are reliable. If the confidence in every explanation is the same, it is impossible to know when to trust the generated explanation and when not to. Our results in Fig. 4 show that, on average, across both explanation methods and five datasets, the verbalized confidence is 94.46%. Our analysis of these methods uncovers that LLMs often exhibit a high degree of overconfidence when verbalizing their uncertainty in explanations. The verbalized uncertainty for commonsense reasoning datasets is lower than math word problem datasets but still very close to 100% with little standard deviation.

RQ2) Less uncertain explanations are more faithful. A model's explanation is said to be faithful if it reflects the true reasoning behind the prediction. For token importance explanations, we swap important words in explanations with synonyms and check if the corresponding replacements are reflected in the new explanation. In Fig. 7, we demonstrate that explanation confidence is correlated with faithfulness, and highly confident (certain) explanations are more

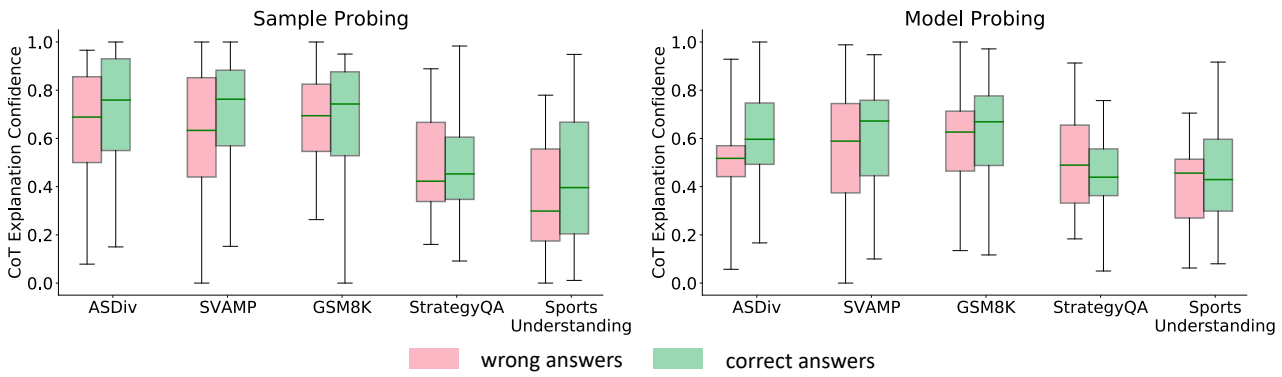

Figure 5: Chain of thought explanation confidence distributions on three math word problems and two commonsense reasoning datasets using GPT-3.5. On average, across two probing strategies and five datasets, correct answers (in green) obtain higher explanation confidence than wrong answers (in red). See Table 1 in appendix for t-test statistics comparing explanation confidence scores of correct and incorrect answers to different datasets.

faithful. In addition, we find a similar trend between the CoT explanation confidence and its faithfulness (see Fig. 6) and find that increased mean explanation confidence lead to an increase in the faithfulness of an explanation for most datasets. Our observations suggest that uncertainty estimation can be used as a test for the faithfulness of NLE, *i.e.,* whether the explanation reflects the true reasoning process of the model.

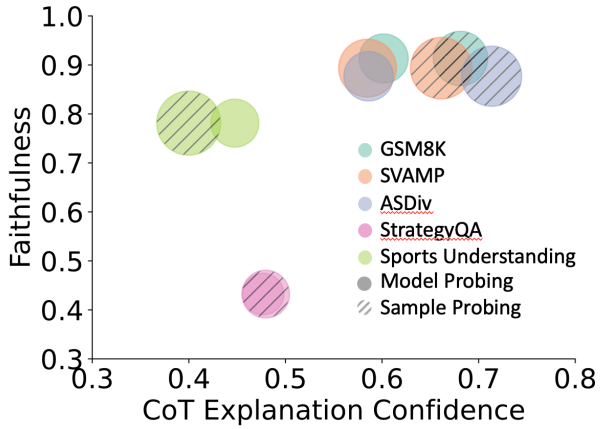

Figure 6: Mean explanation confidence for CoT explanations generated using InstructGPT for five datasets. We find that the explanation confidence is positively correlated with faithfulness for four datasets, *i.e.,* highly confident explanations tend to be more faithful. The circle size denotes the deviation in confidence.

RQ3) Correct answers have more certain explanations. Across five datasets and two probing uncertainty metrics, Fig. 5 shows that explanations of correct answers have higher explanation confidence compared to explanations of wrong answers. Our observation aligns with the general expectation that models tend to provide more reliable and confident explanations when they make correct predictions as opposed to incorrect ones.

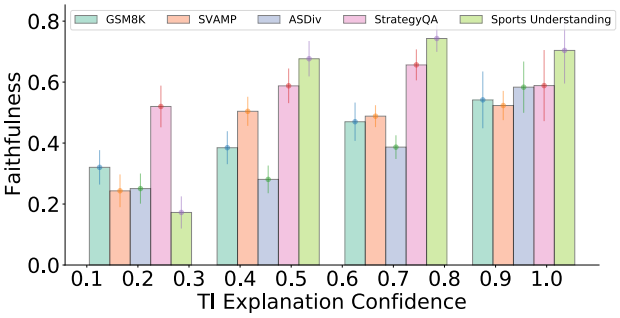

Figure 7: Mean explanation confidence for token importance explanations generated using InstructGPT for five datasets. We find that the explanation confidence is positively correlated with faithfulness, *i.e.,* highly confident explanations tend to be more faithful.

RQ4) Ablation study. We conduct ablation on five key components of our proposed probing metrics i) the number of paraphrases we generate in sample probing, ii) the number of responses we generate at temperature $\tau = 1$ in model probing, and iii) different LLMs (iv) the temperature scaling parameter τ, inherent in most LLM APIs, (v) different entailment models for measuring CoT explanation agreement as defined in Eq. 5. Results in Fig. 8 show that the explanation confidence saturates as we increase the number of paraphrases of the original question Q and our chosen value of 10 is well justified. In addition, we observe that the explanation confidence using our proposed model probing technique shows similar behavior irrespective of the number of responses we generate using the LLM at $\tau = 1$ (Fig. 9). From Figs. 10-11 and Fig. 12, we also observe that the trend of correct answers having less uncertain explanations holds true across different models and temperature scaling values. Moreover, from Fig. 14 and Fig. 13, we also observe that uncertainty

metric for CoT explanations is near-agnostic of the entailment model used. These findings justify our choices of hyperparameters in quantifying the uncertainty in different types of NLEs.

6 CONCLUSION

While improving the explainability of LLMs is crucial to establish user trust, and better understand the limitations and unintended biases present in LLMs, it is also crucial to quantify the reliability of the generated explanations using uncertainty estimates. In this work, we present a novel way to estimate the uncertainty of natural language explanations (NLEs) using verbalized and probing techniques. Specifically, we propose uncertainty metrics to quantify the confidence of generated NLEs from LLMs and compare their reliability. We test the effectiveness of our metrics on math word problem and commonsense reasoning datasets and find that i) LLMs exhibit a high degree of overconfidence when verbalizing their uncertainty in explanations, ii) explanation confidence is positively correlated with explanation faithfulness, and iii) correct predictions tend to have more certain CoT explanations compared to incorrect predictions. Our work paves the way for several exciting future works in understanding the uncertainty of the natural language explanations generated by LLMs.

References

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations, 2023.

Anthropic. Model-card-claude-2.pdf. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf, 2023. (Accessed on 10/12/2023).

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv*, 2021.

Joseph F DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models, 2020.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *TACL*, 2021.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, mar 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145%2F3571730.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2020.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models, 2023.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2017.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp, 2016a.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016b. URL http://arxiv.org/abs/1612.08220.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022.

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

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning, 2023.

James Manyika. An overview of bard: an early experiment with generative ai, 2023.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. *arXiv*, 2021.

R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *KDD*, 2016.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise. *arXiv*, 2017.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv*, 2022.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Hugo Touvron. Llama 2: Open foundation and fine-tuned chat models, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava,

Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15109–15126, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.842. URL https://aclanthology.org/2023.acl-long.842.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Jesse Vig. Visualizing attention in transformer-based language representation models, 2019.

Boxin Wang, Chejian Xu, Xiangyu Liu, Yu Cheng, and Bo Li. Semattack: Natural textual attacks via different semantic spaces, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.383. URL https://aclanthology.org/2020.acl-main.383.

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

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Mahta Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, and Yike Guo. The potential and pitfalls of using a large language model such as chatgpt or gpt-4 as a clinical assistant, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A APPENDIX

A.1 Ablation Studies

A.1.1 Number of Paraphrases in Sample Probing Uncertainty

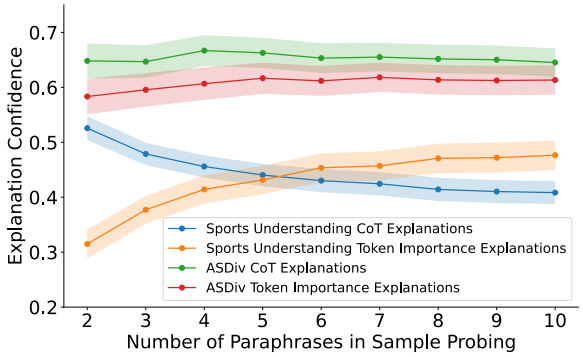

Figure 8: The effect of the number of paraphrased samples of the original Q on the mean explanation confidence of CoT and TI explanations generated from InstructGPT for Sports Understanding and ASDiv datasets. We observe that the confidence saturates as we increase the number of paraphrased samples.

A.1.2 Number of Responses in Model Probing Uncertainty

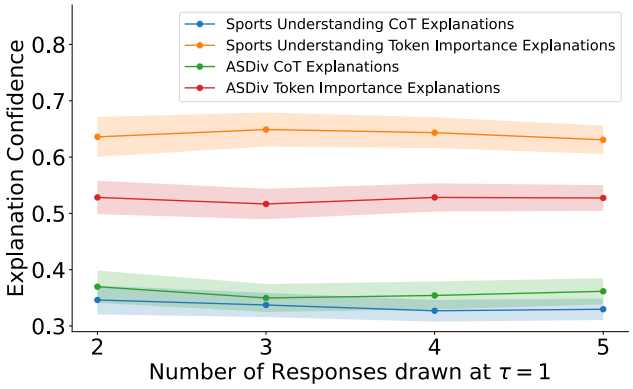

Figure 9: The effect of the number of responses drawn at $\tau = 1$ on the mean explanation confidence of CoT and TI explanations generated from InstructGPT for Sports Understanding and ASDiv datasets. We observe that the confidence remains consistent irrespective of the number of responses generated using InstructGPT.

A.1.3 LLMs

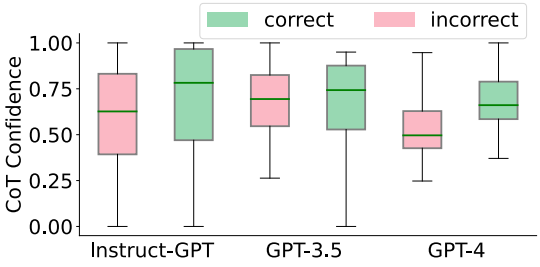

Figure 10: Comparison of chain of thought explanation uncertainty using sample probing across InstructGPT, GPT-3.5, and GPT-4 models on GSM8K dataset. We observe that the trend of correct answers having less uncertain explanations holds true across models.

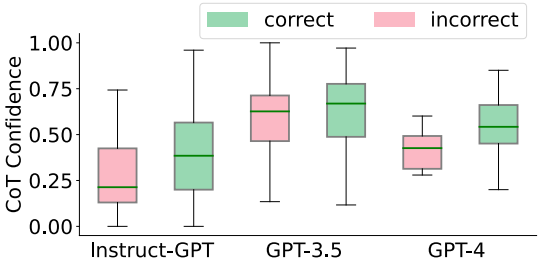

Figure 11: Comparison of chain of thought explanation uncertainty using model probing across InstructGPT, GPT-3.5, and GPT-4 models on GSM8K dataset. We observe that the trend of correct answers having less uncertain explanations holds true across models.

A.2 Temperature Scaling

The temperature scaling parameter, inherent in the APIs of most closed source LLMs, typically defaults to a value of 1. A low value of temperature gives near deterministic responses lacking a discernible signal to measure uncertainty. Temperature value ranges from 0 to 2, and hence we chose 1 (also the default value) as a middle ground. We experimented with 3 different temperature values 0.5, 1.0, 1.5 on ASDiv dataset, and see that the claim of correct answers having less uncertain explanations still holds true.

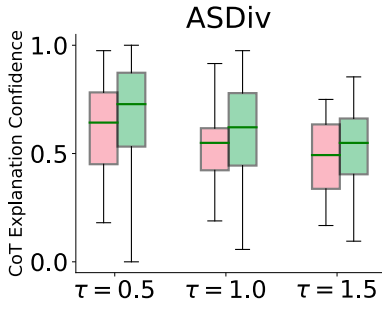

Figure 12: The effect of temperature scaling of the LLM on the mean explanation confidence of CoT explanations generated from InstructGpt on ASDiv datasets. We observe that the trend of correct answers having more certain explanations holds true across temperature values.

A.3 Entailment Models

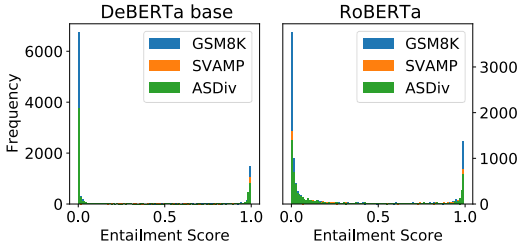

Figure 13: This histogram of entailment model scores on 3 math word problem datasets with 2 entailment models - DeBERTa Base He et al. (2021) and RoBERTa Liu et al. (2019). We can see that the distributions are concentrated at 0 and 1, which implies that these entailment models are confident.

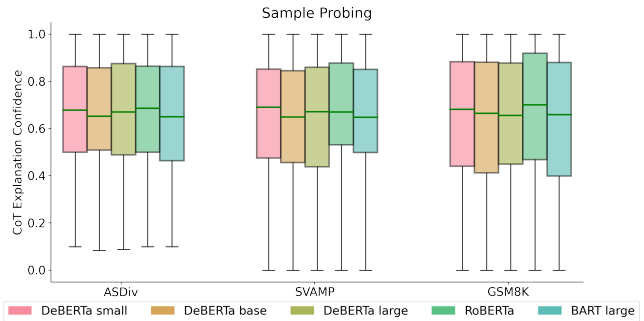

Figure 14: The box plot of explanation confidence scores on 3 math word problem datasets with 5 entailment models - RoBERTa Liu et al. (2019), BART large Lewis et al. (2019), and DeBERTa small, base and large He et al. (2021) As we can see the distributions are near identical for all entailment models indicating that CoT explanation uncertainty metric is near agnostic of the entailment model used.

A.4 Prompts

The questions used to generate token importance and chain of thought explanations are described in Fig. 15 and Fig. 16 respectively. For sample probing and model probing uncertainty, we further tailor the prompt according to the dataset. Tailoring the question prompt helps in parsing answers and explanations from generated responses. The prompts used for each dataset are as follows GSM8K Figs. 17 18 , ASDiv Figs. 19 20 , SVAMP Figs. 21 22, StrategyQA Figs. 23 24, and Sports Understanding Figs. 25 26.

A.5 Paraphrased Questions in Sample Probing

To quantify uncertainty using sample probing, semantically equivalent paraphrased questions are generated for a question using InstructGpt using the following prompt - *"Paraphrase the question into 10 different forms with the same meaning, and share them as a Python list of double quotes enclosed strings"*. An example is shown in Fig. 2.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:

1. [Word 1 here]
2. [Word 2 here]
...
...
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 15: The prompt Q_e prepended to the question Q to elicit a token importance explanation TI along with an answer A.

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here]
Step 2: [Your reasoning here]
Step 3:
...
...
Step N: [Your reasoning here]
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%
Note: The confidence indicates the degree of certainty you have about your reasoning. For instance, if your confidence level is 80%, it means you are 80% certain that your reasoning is correct. Provide the answer in aforementioned format, and nothing else.

Figure 16: The prompt Q_e prepended to the question Q to elicit a chain of thought explanation CoT along with an answer A.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:

1. [Word 1 here]
2. [Word 2 here]
...
...
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 17: **GSM8K** dataset. The prompt Q_e prepended to the paraphrased question Q to generate a token importance explanation TI along with an answer A in sample probing and model probing uncertainty experiments.

Table 1: T-Test Result Comparing Explanation Confidence Scores of Correct and Incorrect Answers using GPT-3.5 and InstructGPT models for Chain of Thought Explanations of GSM8K dataset.

Dataset	Uncertainty Metric	T-Statistic	P-Value
GSM8K	Sample Probing	-0.0977	0.9224
	Model Probing	0.7400	0.4611
SVAMP	Sample Probing	1.7913	0.0763
	Model Probing	1.2307	0.2214
ASDiv	Sample Probing	1.3031	0.1959
	Model Probing	1.7922	0.0765
StrategyQA	Sample Probing	-0.2752	0.7838
	Model Probing	-0.9779	0.3305
Sports Understanding	Sample Probing	1.3941	0.1665
	Model Probing	1.0851	0.2806

(i) GPT-3.5

Dataset	Uncertainty Metric	T-Statistic	P-Value
GSM8K	Sample Probing	1.5694	0.1198
	Model Probing	3.2404	0.0016
SVAMP	Sample Probing	2.6388	0.0097
	Model Probing	0.7660	0.4455
ASDiv	Sample Probing	3.7558	0.0003
	Model Probing	5.1783	0.0000
StrategyQA	Sample Probing	-0.1642	0.8699
	Model Probing	-0.1015	0.9194
Sports Understanding	Sample Probing	-0.8499	0.3975
	Model Probing	0.6971	0.4874

(ii) InstructGPT

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], Confidence: [Your confidence here]%
Step 2: [Your reasoning here], Confidence: [Your confidence here]%
Step 3:
...
...
Step N: [Your reasoning here], Confidence: [Your confidence here]%
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%
Note: The confidence indicates the degree of certainty you have about your answer. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct. Provide the answer in aforementioned format, and nothing else.

Figure 18: **GSM8K** dataset. The prompt Q_e prepended to the paraphrased question Q to elicit a chain of thought explanation CoT along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:

1. [Word 1 here]
2. [Word 2 here]
. . .
. . .
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 19: **ASDiv** dataset. The prompt Q_e prepended to the paraphrased question Q to generate a token importance explanation TI along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], Confidence: [Your confidence here]%
Step 2:
...
Step 3:
...
...
Step N:
...
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%
Note: The confidence indicates the degree of certainty you have about your answer. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct. Provide the answer in aforementioned format, and nothing else.

Figure 20: **ASDiv** dataset. The prompt Q_e prepended to the paraphrased question Q to elicit a chain of thought explanation CoT along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:

1. [Word 1 here]
2. [Word 2 here]
. . .
. . .
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 21: **SVAMP** dataset. The prompt Q_e prepended to the paraphrased question Q to generate a token importance explanation TI along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], Confidence: [Your confidence here]%
Step 2:
...
Step 3:
...
...
Step N:
...
Final Answer and Overall Confidence (0-100): [Your answer as a number here], [Your confidence here]%
Note: The confidence indicates the degree of certainty you have about your answer. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct. Provide the answer in aforementioned format, and nothing else.

Figure 22: **SVAMP** dataset. The prompt Q_e prepended to the paraphrased question Q to elicit a chain of thought explanation CoT along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:
1. [Word 1 here]
2. [Word 2 here]
...
...
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer Yes/No here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 23: **StrategyQA** dataset. The prompt Q_e prepended to the paraphrased question Q to generate a token importance explanation TI along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], Confidence: [Your confidence here]%
Step 2:
...
Step 3:
...
...
Step N:
...
Final Answer and Overall Confidence (0-100): [Your answer Yes/No here], [Your confidence here]% Note: The confidence indicates the degree of certainty you have about your answer. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct. Provide the answer in aforementioned format, and nothing else.

Figure 24: **StrategyQA** dataset. The prompt Q_e prepended to the paraphrased question Q to elicit a chain of thought explanation CoT along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, and output the words important for your final answer, sorted in descending order of importance. The output format is as follows:

1. [Word 1 here]
2. [Word 2 here]
...
...
N. [Word N here]
Final Answer and Overall Confidence (0-100): [Your answer plausible / implausible here], [Your confidence here]%. Provide the answer in aforementioned format, and nothing else.

Figure 25: **Sports Understanding** dataset. The prompt Q_e prepended to the paraphrased question Q to generate a token importance explanation TI along with an answer A in sample probing and model probing uncertainty experiments.

Read the question, give your answer by analyzing step by step, and assign a confidence level to each step and the final answer. The output format is as follows:
Step 1: [Your reasoning here], Confidence: [Your confidence here]%
Step 2: ...
Step 3: ...
...
Step N: ...
Final Answer and Overall Confidence (0-100): [Your answer plausible / implausible here], [Your confidence here]% Note: The confidence indicates the degree of certainty you have about your answer. For instance, if your confidence level is 80%, it means you are 80% certain that your answer is correct. Provide the answer in aforementioned format, and nothing else.

Figure 26: **Sports Understanding** dataset. The prompt Q_e prepended to the paraphrased question Q to elicit a chain of thought explanation CoT along with an answer A in sample probing and model probing uncertainty experiments.

Table 2: Paraphrased Samples for a question in GSM8K math word problem dataset. The original question is *"How many signatures do the sisters need to collect to reach their goal?"*

What is the number of signatures the sisters need to collect to reach their goal?
How many signatures must the sisters acquire to reach their goal?
What is the amount of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters have to collect to reach their goal?
What is the total number of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters require to reach their goal?
What is the quantity of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters need to gather to reach their goal?
What is the sum of signatures the sisters need to collect to reach their goal?
How many signatures do the sisters need to acquire to reach their goal?