CLUE: Concept-Level Uncertainty Estimation for Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable proficiency in various natural language generation (NLG) tasks. Previous studies suggest that LLMs' generation process involves uncertainty. However, existing approaches to uncertainty estimation mainly focus on sequence-level uncertainty, overlooking individual pieces of information within sequences. These methods fall short in separately assessing the uncertainty of each component in a sequence. In response, we propose a novel framework for **C**oncept-**L**evel **U**ncertainty **E**stimation (CLUE) for LLMs. We leverage LLMs to convert output sequences into concept-level representations, breaking down sequences into individual concepts and measuring the uncertainty of each concept separately. We conduct experiments to demonstrate that CLUE can provide more interpretable uncertainty estimation results compared with sentence-level uncertainty, and could be a useful tool for various tasks such as hallucination detection and story generation.

1 Introduction

Large Language Models (LLMs) have demonstrated powerful abilities in generating human-like text and attaining exceptional performance in various Natural Language Processing (NLP) tasks. Previous studies indicate that the generation process of LLMs involves uncertainty (Manakul et al., 2023, Huang et al., 2023b). This uncertainty arises from the stochastic nature of the sampling process in LLMs, leading to the generation of different outputs for the same given input.

Measuring the uncertainty in LLM generation is important, as it can serve as a crucial indicator, offering insights into the reliability or diversity aspects of specific tasks. For example, in a question-answering (QA) task, high uncertainty in the model's output could be interpreted as a form of hallucination, deviating from the expectation of producing consistent answers. In contrast, in the context of a story generation task, high uncertainty could become a favorable characteristic, contributing positively to the diversity of the generated stories. Therefore, understanding and quantifying uncertainty in LLM outputs become essential, allowing for task-specific evaluations and ensuring the desired outcomes in various applications.

Various methods exist for measuring the uncertainty of LLMs' output. Previous approaches have primarily focused on measuring uncertainty at the sequence level (Manakul et al., 2023, Huang et al., 2023b), treating an entire generated sequence as a single unit. These methods are often used to detect hallucinations by identifying output sequences with high uncertainty. However, a single sequence may contain multiple pieces of information, each with different uncertainty levels. Therefore, these methods encounter the "information entanglement issue", where they can only measure the overall uncertainty of an entire sequence. This limitation hinders a nuanced evaluation of individual components. For example, as illustrated in Table 1, the output sequence in each sample may include both consistent information and distinct details. Sequence-level methods fail to discern the uncertainty of each component.

To address the information entanglement issue, we proposed a framework for **C**oncept-**L**evel **U**ncertainty **E**stimation (CLUE) for LLMs. Concepts represent the fundamental meaning of the text, independent of sequence structure or individual lexicons. We use LLMs with handcrafted one-shot example to extract comprehensive concepts from the generated output sequences. Each extracted concept is treated as an independent unit, and its uncertainty is measured separately. The extracted concepts are then evaluated by an NLI-based zero-shot text classifier, which assigns the predicted entailment score as the concept score. Lastly, the uncertainty is determined by the aver-

Prompt	Answer the question in one single sentence with details: Who is the founder of Apple?	Sequence-level uncertainty	
		Sample VRO (Huang et al., 2023b)	SelfCheckGPT-NLI (Manakul et al., 2023)
Output 1	The co-founder of Apple is Steve Jobs, who, along with Steve Wozniak and Ronald Wayne, established the company on April 1, 1976, in Cupertino, California.	0.811	0.13
Output 2	Steve Jobs, along with Steve Wozniak and Ronald Wayne, co-founded Apple Inc. in 1976, revolutionizing the technology industry with iconic products like the iPhone and MacBook.	0.81	0.614
Output 3	Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne, originating in a garage in Los Altos.	0.822	0.904
Output 4	Apple's inception in 1976 was marked by the collaboration of Steve Jobs, Steve Wozniak, and Ronald Wayne, but Wayne sold his stake shortly after, missing out on Apple's immense success.	0.835	0.862

	Extracted concepts	Concept-level uncertainty CLUE
Concept 1	Co-founders of Apple (Steve Jobs, Steve Wozniak, Ronald Wayne)	0.004
Concept 2	Apple's establishment in 1976	**0.175**
Concept 3	Location of Apple's establishment (Cupertino, California)	**3.554**
Concept 4	Iconic Apple products: iPhone and MacBook	**1.629**
Concept 5	Origination in a garage in Los Altos	**4.411**
Concept 6	Ronald Wayne's stake sale	**7.965**
Concept 7	Missed opportunity for Ronald Wayne	**6.572**

Table 1: An example of sequence-level and concept-level uncertainty in output sequences generated by LLM. The output sequences may contain both consistent information (co-founders of Apple) and varied details across individual samples. Sequence-level uncertainty (Sample VRO and SelfCheckGPT-NLI) falls short in considering each piece of information separately, therefore the produced uncertainty scores for the whole generated sentence becomes less meaningful. In contrast, by breaking down sequences into concepts, our method effectively captures concepts with high uncertainty (highlighted in colored underline), while still identifying the consistent concept (Co-founers of Apple).

age negative logarithm of the concept score with respect to each output sequence. The details of the framework are presented in Section 4.

We demonstrate the effectiveness of CLUE in concept-level hallucination detection and its application as a conceptual diversity metric for story generation. Our experimental results validate the assumption that highly uncertain concepts are more likely to be hallucinations in tasks requiring consistent output. Furthermore, CLUE demonstrates a 21% improvement in macro AUROC over the baseline method in detecting hallucinations on QA datasets. To evaluate CLUE's efficacy in addressing the information entanglement issue, we compare its accuracy in predicting human judgments with sequence-level methods using Amazon Mechanical Turk (AMT). The results reveal that it exhibits a 33% higher accuracy, indicating that our concept-level method better aligns with human judgments and is thus easier for humans to understand. We also introduce the utility of CLUE as a conceptual diversity metric for story generation.

2 Motivation

2.1 Information Entanglement Issue

Previous sequence-level uncertainty methods are limited to assessing uncertainty for the entire sequence. Given that paragraph-length sequences encompass vast amounts of information, prior methods primarily focus on sequences of sentence length. Nonetheless, even a single sentence can be lengthy and filled with extensive information. As shown in Table 1, a sentence-long sequence may still encompass multiple pieces of information simultaneously. Addressing this challenge necessitates breaking sequences down into distinct pieces of information and evaluating their uncertainty individually.

2.2 Breaking Down Sequences

To extract information contained in each sequence, it is essential to break down sequences into their constituent components. Various methods exist for sequence breakdown, such as tokenization, named-entity recognition (NER), and syntax tree parsing. Different methods lead to varying levels of infor-

mation. For example, tokenization breaks down sequences into tokens, representing the lowest level of information in natural language. To enhance generalization ability, we employ LLM prompting to break down sequences into information pieces. By designing few-shot examples for LLMs, we can easily adjust the information level. In this paper, we focus on extracting high-level concepts, which effectively capture key meanings or ideas from the given text while disregarding lexical information and sequence structure.

3 Related Work

3.1 Uncertainty Estimation for LLMs

There are numerous methods to measure uncertainty in LLMs. From the algorithmic aspect, uncertainty estimation can be categorized into two types: token-based and sampling-based methods. Token-based uncertainty relies on the output probabilistic distribution for each token from LLMs (Yuan et al., 2021, Kuhn et al., 2023, Manakul et al., 2023, Zhang et al., 2023b, Huang et al., 2023b). These methods directly measure the uncertainty of the generated sequence based on this distribution. However, they cannot be used for black-box LLMs when the output probabilistic distribution is not available. Further, the output probability is often over-confident and may not reflect the actual uncertainty. In contrast, sampling-based uncertainty methods generate multiple samples from the same input prompt and calculate the uncertainty based on these output sequences (Manakul et al., 2023, Huang et al., 2023b). For example, Huang et al., 2023b propose Sample VRO, which is calculated based on the similarity between multiple output samples. Sampling-based methods only require output sequences to calculate uncertainty, thereby making them more applicable across a wider range of LLMs.

From the uncertainty level aspect, previous uncertainty methods can be categorized into three levels: sequence-level, token-level, and word-level. Sequence-level methods treat the entire output sequence as a single unit and assess its uncertainty (Manakul et al., 2023, Huang et al., 2023b, Kuhn et al., 2023 Yang et al., 2023b, Duan et al., 2023, Chen and Mueller, 2023, Lin et al., 2023, Zhang et al., 2023b, Hou et al., 2023, Rivera et al., 2024). Notably, most of the sampling-based sequence-level methods can only handle single-sentence sequences. Token-level approaches directly measure the uncertainty of individual output tokens (Tanneru et al., 2023, Duan et al., 2023, Yang et al., 2023a). Most of them leverage output token probabilities and employ functions such as entropy or the negative logarithm of the probability for uncertainty estimation. Word-level methods involve extracting keywords from output sequences and subsequently evaluating the uncertainty associated with each identified keyword (Varshney et al., 2023). The distinction between word-level and concept-level approaches lies in their functionality. Word-level methods only identify keywords present in the output sequences, whereas concept-level methods directly generate concepts based on the key meaning of the output sequence.

3.2 Hallucination in LLM Generation

Hallucination in LLMs refers to the generation of content that deviates from the input prompt or may lack grounding in reality. It is important to note that hallucination may present as a factual output but is not relevant to the input prompt. Several comprehensive surveys have been conducted to explore hallucination in LLMs (Huang et al., 2023a, Zhang et al., 2023c, Rawte et al., 2023, Ye et al., 2023). In order to improve the reliability of LLMs, extensive studies have been dedicated to the detection of hallucinations (Chen et al., 2023, Bang et al., 2023, Mündler et al., 2023, Chuang et al., 2023, Sadat et al., 2023, Mishra et al., 2024, Wang et al., 2023, Choi et al., 2023, Forbes et al., 2023, Zhang et al., 2023a, Chen et al., 2024). Specifically, some approaches leverage the uncertainty in LLMs to identify unreliable content as hallucinations (Manakul et al., 2023, Varshney et al., 2023, Zhang et al., 2023b). Furthermore, numerous studies focus on mitigating hallucinations through self-refinement by LLMs (Varshney et al., 2023, Mündler et al., 2023, Dhuliawala et al., 2023, Kang et al., 2023, Liang et al., 2024, Ji et al., 2023, Guan et al., 2023, Feldman et al., 2023). In this paper, we focus on utilizing concept-level uncertainty to detect hallucinations that deviate from the input prompt.

4 Methodology

We propose a novel framework, CLUE, to measure the uncertainty of LLMs at the concept level. CLUE extracts concepts from output sequences in each sample and then assesses concept uncertainty based on the corresponding concept score to each output sequence. An overview of our framework is

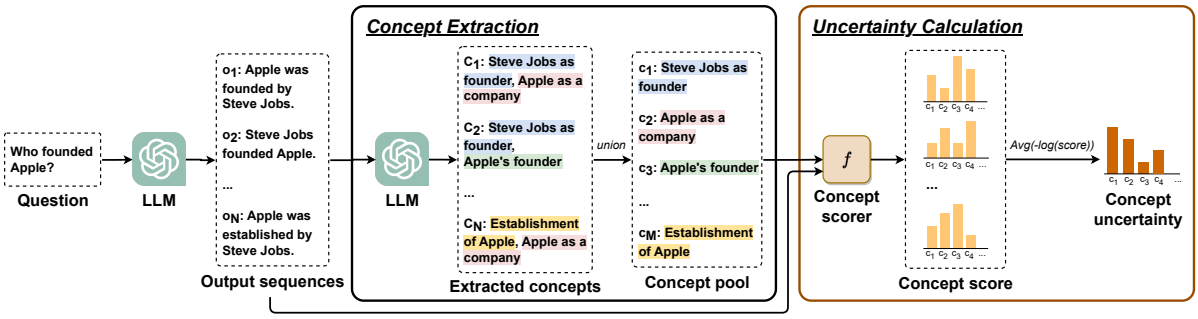


Figure 1: Our proposed framework of concept-level uncertainty. o_i denotes the i-th output sequence, C_i denotes the extracted concepts from o_i, and c_i denotes the i-th concept in the concept pool.

presented in Figure 1.

4.1 Concept Extraction

Concepts are high-level representations of texts, reflecting the meaning of sequences. To measure the uncertainty at the concept level, we extract concepts from the generated sequences by prompting LLMs. Inspired by Brown et al., 2020, we feed handcrafted one-shot example to guide LLMs in generating concepts consistently, as presented in Table 7 in the Appendix. Our analysis reveals that the length, subject, and quantity of examples barely affect the consistency of extracted concepts. We present some examples of generated sequences alongside their corresponding extracted concepts in Table 8 in the Appendix.

We extract a set of concepts for each output sequence. Since each output sequence is different, the extracted concepts also vary. To comprehensively capture the information that may be generated by the LLM, we combine the sets of concepts extracted from each output sequence to form a unified concept pool. The concept pool is composed of the possible concepts generated by the LLM based on the given prompt. Since some extracted concepts may exhibit high similarity, we use an NLI-based zero-shot text classifier to automatically consolidate similar concepts, retaining only one instance. For example, consider the two closely related concepts: "Limited competition among ISPs" and "Lack of competition in broadband market", we randomly select one of these concepts to condense the concept pool. The zero-shot text classifier is employed to measure the similarity between concepts by computing their mutual entailment scores. The two concepts are regarded as equivalent if both entailment scores are higher than the predefined threshold. The threshold is set at 0.99 to ensure stringent selection, allowing only very similar concepts to be considered equivalent. The details of the classifier are presented in Appendix A.

4.2 Concept-level Uncertainty Calculation

4.2.1 Concept Scorer

To measure the concept score based on the relevance between concepts and each output sequence, we design a concept scorer f using an NLI-based zero-shot text classifier. Given a sequence o_i and a concept c_j, the NLI-based zero-shot text classifier determines whether o_i entails c_j and outputs a probability of entailment. High entailment probability indicates that c_j is a concept of o_i. We adopt the entailment probability as the concept score s_{ij}. The details of the classifier are presented in Appendix A.

$$s_{ij} = f(o_i, c_j). \tag{1}$$

4.2.2 Uncertainty Calculation

We measure the concept score for each concept with respect to each sampled output sequence using the concept scorer. The concept uncertainty is determined by calculating the average of the negative logarithm of the concept score

$$U(c_j) = \text{Avg}_i(-\log(s_{ij})) = -\frac{1}{N}\sum_i \log(s_{ij}), \tag{2}$$

where $U(c_j)$ denotes the uncertainty of the concept c_j, and N is the number of samples. Since we employ a sampling-based method for uncertainty calculation, our approach is applicable to both white-box and black-box LLMs.

5 Experiments

We conduct experiments on various NLP tasks to demonstrate the utility of the proposed framework. In Section 5.2, we illustrate how CLUE detects hallucination at the concept level, which is more

intuitive for humans to comprehend compared to sequence-level methods. In Section 5.3, we extend our framework to another application as a conceptual diversity metric for story generation.

5.1 Experimental Settings

We evaluate the effectiveness of CLUE using question-answering (QA) datasets, which comprise multiple positive and negative instances. In the context of QA, high uncertainty indicates unpredictability in the generated output. Since stability and consistency are expected in QA tasks, high uncertainty implies potential hallucinations. To prove this statement, we partition the QA datasets into three derivative subsets: the relevant subset D_R, the less relevant subset D_L, and the irrelevant subset D_I. D_R and D_L consist of positive and negative instances, respectively, while D_I contains questions paired with answers randomly selected from other instances. It is noteworthy that the answers in D_L are more accurate than those in D_I, as the incorrect answers of QA datasets are still crafted to respond to the corresponding questions. An illustrative example of the distinctions among the three subsets is presented in Table 2. We subsequently compute the answer concept score S_j^a for each subset to represent the relevance between the answer a and the concept c_j using the Concept Scorer f:

$$S_j^a = f(a, c_j). \tag{3}$$

These answer concept scores then serve as the ground truth for the subsequent evaluation.

	What county is Farmington Hills, MI in?
relevant D_R	It is the second largest city in Oakland County in the U.S. state of Michigan.
less relevant D_L	In 2010, the area ranked as the 30th safest city in America.
irrelevant D_I	The books have since been published by many publishers worldwide.

Table 2: An example illustrating the three dataset subsets.

5.1.1 Models

We conduct experiments using OpenAI's GPT-3.5-turbo-instruct model. During the sampling stage, we set the temperature to 1 and generate $N = 5$ samples to produce different outputs while preserving the necessary contextual information for coherent and meaningful responses. In the Concept Extraction stage, we set the temperature to 0 to ensure more stable and deterministic results for the extracted concepts. Additionally, we adopted the NLI-based zero-shot text classifier "bart-large-mnli" [1] for our concept scorer. It is based on the bart-large model (Lewis et al., 2020), pretrained on the MNLI dataset (Williams et al., 2018).

5.1.2 Datasets

We select three datasets with different characteristics for a thorough evaluation. ELI5-Category is a long-form QA dataset with paragraph-like answers. WikiQA consists of simple answers, each sequence comprising only one sentence. QNLI is an NLI-based QA dataset that includes answers categorized as either entailing the corresponding questions or not. We construct three subsets D_R, D_L, and D_I for each dataset.

ELI5-Category The ELI5-Category dataset (Gao et al., 2021) is a more recent and compact variant of the original ELI5 dataset (Fan et al., 2019). It is constructed by collecting questions and their answers from `r/explainlikeimfive` subreddit. Each instance contains a single question paired with multiple answers, with each answer being assigned a score. The score is determined by subtracting the number of downvotes from the number of upvotes given by annotators. A higher score indicates a better answer. In our experiment, we select answers with the highest and lowest scores for D_R and D_L. As for D_I, we randomly choose an answer from another instance to serve as the irrelevant answer.

WikiQA The WikiQA dataset (Yang et al., 2015) consists of 3,047 questions initially sampled from Bing query logs. Each instance comprises a single question along with multiple answers, where the answers are sentences extracted from the corresponding Wikipedia page related to the question's topic. Annotators have labeled each answer as either correct or incorrect. In our experiment, we randomly choose one correct answer, one incorrect answer, and one irrelevant answer from another instance to form D_R, D_L, and D_I, respectively.

QNLI The QNLI (Question-answering Natural Language Inference) dataset (Wang et al., 2018) is a Natural Language Inference dataset derived from the Stanford Question Answering Dataset v1.1 (SQuAD) (Rajpurkar et al., 2016). Each instance consists of a question associated with a sen-

[1] https://huggingface.co/facebook/bart-large-mnli

Dataset		Pearson Correlation
Eli5-Category	relevant D_R	**-0.425**
	less relevant D_L	-0.374
	irrelevant D_I	-0.079
Wiki-QA	relevant D_R	**-0.488**
	less relevant D_L	-0.33
	irrelevant D_I	-0.062
QNLI	relevant D_R	**-0.488**
	less relevant D_L	-0.284
	irrelevant D_I	-0.092

Table 3: Correlation across the three different subsets of the datasets. As expected, the relevant subset has the lowest correlation and the irrelevant subset has the highest correlation. This pattern validates our assumption that concepts with high uncertainty tend to be hallucinated concepts.

tence labeled either as "entailment" or "not entailment". In our experiment, we select instances with "entailment" sentences as D_R and those with "not entailment" sentences as D_L. For D_I, we arbitrarily choose a sentence from another instance as the answer.

5.2 Uncertainty-based Concept-level Hallucination Detection

To demonstrate the application of our method for concept-level hallucination detection, we first validate the assumption that high uncertainty in output suggests hallucination. Building upon this assumption, we evaluate the effectiveness of CLUE in detecting hallucinations. We further conduct a human study showing that concept-level uncertainty is better than previous sequence-level uncertainty as it is easier for humans to understand.

5.2.1 Motivating Experiment

To verify the assumption that high uncertainty in outputs suggests hallucination, we examine the correlation between the concept uncertainty $U(c_j)$ and the answer concept score S_j^a across all concepts for each instance. We then compute the average correlation across all instances for three dataset subsets D_R, D_L, and D_I. Since the answer concept score indicates the relevance between the concept and the answer, a low correlation implies that concept uncertainty can serve as an indicator of the concept's irrelevance to the answer. In D_R, where answers are logically connected to the questions, concepts irrelevant to the answer are considered hallucinations. We expect a low correlation if the assumption holds. Conversely, in D_I, where answers are randomly selected from other instances, the answer concept score is not expected to exhibit a clear linear relationship with uncertainty. Therefore, we anticipate the correlation for D_I to approach 0. Regarding D_L, the correlation is expected to fall between that of D_R and D_I, given its intermediary relevance to the questions.

We present the experiment results of Pearson correlation between concept uncertainty and answer concept score in Table 3. As expected, across three subsets of the datasets, the correlation trend adheres to the following pattern: D_R exhibits a lower correlation than D_L, and D_L shows a lower correlation than D_I. The results demonstrate that across QA datasets with various characteristics, they consistently validate our assumption that concepts with high uncertainty tend to be hallucinated concepts. This suggests that the uncertainty of the LLM is an effective measure for assessing the faithfulness of the output across diverse circumstances. We present an example of the correlation experiment in Table 9 in the Appendix.

5.2.2 Concept-level Hallucination Detection

Based on the assumption that high uncertainty in outputs suggests hallucination, we proceed to evaluate the efficacy of uncertainty in detecting hallucination. We formulate this as a classification task and use concept uncertainty to conduct classification. To achieve this, we first construct a concept set to be classified, and the label of each concept is determined by its answer concept score, as illustrated in Equation 3. To enhance precision in concept labeling, we employ two thresholds, a high threshold θ_h and a low threshold θ_l, applied to the

Dataset	Method	Macro		Micro	
		AUROC	AUPRC	AUROC	AUPRC
Eli5-Category	CLUE	**0.871**	**0.894**	**0.795**	**0.826**
	bart-large-mnli	0.661	0.705	0.602	0.622
Wiki-QA	CLUE	**0.881**	**0.911**	**0.838**	**0.877**
	bart-large-mnli	0.712	0.748	0.677	0.701
QNLI	CLUE	**0.867**	**0.899**	**0.841**	**0.884**
	bart-large-mnli	0.761	0.798	0.76	0.789

Table 4: Experiment results of concept-level hallucination detection. CLUE consistently outperforms bart-large-mnli model across all datasets, showcasing substantial superiority in performance.

concept scores to determine the concept labels:

$$\text{label of concept } c_j = \begin{cases} 0 & \text{if } S_j^a > \theta_h \\ 1 & \text{if } S_j^a < \theta_l \\ -1 & \text{otherwise.} \end{cases}$$

A concept is categorized as an "entailed concept" (label 0) if its score surpasses the threshold θ_h. Conversely, if the score falls below θ_l, the concept is designated as a "hallucinated concept" (label 1). For this experiment, we do not consider other concepts (label -1). We exclusively apply this task on D_R since we require accurate answers from positive instances to label concepts.

As for the metrics, we employ AUPRC (Area Under Precision-Recall Curve) along with AUROC (Area Under the Receiver Operating Characteristic Curve) to evaluate the classification performance. Given that each instance contains a concept pool with multiple concepts to be classified, it can be viewed as an independent classification task. We present both macro and micro versions of these two metrics to provide insights into the overall performance across all classifications. Additionally, we compare CLUE to the NLI-based zero-shot classifier "bart-large-mnli" to demonstrate the efficacy of our approach. The details of the classifier are presented in Appendix A.

The results of the concept-level hallucination detection experiment are presented in Table 4. CLUE achieves remarkable performance, significantly outperforming the baseline method in detecting hallucinations. Due to the disparity in units between our method and sequence-level uncertainty, direct comparisons of hallucination detection performance with previous methods are not feasible. Table 1 provides an example to illustrate that the primary issue with sequence-level uncertainty lies not in its performance but in its unit. The ablation studies on the thresholds of concept scores are presented in Appendix B.

5.2.3 Human Study

To show that concept-level uncertainty is easier for humans to comprehend, we conduct an experiment directly comparing it with sequence-level uncertainty through human evaluation. We generate 100 instances, each comprising a question, along with 2 output sequences and 2 extracted concepts. One sequence and concept exhibit high uncertainty, while the other sequence and concept demonstrate low uncertainty. We treat this task as a binary classification problem and assess the accuracy of using uncertainty to predict the irrelevant option. We employ SelfCheckGPT-NLI (Manakul et al., 2023) as the sequence-level method for comparison. The instances are labeled using Amazon Mechanical Turk (AMT), where MTurkers are asked to select the concept and sequence they deem more relevant to the given question, as presented in Figure 2 and Figure 3 in the Appendix. To ensure the reliability of human annotations, we assign five distinct MTurkers to each instance. The label of each instance is determined based on the option selected by the majority of the MTurkers, i.e. more than 2.

Uncertainty Method	Accuracy
CLUE	**0.91**
SelfCheckGPT-NLI	0.58

Table 5: Accuracy comparison between concept-level and sequence-level uncertainty in the MTurk experiment. Our approach aligns more closely with MTurkers' judgments.

The results are presented in Table 5. Our concept-level method exhibits a 33% higher accuracy compared to the sequence-level approach. Our findings indicate that concept-level uncertainty correlates more closely with MTurkers' judgments. This suggests that CLUE serves as a more effective

indicator of the relevance of generated information to the question.

5.3 Conceptual Diversity Metric for Story Generation

As detailed in Appendix C.1, previous diversity metrics fall short in capturing high-level features such as tone or genre of generated stories. In this section, we extend the application of our framework to serve as a conceptual diversity metric in story generation.

5.3.1 Method

Since uncertainty cannot directly be used to represent diversity, we define a two-level concept structure: an upper-level concept representing a conceptual feature of generated stories, with lower-level concepts as its subclasses. For example, consider the overarching concept "tone", which includes more specific sub-concepts like "happy tone", "sad tone", "humorous tone", and so forth. We measure the diversity of the upper-level concept by aggregating the uncertainty of its lower-level concepts. Given that high uncertainty in lower-level concepts indicates that fewer generated stories are considered as the same subclasses, the aggregated uncertainty of lower-level concepts can be regarded as the diversity of the upper-level concept. We further propose two aggregation functions: the harmonic mean and entropy. The former directly measures the harmonic mean of the uncertainty of all lower-level concepts, while the latter treats it as a multi-class classification problem and measures the entropy of the classes. The equations are listed below:

$$\text{Harmonic mean} = \frac{M}{\sum_{j=1}^{M} \frac{1}{U(c_j)}}, \tag{4}$$

$$\text{Entropy} = -\sum_{j=1}^{M} \frac{n(c_j)}{N} log(\frac{n(c_j)}{N}), \tag{5}$$

where c_j denotes the j-th lower-level concept in this experiment, N is the number of samples, M is the number of concepts, and $n(c_j)$ is the number of samples classified as c_j:

$$n(c_j) = \sum_i \arg\max_k (f(o_i, c_k)) * \delta_{jk}, \tag{6}$$

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

5.3.2 Qualitative Analysis

To illustrate, we create 1000 stories by prompting LLMs to generate stories with a happy tone. We define a set of two-level concepts with an upper-level concept "tone" and 5 lower-level concepts "happy tone", "sad tone", "humorous tone", "serious tone", and "romantic tone". As depicted in Table 6, the concept scorer effectively identifies the stories with a happy tone, resulting in significantly lower uncertainty compared to the other lower-level concepts. Consequently, in the harmonic mean function, the low uncertainty term predominates in the denominator, leading to low diversity. We further create datasets with different diversity to evaluate our metrics. The experimental details are listed in Appendix C.2.

Lower-level Concept	Uncertainty
Happy tone	**0.037**
Sad tone	7.216
Humorous tone	0.284
Serious tone	2.949
Romantic tone	0.241

Table 6: Uncertainty of the lower-level concepts. Our concept uncertainty score can successfully identify the ground truth (Happy tone).

6 Conclusion

In this paper, we propose a novel framework for Concept-Level Uncertainty Estimation (CLUE) for LLMs. Our framework separates sequences into multiple concepts and measures their uncertainty individually, successfully addressing the information entanglement issue. We showcase the versatility of our framework by applying it to hallucination detection and as a conceptual diversity metric for story generation. We hope the proposed concept-based approach can achieve a more "interpretable" uncertainty estimation and can facilitate the interaction between human and LLMs.

Limitations

First, a key limitation of CLUE is its dependency on the chosen LLM for concept extraction and the specific concept scorer utilized. In this work, we generate a prompt with a one-shot example to improve the consistency of concept extraction. In future work, we will explore employing alternative white box methods for concept extraction to enhance the reliability of our framework.

Second, the lack of high-level feature diversity metrics for story generation prevents us from benchmarking CLUE's performance. However, given the customizable nature of our framework's two-level concept structure, it remains applicable across more scenarios. In future work, we aim to propose a benchmark for high-level feature diversity measurement in story generation, with CLUE serving as the baseline.

Ethical Consideration

We propose a framework for LLMs to estimate the concept-level uncertainty of generated content. The method is designed to improve LLMs' interpretability and improve human-LLM interactions. However, we do believe there could be certain risks if human over-trust the proposed uncertainty estimation tool. For example, there could be implicit biases in LLMs so that the generated biased content will be associated with low uncertainty. Therefore, when using the uncertainty estimation tool, we need to keep in mind that the estimation is measuring the LLM-generated uncertainty, not the true uncertainty of a particular concept. On the other hand, it is also possible that uncertainty estimation is manipulated by adversarial attacks, and further studies are required to improve the robustness of uncertainty estimation against those attacks.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection.

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

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 245–255, New York, NY, USA. Association for Computing Machinery.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Wenchao Du and Alan W Black. 2019. Boosting dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 38–43, Florence, Italy. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.

Philip Feldman, James R. Foulds, and Shimei Pan. 2023. Trapping llm hallucinations using tagged context prompts.

Grant C. Forbes, Parth Katlana, and Zeydy Ortiz. 2023. Metric ensembles for hallucination detection.

Jingsong Gao, Qingren Zhou, and Rui Qiu. 2021. ELI5-Category: a categorized open-domain qa dataset.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification.

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*.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation.

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

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

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

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

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying uncertainty in natural language explanations of large language models.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Lifu Tu, Xiaoan Ding, Dong Yu, and Kevin Gimpel. 2019. Generating diverse story continuations with controllable semantics. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 44–58, Hong Kong. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Hallucination detection for generative large language models by Bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371, Singapore. Association for Computational Linguistics.

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

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

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

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023b. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. 2023a. Sac^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus.

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. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

A NLI-based Zero-shot Text Classifier

An NLI-based zero-shot text classifier operates by predicting three logits, each representing the degree of the relationship between the premise and the hypothesis for the labels: "entailment", "contradiction" and "neutral". Following the instructions from the bart-large-mnli website, we disregard the "neutral" label and apply a softmax layer to the remaining two logits to derive the probability associated with the "entailment" label:

$$f(o_i, c_j) = \sigma(\text{cls}(o_i, c_j) == \textit{entailment}) \quad (8)$$

$$\sigma(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (9)$$

where σ denotes softmax function and cls denotes the classifier.

In our framework, we employ the NLI-based zero-shot text classifier for three purposes: concept consolidation, the concept scorer, and the baseline method for the hallucination detection task. For concept consolidation, the classifier computes the mutual entailment score of extracted concepts as

Extract high-level concepts like the following example:
paragraph: “Basketball, a beloved sport worldwide, has come a long way since its humble beginnings in the late 19th century. The game was originally created by Dr. James Naismith in 1891 as a way to keep his students active during the winter months. Back then, players used a soccer ball and peach baskets as makeshift goals. Fast forward to the modern era, and basketball has transformed into a high-paced, adrenaline-pumping spectacle. With legendary athletes like Michael Jordan, LeBron James, and Kobe Bryant gracing the courts, and the introduction of the slam dunk, three-point shot, and shot clock, the sport has evolved into an art form that captivates fans around the globe. The NBA, with its star-studded roster and global reach, is a testament to basketball’s enduring popularity and its remarkable journey from humble beginnings to a multimillion-dollar industry.”
concepts:“’Basketball’s origins’, ’Evolution of basketball’, ’Modern era of basketball’, ’Legendary basketball athletes’, ’Basketball’s global popularity’, ’Basketball as an art form’, ’Basketball as a multimillion-dollar industry”’

paragraph: *TARGET_OUTPUT_SEQUENCE*
concepts:

Table 7: The used prompt with the handcrafted one-shot example.

Output sequences	**Extracted concepts**
Saliva production increases when we yawn to keep the mouth lubricated and prevent dryness. When we yawn, the muscles in our face and throat contract, causing the movement of saliva and sometimes causing it to shoot out of our mouths. This is a normal and harmless bodily reaction.	“Saliva production and yawning”, “Purpose of saliva production during yawning”, “Muscles involved in yawning”, “Saliva shooting out during yawning”, “Normal bodily reaction”
The sun does not get oxygen from space. The sun is primarily composed of hydrogen and helium, and the process of nuclear fusion creates the energy and light that we see as sunlight. It does not rely on oxygen for this process.	“Composition of the sun”, “Nuclear fusion in the sun”, “Sunlight as a product of nuclear fusion”, “Sun’s energy source”, “Sun’s lack of reliance on oxygen”
When we are asleep, our eyes are still functioning, but our brain switches off our ability to send visual signals to the conscious mind. This means that while we may still react to bright lights or movements in our sleep, we are not consciously seeing in the same way as we do while awake.	“Sleep and vision”, “Brain activity during sleep”, “Consciousness and visual signals”, “Reactions during sleep”, “Difference between awake and asleep vision”
Darkness is the absence of light, and cannot be directly measured as a physical quantity. Light can be measured using units such as lumens or lux, but darkness cannot be quantified in the same way.	“Darkness’ as the absence of light”, “Measurement of light”, “Inability to measure darkness”

Table 8: Examples of generated output sequences and their corresponding extracted concepts.

their similarity. One concept serves as the premise, and the hypothesis is generated by transforming the other concept into the following format: “This concept is similar to *PREMISE_CONCEPT*”. Regarding the concept scorer, the classifier treats the output sequence as the premise and generates the hypothesis for each concept by transforming it into the following format: “This example is about *CONCEPT*”. As for the baseline method, the classifier considers the question as the premise and generates the hypothesis for each concept by transforming it into the following format: “This question is relevant to *CONCEPT*”.

Question		What does the name “Meister” mean in German?
Answer	$\mathbf{D_R}$	Meister means master in German (as in master craftsman, or as an honorific title such as Meister Eckhart).
	$\mathbf{D_R}$	Many modern day German police forces use the title Meister.
	$\mathbf{D_R}$	A rocket engine, or simply “rocket”, is a jet engine that uses only stored propellant mass for forming its high speed propulsive jet.

Concept	**Uncertainty**	**Answer concept score** $\mathbf{D_R}$	$\mathbf{D_L}$	$\mathbf{D_I}$
Skill and expertise associated with the name Meister	0.01	0.978	0.419	0.003
German origin of the name Meister	0.018	0.976	0.732	0.003
German cultural influence	0.062	0.896	0.897	0.005
Achievement and recognition	0.302	0.525	0.889	0.037
Origin of Meister	0.766	0.226	0.294	0.012
Definition of Meister	0.876	0.251	0.474	0.012
Use of Meister as a surname	1.371	0.073	0.471	0.005
Pearson Correlation		-0.954	-0.529	0.041

Table 9: An example of the correlation between concept uncertainty and answer concept score across three dataset subsets. In this instance, the concept uncertainty effectively represents the concepts' relevance to the question, resulting in a low correlation for D_R and D_L. In the case of D_I, where the answer is not accurate to the question, the uncertainty does not exhibit a linear relationship with the answer concept scores.

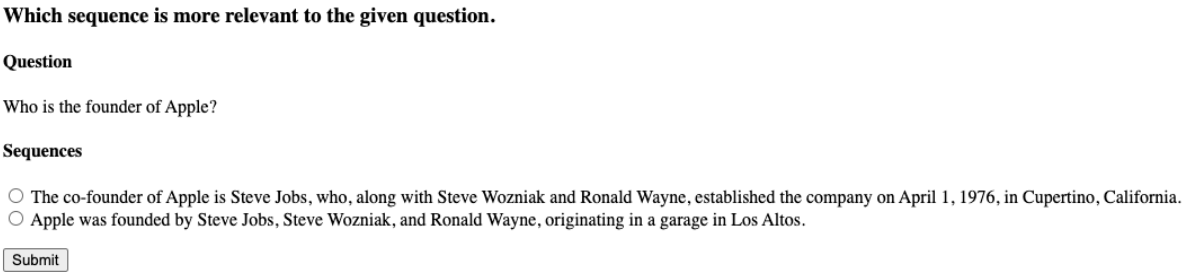


Figure 2: Screenshot of sequence-level human annotation interface presented to MTurkers.

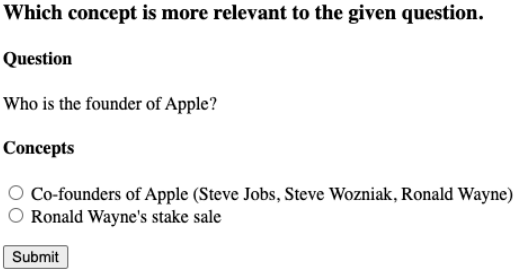
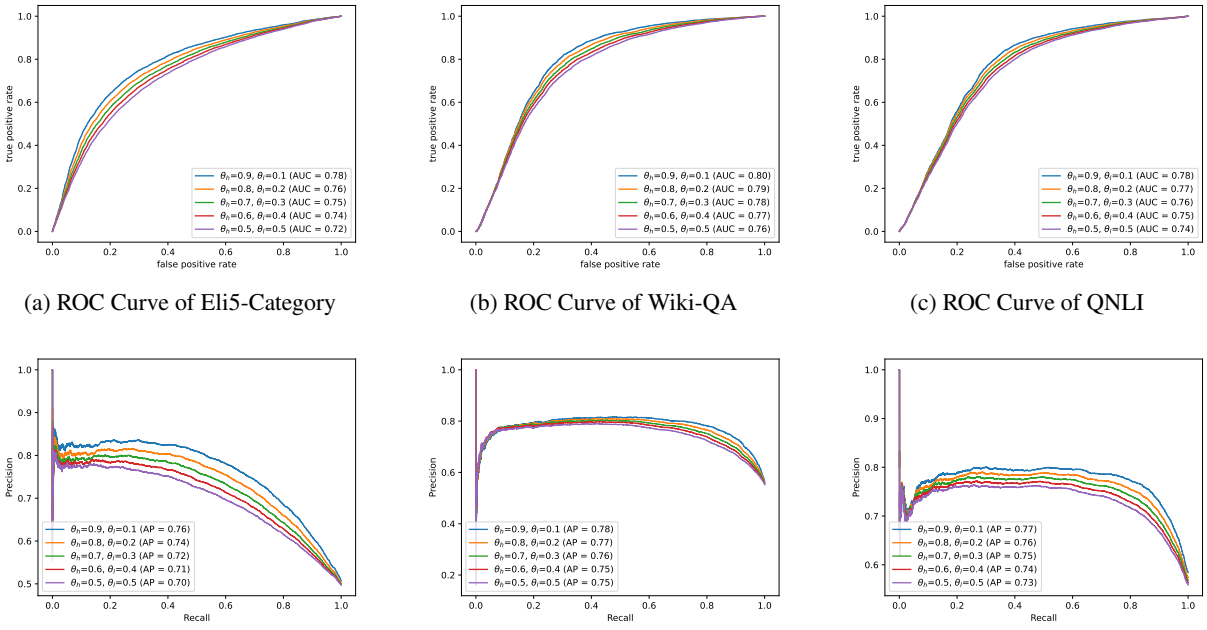


Figure 3: Screenshot of concept-level human annotation interface presented to MTurkers.

B Ablation Studies of Hallucination Detection

We further conduct ablation studies on the thresholds of concept scores. As illustrated in Figure 4, our method demonstrates better performance across all datasets when employing tighter thresholds – specifically, a higher θ_h and a lower θ_l. This observation implies that the scores predicted by our concept scorer effectively reflect the concept's faithfulness.



(a) ROC Curve of Eli5-Category (b) ROC Curve of Wiki-QA (c) ROC Curve of QNLI

(d) PR Curve of Eli5-Category (e) PR Curve of Wiki-QA (f) PR Curve of QNLI

Figure 4: ROC Curves and PR Curves on different thresholds of concept score. The results indicate that our method demonstrates better performance when utilizing tighter thresholds.

	Number of stories in each tone				
Dataset	**Happy**	**Sad**	**Humorous**	**Serious**	**Romantic**
Single-class dataset	1000	0	0	0	0
Biased dataset	600	100	100	100	100
Uniform distribution dataset	200	200	200	200	200

Table 10: Overview of story generation dataset.

	Diversity	
Dataset	**Harmonic mean**	**Entropy**
Single-class dataset	0.142	0.319
Biased dataset	0.903	1.286
Uniform distribution dataset	1.215	1.594

Table 11: Results of our proposed diversity metrics. Both metrics successfully capture the diversity of "tone" across three datasets.

C Diversity Metric for Story Generation

C.1 Related Work

Extensive research has leveraged LLMs for story generation tasks, and various metrics have also been introduced to evaluate the diversity of generated stories. Existing metrics commonly rely on quantifying diversity through measures such as the count of distinct n-grams (Yao et al., 2019, Tevet and Berant, 2021, Li et al., 2016, Goldfarb-Tarrant et al., 2020), or by employing BLEU or ROUGE scores (Papineni et al., 2002, Zhu et al., 2018, Shu et al., 2019, Xie et al., 2023, Tu et al., 2019). However, these metrics are confined to measuring lexical diversity and fail to capture high-level features such as tone or genre in story generation. While some diversity metrics based on text embeddings have been proposed to address this limitation (Lai et al., 2020, Du and Black, 2019), their applicability to story generation tasks remains unexplored.

C.2 Evaluation of Diversity Metric

To evaluate the effectiveness of our method as a diversity metric, we create three small datasets containing stories generated in different tones, as illustrated in Table 10. These datasets exhibit distinct

distributions, with the highest expected diversity in the uniform distribution dataset and the lowest diversity in the single-class dataset. We utilize the prompt "Generate a story in happy/sad/humorous/serious/romantic tone in five sentences." to generate the stories. The experiment results are presented in Table 11, demonstrating that the two proposed diversity metrics both effectively capture the diversity of the upper-level concept 'tone'.