# Scenario-independent Uncertainty Estimation for LLM-based Question Answering via Factor Analysis

Anonymous Author(s)
Submission Id: 551*

## Abstract

Large language models (LLMs) demonstrate significant potential in various applications; however, they are susceptible to generating hallucinations, which can lead to the spread of misinformation online. Existing studies address hallucination detection by (1) employing reference-based methods that consult external resources for verification or (2) utilizing reference-free methods that mainly estimate answer uncertainty based on LLM's internal states. However, reference-based methods incur significant costs and can be infeasible for obtaining reliable external references. Besides, existing uncertainty estimation (UE) methods often overlook the impact of scenario backgrounds inherited from the query's lexical resources, leading to noise in UE. In almost all real-world applications, users care about the uncertainty concerning semantics or facts instead of the query's scenario information. Therefore, we argue that mitigating scenario-related noise and focusing on semantic information can yield a more desirable UE. In this paper, we introduce a plug-and-play scenario-independent framework to enhance unsupervised UE in LLMs by removing scenario-related noise and focusing on semantic information. This framework is compatible with most existing UE methods, as it leverages only the existing UE methods' outputs. Specifically, we design a scenario-specific sampling to paraphrase queries, maintaining their common semantics while diversifying the scenario distribution. Subsequently, to estimate the contribution of the common semantics, we design a factor analysis (FA) model to disentangle the UE score obtained from the given UE method into a combination of multiple latent factors, which represent the contribution of the common semantics and scenario-related noise. By solving the FA model, we decompose the impact of the most significant factor to approximate the uncertainty caused by the common semantics, thus achieving scenario-independent UE. Extensive experiments and analysis across multiple models and datasets demonstrate the effectiveness of our approach.

## CCS Concepts

• **Computing methodologies** → **Natural language generation**.

## Keywords

Large Language Models, Hallucination, Uncertainty Estimation


*WWW'25, April 28–May 2, 2025, Sydney, Australia*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

**ACM Reference Format:**
Anonymous Author(s). 2024. Scenario-independent Uncertainty Estimation for LLM-based Question Answering via Factor Analysis. In *The 2025 ACM Web Conference, 28 April – 2 May, 2025, Sydney, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Large language models (LLMs) show potential in extensive applications across various domains [22]; however, they are prone to generating hallucinations, leading to the dissemination of factually incorrect contents across the Internet [6]. Detecting and mitigating LLM-generated hallucinations are important to building a responsible and trustworthy internet. Particularly, for online service providers of LLMs, such as OpenAI, it is crucial to avoid misleading users with misinformation.

Efforts to detect such hallucinations are categorized into two main directions [22]. (1) Initially, researchers delved into reference-based methods, which rely on consulting evidence from external resources to verify the accuracy of the information generated by LLMs [7, 18, 55]. Min et al. [40] deconstructs an LLM generation into a sequence of atomic facts and calculates the percentage of those facts that are grounded by a reliable knowledge source. Wan et al. [53] mitigate hallucinations by verifying and minimizing knowledge inconsistency between external knowledge and the intrinsic knowledge embedded in LLMs. Although these methods are effective and explainable, obtaining reliable external resources in specific tasks incurs significant costs and is sometimes inaccessible [36]. (2) Consequently, researchers explore reference-free methods that detect hallucinations by estimating the LLM's confidence in the correctness of its output. While a few researchers try to achieve this by observing LLM behaviors [27, 49] (e.g. LLM debate), a more prevalent approach is to analyze the internal states of LLMs, a concept known as uncertainty estimation (UE) [36].

The research community studies UE in both supervised and unsupervised manners. Some researchers train LLMs with carefully crafted data to achieve UE with human supervision to output uncertainty scores. Amayuelas et al. [1] construct a dataset containing known-unknown questions and fine-tune an LLM-based detector to distinguish between known and unknown queries. Cheng et al. [9] fine-tune an LLM-based detector to leverage external tools with hallucination detection trajectory data. However, these supervised approaches are costly and have been demonstrated to be sensitive to distribution shifts of the tuned detector, meaning that these methods may malfunction in applications whose data distribution is different from that of the trained detector [32]. Another group of researchers study unsupervised UE without additional supervised training. These methods freeze LLM parameters and leverage values of the model's activations or outputs to calculate token similarity [16, 34], semantics [14, 32, 46], or entropy [38] for uncertainty estimation. Duan et al. [14] distinguish semantics-relevant

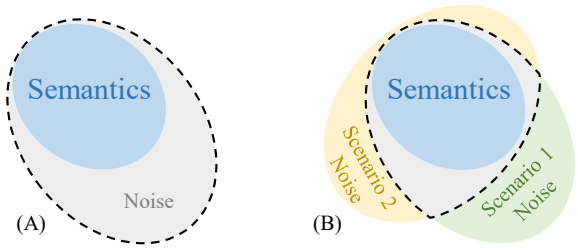


**Figure 1: The idea of our method. Black dashed lines in both subfigures indicate the UE scores obtained with (right) and without (left) applying our method. (A) The traditional scenario-dependent UE method contains both semantic information (the blue area) and various kinds of noise (the gray area). (B) Our method extracts the shared common semantic information across diverse scenarios, thus alleviating the influence of scenario-related noise (the green and yellow area) on UE.**

keywords and highlight their token entropy in UE. Kuhn et al. [32] cluster answers with similar semantics and aggregate their generation entropy to eliminate lexical influence. These studies primarily concentrate on enhancing UE by exploiting model-intrinsic information such as token probabilities, relational coherence, and semantic consistency, while often overlooking the impact of the background or scenario in which the QA is presented.

Intuitively, when a query is expressed in a different background or conversational scenario (e.g., debate or casual conversation), the human perception of the answer's uncertainty varies.[1] Given that LLMs easily capture strong style biases unrelated to factual content during training [59], we argue that similar to humans, the perception of LLMs' uncertainty may be heavily influenced by the scenario-related noise (e.g., language style, wording, and syntax). In nearly all real-world situations, users care more about how confident the LLM is about the semantics of their query, such as the meaning, facts, or knowledge involved. They are less concerned about the model's uncertainty of the conversational scenario or speaking background. Therefore, alleviating the impact of scenario backgrounds in UE and focusing mainly on the contained semantics can lead to a more desirable estimation of the LLM's uncertainty for real-world applications.

In this paper, we propose a plug-and-play scenario-independent framework to augment unsupervised UE for LLM, which disentangles the semantic information and scenario-specific information via factor analysis and focuses on semantic information for UE while ignoring scenario-specific noises[2]. The plug-and-play design means that our method is adaptive to almost all existing UE methods, including black-box LLMs (i.e. GPT-4) based UE methods, since our method only uses the traditional UE's outputs instead of accessing its model structure. Specifically, given any existing UE model, to capture that model's output distribution in various scenarios, we design a scenario-specific sampling approach for a given QA pair to generate multiple paraphrases containing the shared common semantics while differing in their scenario backgrounds (see the demonstration of our intuition in Figure 1). Subsequently, we propose a factor analysis (FA) model to disentangle the given UE model's outputs into a combination of multiple latent factors, representing the contribution of the common semantics and scenario-related noise. By solving the FA model, we decompose the impact of the most influential latent factor to approximate the uncertainty that originates from the common semantics, thus achieving a scenario-independent UE. Extensive empirical evidence verifies the effectiveness and robustness of our framework.

Our contributions are as follows: (1) To the best of our knowledge, we are the first to consider scenario information to enhance uncertainty estimation for detecting the hallucinations in LLMs; (2) We model scenario-independent uncertainty estimation with factor analysis and disentangle the semantic information from scenario-related noise to improve UE; (3) Our method achieves SOTA performance across multiple datasets and model families. Empirical evidence verifies the effectiveness of our proposed approach.

## 2 Related Work

### 2.1 Reference-based Hallucination Detection

Reference-based hallucination detection uses external references (e.g., Documents [7], knowledge graph [44]) to compare the generated outputs from LLMs with known facts, thus determining the presence of hallucinations. Huo et al. [23] combine the question with the LLM's generated answer to retrieve supporting evidence from a corpus, thereby improving the detection of hallucinations in their responses. Sansford et al. [44] introduce GraphEval and leverage knowledge graphs to provide well-structured, interpretable evaluations and corrections for hallucination detection. Min et al. [40] introduce FACTSCORE, a fine-grained evaluation metric for factual precision in long-form text generation by decomposing text into atomic facts and assessing their veracity against reliable knowledge sources. Chern et al. [10] propose a versatile framework for detecting factual errors across multiple tasks and domains by leveraging external tools, such as Google Search. These approaches rely on high-quality external references to ensure that the generated outputs are aligned with verified facts.

### 2.2 Hallucination Detection via Uncertainty Estimation

UE assesses the likelihood that the LLM-generated content is factual without relying on external references, which can be categorized into two types based on the need for supervision [32, 36].

*2.2.1 Supervised Approach for Uncertainty Estimation.* Supervised approaches train the LLM to generate a confidence score alongside its responses [1, 56] or train an additional detector [8, 9, 11, 48] to recognize hallucinations in the LLM's output. Zhang et al. [57] introduce an automated approach for creating synthetic data to train hallucination detectors, enhancing detection accuracy and latency without manual annotations. Ji et al. [26] propose an iterative self-training framework that progressively scales hallucination annotation datasets, thereby improving the accuracy of hallucination annotators in LLMs. Chen et al. [8] suggest training a discriminator

[1]For the question "Is drinking eight glasses of water a day necessary for good health?", humans may be less certain about their answer when in the scenario of an academic symposium than in a daily chit-chat.
[2]Our code is available at: anonymous.4open.science/r/WWW551.

on bilingual QA datasets to more effectively detect hallucinations in LLMs' generated answers. These supervised approaches require additional training data and computational resources. They are also sensitive to distribution shifts.

*2.2.2 Unsupervised Approach for Uncertainty Estimation.* Unsupervised methods do not require additional training and typically rely on internal signals from the LLM [5, 14, 25, 32, 46] to estimate the certainty of its outputs. Researchers explore various strategies using entropy [38], similarity [16, 34], semantic features [14, 32, 46], and information from logits or hidden states [5, 30, 42, 48] to derive uncertainty metrics [36]. Vazhentsev et al. [52] propose modeling the conditional dependency between multiple generation steps, adjusting the current generation step's uncertainty based on the previous step's uncertainty. Chuang et al. [11] identify contextual hallucinations by analyzing the ratio of attention weights between context and generated tokens. Da et al. [13] introduce a directional entailment graph and claim-level response augmentation for quantifying uncertainty, considering both semantic and logical directional information. Chen et al. [5] leverage the internal states of LLMs to detect hallucinations using an EigenScore metric from the sentence embeddings of multiple responses. When dealing with black-box models, whose internal signals are inaccessible, researchers approximate these metrics by sampling multiple outputs from the LLMs [35, 38, 39]. Manakul et al. [39] evaluate factual consistency among multiple sampled outputs. Zhang et al. [58] enhance hallucination detection performance in commonsense QA by checking response consistency across different models.

Aside from UE, researchers also try to detect hallucination by observing LLM behavior [27, 49]. Sun et al. [49] propose a Markov Chain-based multi-agent debate framework to improve hallucination detection accuracy. Cohen et al. [12] leverage interactions between two LLMs where an "examiner" LLM questions an "examinee" LLM, aiming to unveil inconsistencies of responses through multi-round conversations. While some previous work also employs query paraphrasing [60], they overlook the influence of scenario information on the hallucination detection results.

Our approach belongs to unsupervised UE methods. We explicitly consider diverse scenario information during paraphrasing and construct a factor analysis model to eliminate scenario-related noise, thus achieving a more accurate uncertainty estimation.

# 3 Method

## 3.1 Overview

We first construct a unified framework to identify the impact of scenario information on uncertainty estimation (UE) methods and then implement our algorithm. Our framework comprises two modules (see Figure 2). First, we perform scenario-specific paraphrase sampling of the query, using a scenario-dependent UE method to obtain scenario-dependent UE scores. Then, we propose a factor analysis (FA) model to achieve scenario-independent UE by disentangling and decomposing the contribution of common semantics from scenario-specific noise.

## 3.2 Plug-and-play Scenario-independent Uncertainty Estimation Framework

We propose a scenario-independent UE framework to identify the impact of scenario information on traditional UE and then mitigate its effect. Traditional UE methods $g(x, y)$ predict the probability of an LLM's answer $y$ being correct to the query $x$. However, when using any lexical representation $x$ to express the semantic $s$ of a query, the choice of lexical resources in $x$ is inherently influenced by scenario information $c$: $x \sim p(x \mid c, s)$. Here, $c$ is associated with the background in which $x$ is posed (e.g., social media, academic conferences, etc.). Note that users typically care about the uncertainty of the semantic $s$ rather than the uncertainty of $c$. However, existing (scenario-dependent) UE suffers from the noise brought by this scenario-related noise (we observe that scenario information has a significant impact on existing UE algorithms in Sec. 4.3).

Therefore, we propose a plug-and-play framework $G(C, g(x, y))$ to achieve scenario-independent UE, where $C = \{c_1, \ldots, c_m\}$ denotes a set of different scenarios. We first employ a scoring function for an LLM-based task (e.g. QA) that evaluates the quality of the generated response $f(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. For each pair of $(x, y)$, the evaluation function rates the response with the score $f(y_{true}, y)$, where $y$ is the LLM response for the query $x$ and $y_{true}$ is the ground truth. A larger score represents a more reliable answer. $G$ approximates the value of $f(y^{\text{true}}, y)$ without relying on the ground truth $y^{\text{true}}$, as shown in Eq. 1. The goal of $G$ is to minimize the prediction error relative to the true evaluation function, considering the scenarios $C$ and UE's output $g(x, y)$:

$$G(C, g(x, y)) \approx f(y^{\text{true}}, y),$$

$$\min_{G} \mathbb{E}_{(x,y,y^{\text{true}})\sim\mathcal{D}} \left\| G(C, g(x, y)) - f(y^{\text{true}}, y) \right\|_2^2 . \tag{1}$$

Our framework $G$ can build upon any scenario-dependent UE method in a plug-and-play fashion because it is agnostic to the implementation of $g(x, y)$, relying solely on its output. $G$ considers multiple possible scenarios for a given QA pair to remove the scenario-related noise and achieve a more reliable UE. We realize $G$ is in Sec. 3.3 and Sec. 3.4.

## 3.3 Scenario-specific Sampling for Query Paraphrases

We conduct a scenario-specific sampling to diversify the scenario distribution among paraphrases while retaining the query's semantics. Each paraphrased sentence corresponds to each scenario. Specifically, we instruct GPT-4o-mini [41] to generate various scenario-specific paraphrases of the query. First, we pre-define $m$ diverse scenarios based on various real-world situations, ensuring a significant divergence between them, such as chatting on social media and reporting at academic conferences, etc. To ensure that each synthesized paraphrase $\hat{x}_i$ retains the semantic meaning of the original query $x$ while only altering information related to the corresponding scenario $c_i$, we explicitly constrain this in the instruction:

$$\hat{x}_i \sim P_{LLM}(\hat{x}_i | x, c_i), \quad \forall i \in \{1, 2, \ldots, m\} \tag{2}$$

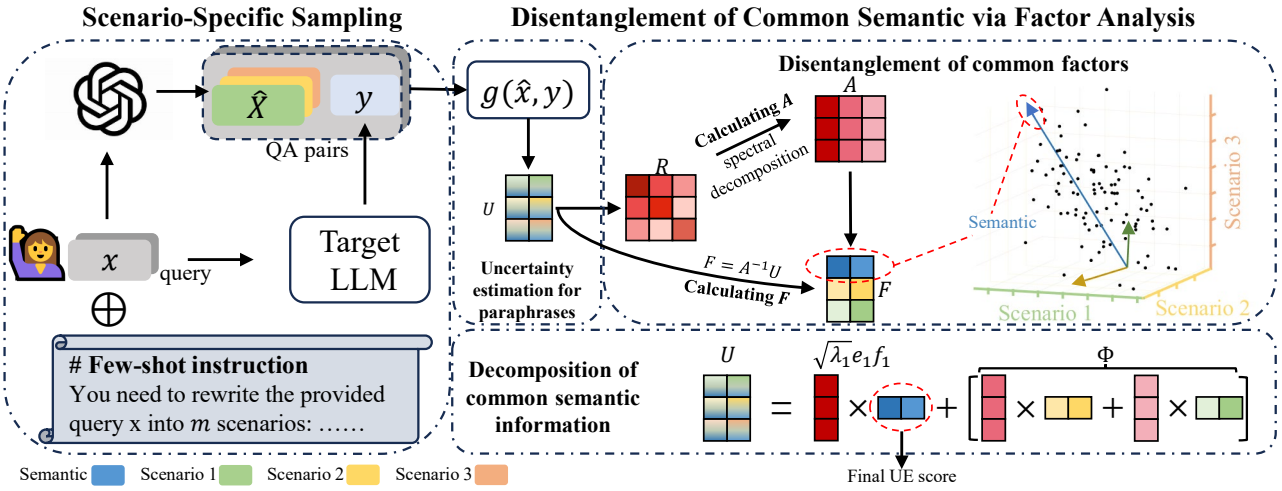


**Figure 2: The overview of our framework. For a query $x$ with an answer $y$ given by the target model, GPT-4 paraphrases $x$ considering 3 different scenarios, resulting in QA pairs $(\hat{X}, y)$. Given any scenario-dependent UE algorithm $g$, our framework disentangles its output sample matrix $U$ into multiple common factors (the row vectors of $F$). Then, we decompose the contribution of common semantic information (the blue row vector) from noise information (green and yellow vectors) to produce the final UE score.**

Our instruction template is: *You need to rewrite the provided sentences into the scenario: [SCENARIO]. Here are three examples: [EXAMPLES]. You need to paraphrase this sentence: [SENTENCE].* *"[EXAMPLES]"* represents the few-shot examples and response format we show, *"[SENTENCE]"* represents the given query that needs to be paraphrased, and "[SCENARIO]" is the scenario for paraphrasing.

Finally, for each sample $(x, y)$ in the dataset $\mathcal{D} = \{(x_i, y_i^{\text{true}}) \mid i = 1, 2, \ldots, N\}$, we instruct the LLM to paraphrase the input query $x$ based on each scenario, resulting in a set of scenario-specific paraphrases $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_m\}$. By maintaining the semantics during scenario-specific sampling, we ensure that the underlying semantics are the core of the information they have in common. Besides, by diversifying the scenario backgrounds, we reduce the similarity of different types of scenario-related noise information among paraphrases, thereby making it easier to separate them from the core semantic information (see demonstration in Fig 1).

## 3.4 Disentanglement of Common Semantic via Factor Analysis

Based on factor analysis, our method disentangles the common semantics and scenario-related noise information within UE scores. Factor analysis is a statistical technique that identifies underlying latent variables or factors. These factors explain the relationships among observed variables by expressing them as linear combinations of multiple latent variables. As we maintain the same semantics while making the noise more dissimilar in scenario-specific sampling, the contribution of the shared semantics in the common factor is larger than that of common noise. Therefore, it is natural to assume that these paraphrases' most significant common factor is their shared semantics. Consequently, by disentangling the proportion of the most influential factor in uncertainty scores from others, we can distill UE to concentrate on semantics.

Specifically, our algorithm consists of three stages. We first conduct scenario-dependent UE for paraphrases obtained in Sec. 3.3 and then disentangle the UE scores into multiple latent factors. Finally, we decompose the contribution of semantic information from scenario-related noise to achieve scenario-independent UE.

(1) **Uncertainty estimation for paraphrases.** We first feed the query $x$ into the target LLM for an answer $y$. We then apply a scenario-dependent (traditional) UE algorithm $g(x, y)$ for each group of scenario-specific paraphrases $\{(\hat{x}_i, y) | \hat{x}_i \in \hat{X}\}$, to obtain their corresponding UE scores, which we organize into a vector: $u = [g(\hat{x}_1, y), \ldots, g(\hat{x}_m, y)]^T$. This process allows us to compile a sample matrix $U_{m \times n} = [u_1, ..., u_n]$, where $n$ is the number of data samples and $m$ is the number of scenarios.

(2) **Disentanglement of common factors.** We construct a factor analysis model $U = A_{m \times m} F_{m \times n}$ to disentangle the uncertainty scores into a linear combination of a series of common factors. We denote $F = [f_1, \ldots, f_m]^T$, where each $f_i$ describes a common factor's scores across $n$ samples. The columns of loading matrix $A$ represents the loading weights of common factors, and their weighted sum yields the uncertainty scores $U$ under different scenarios. The larger the weights in $A$, the more its corresponding factor contributes to UE. It consists of the following two steps:

- **Calculating the loading matrix** $A$. We first consider the UE scores across these $m$ scenarios as the observed variables $U$ and compute its covariance matrix:

$$R_{m \times m} = \frac{1}{n-1} U U^T, \tag{3}$$

which measures the pairwise correlation between each scenario. Inspired by the PCA algorithm that reduces the dimensionality of the original data using eigen decomposition, we perform spectral decomposition [47] on the sample covariance matrix $R$ to obtain the loading matrix $A$: $R = A_{m\times m}A^T_{m\times m}$, where $A = [\sqrt{\lambda_1}e_1, \ldots, \sqrt{\lambda_m}e_m]$ (eigenvalues $\lambda$s are arranged in descending order corresponding to the $m$ common factors from most to least influential). The advantage of using the FA model is that it allows us to preserve the correlation information among UE scores from different perspectives while ensuring that different factors are orthogonal. Besides, FA helps rank the contribution of obtained factors, with larger eigenvalues in $A$ indicating greater importance of corresponding factors in $F$, thereby extracting the most significant information from $U$, i.e. common semantics.

- **Calculating the common factors matrix $F$.** Since the loading matrix $A$ is invertible (full rank), we derive the common factor matrix $F$ according to our FA model $U = AF$ as follows:

$$F = A^{-1}U. \tag{4}$$

In this way, we map the original UE scores to common factors of varying importance based on their eigenvalues. The value of each $f$ quantifies the amount of common information across all scenarios from different perspectives. The larger the eigenvalue corresponding to $f$, the more critical $f$ is in describing the shared information among paraphrases.

(3) **Decomposition of common semantic information.** We consider the common factor $f_1$ (i.e. the first row vector in $F$), corresponding to the largest eigenvalue, as representing the shared information across all sampled paraphrases for UE. Since the eigenvalue quantifies the contribution of common factors, and common semantics is the most significant factor affecting different paraphrases, $f_1$ can be used to measure the contribution of common semantics to the UE scores. Below, we decompose the common semantic information based on $f_1$ as shown in Eq.5. Specifically, we decompose the matrix $U$ into a product of column vectors of $A$ and row vectors of $F$. Further, We consider $U$ to comprise two components: (1) the product of common semantic information $f_1$ with its corresponding loading weights $\sqrt{\lambda_1}e_1$ that quantifies the proportion of $f_1$ under different scenarios; (2) the non-semantic noise information $\Phi$ that is unhelpful in UE.

$$\begin{aligned} U &= \sum_{i}^{m} A_{(:,i)}F_{(i,:)} \\ &= A_{(:,1)}F_{(1,:)} + A_{(:,2:m)}F_{(2:m,:)} \\ &= \sqrt{\lambda_1}e_1 f_1^T + \Phi. \end{aligned} \tag{5}$$

In this way, we derive the scenario-independent UE score $f_1$, which mainly considers the semantic information for each sample, to evaluate the quality of the model's responses.

Our framework can be applied in a plug-and-play manner to any unsupervised UE method. In our framework, we first apply scenario-specific sampling on the query across diverse scenarios (Sec.3.3). Next, we estimate the uncertainty for these paraphrases with arbitrary scenario-dependent UE algorithms in a plug-and-play fashion (the first step in Sec.3.4). Based on factor analysis, we then disentangle the UE scores across different scenarios into a combination of multiple common factors with varying weights (the second step in Sec.3.4). Finally, we decompose the uncertainty information related to semantics (the one with the largest eigenvalue) from scenario-related noise information (the final step in Sec.3.4). This ultimately achieves scenario-independent UE.

# 4 Experiments

## 4.1 Experimental Setting

**Implementation Details.** We conduct our experiments on a single A100 40GB GPU, using LLaMA-3 [15] 8B model and Qwen2 [2] models (0.5B, 1.5B, and 7B) as the target LLMs for uncertainty estimation. We set the number of scenarios $m$ as 6 and ask GPT-4o-mini to paraphrase the query based on the following scenarios: Chat, Academic, Research, Report, Podcast, Informal. See detailed explanations for each scenario in App. E.1. In our experiments, we use greedy decoding for the target LLM to generate the most probable answers and compare them with the gold answers to evaluate their correctness. For different UE applied in our method, we keep other hyperparameters consistent with those in their original papers. In inference, we apply 5-shot prompting for the target LLM to answer the given question. We select the first 5 samples from each dataset as demonstrations, which are not included in the testing set. See details in App. F.

**Dataset.** We apply our method on two datasets: EntityQuestions [45] and TriviaQA [28], to evaluate its performance in enhancing UE on QA tasks. EntityQuestions is a closed-book QA dataset comprising approximately 221k QA pairs, with the test set consisting of around 22k samples. The dataset consists of 24 subsets covering various knowledge types, such as places of birth, city locations, and music genres. The gold answers are typically unique and consist of a single word or phrase[3]. We evaluate our method on the test set. TriviaQA is an open-book QA dataset primarily derived from Wikipedia and the Web, containing about 95k samples. Following Kuhn et al. [32] and Welbl et al. [54], we evaluate our approach on the validation set with approximately 17k samples. The dataset typically contains gold answers in the form of a word, phrase, or short sentence, along with aliases that are also considered correct[4].

**Metric.** We adopt the widely adopted **AUROC**[20, 29, 32] to evaluate the performance of our method in measuring LLM uncertainty. See its calculation details in App. A. Following the common practice [17, 32, 46], we apply a DeBERTa-based [21] natural language inference (NLI) model[5] to evaluate whether the LLM response aligns with the gold answer. It categorizes the relationship between two sentences as either entailment, neutrality, or contradiction. We consider an LLM response to align with the gold answer if it entails the gold answer and vice versa.

**Baseline.** Our baselines consist of two parts. The first part consists of existing scenario-dependent UE algorithms that calculate scenario-dependent UE scores. The second part contains baselines we use to integrate the obtained UE scores. we compare our method with the second part to demonstrate its performance.

[3]For example, for question "Who is Birgit Rosengren married to?", the answer is "Elof Ahrle".
[4]For example, for the question "In which country was the inventor of the machine gun Hiram Maxim born?", gold answers include "America", "the U.S.", and "the USA".
[5]huggingface.co/microsoft/deberta-large-mnli

**Table 1: The performance (AUROC) of different integration strategies with various scenario-dependent baselines. The second row shows the scenario-dependent baselines we use to calculate the UE score, and the second column displays the integrated baselines we compare with ours. , the results in bold indicate the best-performing strategies. Our improvements are significant under the t-test with $p < 0.01$ (see details in App. C).**

| Dataset | Method | UE for White-box LLM | | | | | | UE for Black-box LLM |
|---|---|---|---|---|---|---|---|---|
| | | Answer PPL | P (True) | SE-UE | SE-UE (unnormalized) | PE-UE | PE-UE (unnormalized) | Self-Detect |
| EntityQuestions | **Mean** | 0.779 | 0.705 | 0.787 | 0.783 | 0.789 | 0.825 | 0.668 |
| | **PPL Re-weighting** | 0.787 | 0.743 | 0.825 | 0.821 | 0.797 | 0.835 | 0.726 |
| | **MDS** | 0.549 | 0.575 | 0.512 | 0.585 | 0.519 | 0.529 | 0.514 |
| | **Isomap** | 0.646 | 0.675 | 0.665 | 0.713 | 0.606 | 0.784 | 0.776 |
| | **Our** | **0.797** | **0.778** | **0.842** | **0.839** | **0.807** | **0.844** | **0.836** |
| TriviaQA | **Mean** | 0.826 | 0.659 | 0.787 | 0.785 | 0.829 | 0.824 | 0.711 |
| | **PPL Re-weighting** | 0.832 | 0.700 | 0.868 | 0.867 | 0.835 | 0.827 | 0.736 |
| | **MDS** | 0.558 | 0.534 | 0.519 | 0.532 | 0.565 | 0.532 | 0.537 |
| | **Isomap** | 0.547 | 0.527 | 0.710 | 0.535 | 0.561 | 0.715 | 0.822 |
| | **Our** | **0.846** | **0.762** | **0.884** | **0.882** | **0.848** | **0.837** | **0.836** |

*Scenario-dependent UE Baselines.* We apply our method to the following diverse scenario-dependent UE algorithms to evaluate our effectiveness in augmenting their UE reliability. (1) Inspired by Kadavath et al. [29], **P (True)** directly asks the LLM to output the probability of whether the QA pair is true. (2) Following perplexity-based approaches, **Answer PPL** [24] estimates the model's uncertainty on the query by calculating the perplexity of the response generated by the LLM. In addition, **PE-UE** [29] (predict entropy) samples multiple times to estimate the average perplexity instead of calculating by the greedy decoding response like PPL ("unnormalized" means not normalizing based on the token length of the output answer). (3) Following Kuhn et al. [32], **SE-UE** (semantic entropy) clusters the LLM's responses and calculates the probability of each cluster for UE ("unnormalized" means not normalizing based on the token length of the output answer). (4) Following Zhao et al. [60], we use **Self-Detect** to verify our method is also suitable for black-box UE methods. It estimates the uncertainty by considering the diversity of responses.

*Integrated Baselines.* We compare our method with other baselines by integrating UE scores from scenario-dependent methods across different scenarios to validate our method's efficiency. (1) **Mean.** Inspired by Kadavath et al. [29], we calculate the mean of uncertainty scores for different scenarios for a given scenario-dependent UE algorithm. (2) **PPL Re-weighting.** Inspired by Zhang and Wu [59], we calculate the final UE score by re-weighting and summing the UE scores of different styles based on the perplexity of the query under each scenario. (3) Inspired by Chan et al. [4], the patterns extracted by FA may not necessarily be the most helpful for UE [33, 50]. Thus, we compare our method with two similar dimensionality reduction algorithms for pattern extraction: 1) Multidimensional Scaling (**MDS**) [51] aims to extract the Euclidean distances among all samples and 2) Isometric Mapping (**Isomap**) [50], which preserves the geodesic distances (curved surface distances) information.

## 4.2 Overall Performance

We analyze the effectiveness of our method in LLM uncertainty estimation for QA tasks by comparing it with multiple baselines on LLaMA 3 8B model [15]. In addition, we evaluate Self-Detect's performance through GPT-4o-mini [41] to verify that our method also applies to black-box LLMs. First, we use GPT-4o-mini [41] to generate six additional queries in different scenarios. Then, we select the answer generated by the target LLM and use the NLI model to evaluate their correctness. Subsequently, we estimate the UE scores for paraphrases with various scenario-dependent UE baselines outlined in Sec. 4.1. Finally, we integrate the UE scores using our method alongside other baselines in Sec. 4.1 to obtain the final UE scores and evaluate their performance using AUROC. Table 1 shows the integration performance of different scenario-dependent UE methods under diverse scenarios using various integration methods.

Mean directly uses the average of UE scores for all scenarios, resulting in the worst performance, except for the baselines related to dimensionality reduction. It suggests that averaging UE scores across different scenarios does not effectively eliminate noise and then reveal the common semantics. Answer PPL uses query perplexities to re-weight UE scores. It shows a significant improvement compared to Mean, suggesting that perplexity can evaluate noise intensity by measuring sentence fluency. However, Answer PPL only measures individual sentences' perplexity while failing to adequately analyze the correlations among queries across different scenarios. This insufficiency in removing noise may account for its relative underperformance compared to our method.

Our method disentangles semantic information from scenario-related noise via FA and achieves the SOTA performance across all evaluation datasets and baselines (see examples in App. E). MDS and Isomap aim to reduce the UE scores across all scenarios to one dimension by minimizing changes in Euclidean and geodesic distances, respectively. Both methods yield the poorest results across almost all baselines, indicating that without a proper perspective, merely reducing dimensions is insufficient to derive semantic information to help UE. In addition, we also observe that Isomap performs significantly better than MDS in several baselines, which suggests that different scenarios of UE scores may exhibit manifold characteristics rather than linear correlations in high-dimensional space, which is counterintuitive. Furthermore, our method demonstrates more remarkable improvements in baselines that utilize the

**Table 2: Ablation study on the key components in our method. – Scenario ignores the requirement of diverse scenarios when paraphrasing. – FA indicates removing FA when integrating the scenario-dependent UE results. – Denosing means using insignificant factors that include noise.**

| Dataset | Model Variant | Answer PPL | P (True) | SE-UE | SE-UE (unnormalized) | PE-UE | PE-UE (unnormalized) | Self-Detect |
|---|---|---|---|---|---|---|---|---|
| Entity Question | – Scenario | 0.788 | 0.759 | 0.826 | 0.822 | 0.798 | 0.834 | 0.817 |
| | – FA | 0.775 | 0.705 | 0.787 | 0.776 | 0.789 | 0.829 | 0.668 |
| | – Denosing | 0.534 | 0.520 | 0.534 | 0.534 | 0.508 | 0.528 | 0.504 |
| | Ours | **0.797** | **0.778** | **0.842** | **0.839** | **0.807** | **0.844** | **0.836** |
| TriviaQA | – Style | 0.841 | 0.742 | 0.868 | 0.864 | 0.843 | 0.831 | 0.809 |
| | – FA | 0.825 | 0.661 | 0.788 | 0.786 | 0.831 | 0.823 | 0.711 |
| | – Denosing | 0.509 | 0.522 | 0.534 | 0.522 | 0.501 | 0.526 | 0.543 |
| | Ours | **0.846** | **0.762** | **0.884** | **0.882** | **0.848** | **0.837** | **0.836** |

diversity of answers compared to others with approximately the same AUROC score. Specifically, SE-UE shows a more significant enhancement than both Answer PPL and PE-UE, while Self-Detect exhibits notable improvements over P (True).

## 4.3 Ablation Study

We conduct an ablation study on our proposed method to verify each component's importance in improving uncertainty estimation accuracy (as shown in Table 2). – Scenario asks the model to paraphrase the original query under the same scenario and achieves suboptimal results in our experiment. Without the constraint of the scenario, the restated sentences are highly similar. Consequently, FA may erroneously interpret common noise as common semantics, thus risking the pollution of the distilled common semantics by such noise. Paraphrasing the original query into different scenarios reduces common noise and thus effectively enhances performance. In – FA, we randomly select a scenario's UE score for each test case instead of using the factor analysis model to estimate the common semantics. Compared to our full method, it integrates arbitrary scenario data and exhibits subpar performance. It indicates that scenario information harms UE but can be eliminated by our factor analysis model, thereby improving performance. In – Denoising, we select the common factors without the largest eigenvalue, which means that the patterns related to scenario noise are the major contributors to the final UE score. This approach receives the poorest results among all model variants, with AUROC values consistently ranging between 0.5 and 0.54. It indicates that these non-semantic common factors almost randomly distinguish between reliable and unreliable LLM responses. It aligns with our expectation of noise behavior and our belief that, unlike common semantic information, scenario-related noise information does not aid UE.

## 4.4 Analysis of the Selection of Common Factor

We analyze the correlation between common factors and scenario-dependent UE score to verify that the common factor corresponding to the largest eigenvalue can approximate the contribution of semantics while the other factors are noise. We use box plots to demonstrate the correlation coefficients according to the order of eigenvalue size. As Figure 3 shows, the first common factor correlates strongly with the original UE results across multiple scenarios, with correlation coefficients generally exceeding 0.8. This suggests

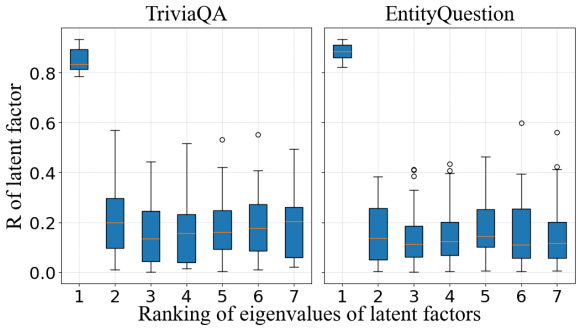


**Figure 3: The correlation between the common factors with each scenario-dependent UE baseline. The horizontal axis represents the descending order of eigenvalue ranks, while the vertical axis displays the correlation value.**

it captures the most prevalent information among paraphrases, i.e., the common semantics (refer to Sec. 3). Excluding the factor corresponding to the largest eigenvalue, the Pearson correlation coefficients of the remaining factors with the original scenario-dependent UE scores are mostly under 0.4, which indicates a weak correlation between non-first common factors and the original data. The characteristics of the remaining factors align with scenario noise, echoing the findings in Sec. 4.3 (the poor performance of – Denoising) that the common factors corresponding to noise could hardly distinguish between correct and incorrect answers.

## 4.5 Uncertainty Estimation with Different Numbers of Scenarios

To study the impact of the number of scenarios on our method, we adjust the number of additional scenarios $m$ in our framework from 1 to 6 and conduct experiments following the steps in Sec 4.2. As in Figure 4, the improvement of our method on all scenario-dependent UE algorithms exhibits a positive correlation with the number of scenarios. The improvement is particularly noticeable in cases with few scenario samples, where adding just one or two scenarios can substantially enhance performance.

**Table 3: The performance of Qwen2 models with different parameter scales. The numbers in parentheses indicate the improvement of our method over the Mean baseline.**

| Method | Model Size | Answer PPL | P (True) | SE-UE | SE-UE (unnormalized) | PE-UE | PE-UE (unnormalized) |
|---|---|---|---|---|---|---|---|
| EntityQuestions | 0.5B | 0.742 (+0.013) | 0.807 (+0.078) | 0.835 (+0.038) | 0.817 (+0.041) | 0.762 (+0.013) | 0.804 (+0.010) |
| | 1.5B | 0.780 (+0.015) | 0.761 (+0.060) | 0.841 (+0.065) | 0.805 (+0.075) | 0.781 (+0.030) | 0.763 (+0.046) |
| | 7B | 0.777 (+0.010) | 0.824 (+0.070) | 0.862 (+0.049) | 0.843 (+0.055) | 0.785 (+0.022) | 0.806 (+0.043) |
| TriviaQA | 0.5B | 0.666 (+0.013) | 0.578 (+0.037) | 0.773 (+0.059) | 0.741 (+0.063) | 0.657 (+0.020) | 0.630 (+0.024) |
| | 1.5B | 0.578 (+0.005) | 0.617 (+0.022) | 0.685 (+0.044) | 0.667 (+0.044) | 0.578 (+0.010) | 0.555 (+0.010) |
| | 7B | 0.834 (+0.020) | 0.821 (+0.079) | 0.885 (+0.048) | 0.883 (+0.054) | 0.830 (+0.031) | 0.786 (+0.037) |

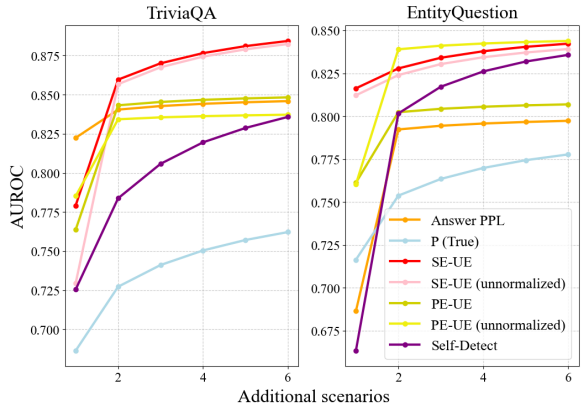


**Figure 4: Our performance on two datasets increases with the number of additional scenarios. We change the number of additional scenarios and measure the changes in our method's performance. The horizontal axis shows the number of additional scenarios, and the vertical axis shows AUROC.**

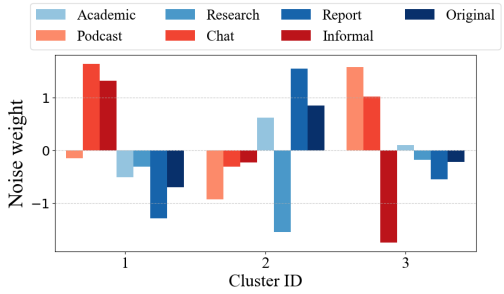


**Figure 5: Average noise weight in clusters. We conduct K-means clustering on loading weights and average them in different clusters. The horizontal axis shows the cluster ID, and the vertical axis shows the average noise weight for each scenario in different clusters. The degree of formality or informality of a scenario increases in proportion to the intensity of blue or red, respectively.**

## 4.6 Analysis of Scenario-related Noise

As the latent factors do not correspond to our pre-defined scenarios one-to-one, we use K-Means to cluster noise-related factors and analyze their patterns. Specifically, we normalize the loading weights (the columns of $A$) for each factor which denote each scenario's influence, and apply the K-Means algorithm [37] to group factors into 3 clusters (see the reasons for our choice of cluster number in App. D). We demonstrate the average loading weights for each scenario in each cluster in Figure 5.

The loading weights indicate the significance of noise in each scenario so that we can infer their causes through their distribution. For Cluster 1, Chat and Informal scenarios exhibit strong positive correlations. In contrast, all formal scenarios show negative correlations, suggesting that the presence or absence of formality may be a major noise source. In Cluster 2, informal scenarios generally have a mild impact, while Research and Report display a strong opposite impact, indicating that the noise may primarily arise from variations of written styles within formal expressions. In contrast, formal scenarios in Cluster 3 have low weights and informal scenarios strongly affect UE in mixed directions. It implies that differences within informal scenarios are the main reason for noise in Cluster 3. Overall, the clustering of noise reveals that scenario noise originates from the presence or absence of formality, writing styles within formal expressions, and differences in informal scenarios.

## 4.7 Analysis of Robustness Across Model Families and Parameter Scales

We conduct experiments on the Qwen2 model family to verify our method's robustness across different model families and parameter scales (see results in Table 3). We enhance scenario-dependent UE baselines with our method on LLM with 0.5B, 1.5B, and 7B parameters respectively. Our approach significantly improves across all model scales and scenario-dependent UE baselines, demonstrating its strong generalization ability.

## 5 Conclusion

We propose a plug-and-play unsupervised method to enhance uncertainty estimation (UE) by eliminating scenario-related noise and focusing on semantic information. We perform scenario-specific sampling, which rephrases each query into various stylistic paraphrases, capturing a range of expressions for the same underlying semantics. We then design a factor analysis model to decompose the original UE scores into multiple latent factors. By isolating the most significant factor, we disentangle the uncertainty caused by common semantics from scenario-related noise. Experiments on multiple models and datasets shows the effectiveness of our method, improving the reliability of existing UE methods.

# References


[1] Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. In *Findings of ACL*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting. https://doi.org/10.18653/v1/2024.findings-acl.383

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL] https://arxiv.org/abs/2309.16609

[3] M. S. Bartlett. 1937. Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 160, 901 (1937), 268–282. http://www.jstor.org/stable/96803

[4] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. 2015. PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE transactions on image processing* 24, 12 (2015), 5017–5032.

[5] 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. In *ICLR*. https://openreview.net/forum?id=Zj12nzlQbz

[6] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=ccxD4mtkTU

[7] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex Claim Verification with Evidence Retrieved in the Wild. In *NAACL*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico. https://doi.org/10.18653/v1/2024.naacl-long.196

[8] 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* (Birmingham, United Kingdom) *(CIKM)*. Association for Computing Machinery, New York, NY, USA, 245–255. https://doi.org/10.1145/3583780.3614905

[9] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector. https://arxiv.org/abs/2406.11277

[10] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).

[11] Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. *arXiv preprint arXiv:2407.07071* (2024).

[12] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. In *EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.emnlp-main.778

[13] Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. 2024. LLM Uncertainty Quantification through Directional Entailment Graph and Claim Level Response Augmentation. arXiv:2407.00994 [cs.CL] https://arxiv.org/abs/2407.00994

[14] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *ACL*. 5050–5063.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve

Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[16] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *TACL* 8 (2020). https://doi.org/10.1162/tacl_a_00330

[17] Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-Shot Text Classification with Self-Training. In *Conference on Empirical Methods in Natural Language Processing*.

[18] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *ICLR*. https://openreview.net/forum?id=Sx038qxjek

[19] Louis Guttman. 1954. Some necessary conditions for common-factor analysis. *Psychometrika* 19, 2 (1954), 149–161.

[20] James A. Hanley and Barbara J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[21] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *2021 International Conference on Learning Representations*. Under review.

[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232 [cs.CL]

[23] Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781* (2023).

[24] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62, S1 (1977), S63–S63.

[25] Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM Internal States Reveal Hallucination Risk Faced With a Query. arXiv:2407.03282 [cs.CL]

[26] Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical Annotation of Hallucinations in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand. https://aclanthology.org/2024.acl-long.442

[27] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.findings-emnlp.123

[28] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.

[29] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[30] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs.CL] https://arxiv.org/abs/2207.05221

[31] Henry F. Kaiser and John Rice. 1974. Little Jiffy, Mark Iv. *Educational and Psychological Measurement* 34 (1974), 111 – 117. https://api.semanticscholar.org/CorpusID:144844099

[32] 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*. https://openreview.net/forum?id=VD-AYtP0dve

[33] Jake Lever, Martin Krzywinski, and Naomi Altman. 2017. Principal component analysis. *Nature methods* 14, 7 (2017).

[34] Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022. Towards Collaborative Neural-Symbolic Graph Semantic Parsing via Uncertainty. In *Findings of ACL*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland. https://doi.org/10.18653/v1/2022.findings-acl.328

[35] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=DWkJCSxKU5

[36] Xiaocheng Li Linyu Liu, Yu Pan and Guanting Chen. 2024. Uncertainty Estimation and Quantification for LLMs: A Simple Supervised Approach. arXiv:2404.15993 [cs.LG] https://arxiv.org/abs/2404.15993

[37] J Macqueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

[38] Andrey Malinin and Mark Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *ICLR*. https://openreview.net/forum?id=jN5y-zb5Q7m

[39] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.emnlp-main.557

[40] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *ACL*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.emnlp-main.741

[41] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ Accessed: 2024-10-07.

[42] Benjamin Plaut, Khanh Nguyen, and Tu Trinh. 2024. Softmax Probabilities (Mostly) Predict Large Language Model Correctness on Multiple-Choice *Q&A*. arXiv:2402.13213 [cs.CL] https://arxiv.org/abs/2402.13213

[43] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[44] Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793* (2024).

[45] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL), 6138–6148.

[46] Lorenz Kuhn Sebastian Farquhar, Jannik Kossen and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. In *Nature*. Nature. https://www.nature.com/articles/s41586-024-07421-0#citeas

[47] C. Spearman. 1904. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology* 15, 2 (1904), 201–292. http://www.jstor.org/stable/1412107

[48] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of ACL*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting. https://aclanthology.org/2024.findings-acl.854

[49] Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. 2024. Towards Detecting LLMs Hallucination via Markov Chain-based Multi-agent Debate Framework. arXiv:2406.03075 [cs.CL] https://arxiv.org/abs/2406.03075

[50] JB Tenenbaum. 1998. Mapping a Manifold of Perceptual Observations.. In *Conference on Neural Information Processing Systems*.

[51] Warren S Torgerson. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17, 4 (1952), 401–419.

[52] Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2024. Unconditional Truthfulness: Learning Conditional Dependency for Uncertainty Quantification of Large Language Models. arXiv:2408.10692 [cs.CL] https://arxiv.org/abs/2408.10692

[53] Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge Verification to Nip Hallucination in the Bud. https://arxiv.org/abs/2401.10768

[54] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics* 6 (2018), 287.

[55] Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. GROVE: A Retrieval-augmented Complex Story Generation Framework with A Forest of Evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3980–3998.

[56] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. arXiv:2306.13063 [cs.CL] https://arxiv.org/abs/2306.13063

[57] Dongxu Zhang, Varun Gangal, Barrett Lattimer, and Yi Yang. 2024. Enhancing Hallucination Detection through Perturbation-Based Synthetic Data Generation in System Responses. In *Findings of ACL*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Bangkok, Thailand and virtual meeting. https://aclanthology.org/2024.findings-acl.789

[58] Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC$^3$: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. In *Findings of EMNLP*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.findings-emnlp.1032

[59] Xiao Zhang and Ji Wu. 2024. Dissecting learning and forgetting in language model finetuning. In *ICLR*. https://openreview.net/forum?id=tmsqb6WpLz

[60] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. In *NAACL*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico. https://doi.org/10.18653/v1/2024.naacl-long.390

# A The Calculation Process of AUROC

In this section, we briefly introduce the calculation process of AUROC (Area Under the Receiver Operating Characteristic Curve). As shown in Eq 6, it is calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings and measuring the area under the resulting curve, where TPR represents the proportion of actual positives correctly identified and FPR represents the proportion of actual negatives incorrectly identified as positives. For a UE algorithm, its AUROC value ranges from 0 to 1, with a higher value indicating better performance in distinguishing trustworthy and untrustworthy answers.

$$\text{AUROC} = \int_{x_{\min}}^{x_{\max}} \text{TPR}(x) d(\text{FPR}(x)) \tag{6}$$

# B Kaiser-Meyer-Olkin Estimate and Bartlett's Test of Sphericity

We use the Kaiser Meyer Olkin (KMO) Measure [31] and Bartlett's Test of Sphericity [3] to examine the feasibility of using factor analysis (FA) in our method. The KMO Test is a method of evaluating whether data is suitable for factor analysis. It calculates the sum of the correlation coefficients of each scenario divided by the sum of the correlation coefficients and partial correlation coefficients of each scenario. Bartlett's Sphericity Test measures the degree of correlation between variables. It first calculates the sample variance and population variance, then estimates the Bartlett statistic, and finally combines the degrees of freedom of the data to calculate the p-value. We examine the scenario-dependent baselines' UE scores on the EntityQuestion dataset and TriviaQA dataset. According to the requirements of using FA [19], our datasets show an impressive KMO Test result of above 0.8, often nearing 0.9. Moreover, Bartlett's Test is less than 0.01, which signifies the mathematical viability of our approach.

# C Significance Test Results

T-test significance testing is a statistical technique for determining if there is a meaningful difference between the means of two groups. The t-test result is usually reported with a $p$ value, expressing the likelihood of observing the data or something more extreme under the null hypothesis (which states that there is no effect or difference). We use the t-test to examine whether the improvement of our method is significant. The $p$ values in Table 5 are all smaller than 0.01, demonstrating that our improvement is significant.

# D Silhouette Coefficient of K-Means Clustering

We use silhouette coefficients [43] to measure the clustering performance and determine the appropriate number of clusters in K-Means clustering. The Silhouette coefficient measures the quality of clustering by evaluating how similar a data point is to its cluster compared to other clusters; the higher the value, the better the performance of the K-Means. As Figure 6 shows, the silhouette coefficient significantly increases when the number of clusters reaches three, and then gradually rises until it attains its maximum value of approximately 0.63 at twenty clusters. To ensure the robustness of the noise weights within each cluster, it is vital to ensure a significant number of samples per cluster. This requires maintaining

**Table 4: The results of Kaiser Meyer Olkin (KMO) Measure and Bartlett's Test of Sphericity on two datasets**

| Dataset | Method | Answer PPL | P(True) | SE-UE | SE-UE (unnormalized) | PE-UE | PE-UE (unnormalized) | Self-Detect |
|---|---|---|---|---|---|---|---|---|
| TriviaQA | **KMO Measure** | 0.949 | 0.873 | 0.939 | 0.935 | 0.950 | 0.911 | 0.924 |
| | **Bartlett's Test** | 8.791e-125 | 1.694e-97 | 6.300e-23 | 4.722e-22 | 1.035e-90 | 0 | 3.654e-97 |
| EntityQuestion | **KMO Measure** | 0.949 | 0.904 | 0.949 | 0.944 | 0.947 | 0.951 | 0.907 |
| | **Bartlett's Test** | 1.148e-14 | 8.056e-9 | 1.993e-4 | 1.465e-3 | 3.242e-11 | 6.39e-260 | 1.756e-29 |

**Table 5: The $p$ values of t-test on our method. The $p$ values are all smaller than 0.01.**

| Dataset | Answer PPL | P (True) | SE-UE | SE-UE (unnormalized) | PE-UE | PE-UE (unnormalized) | Self-Detect |
|---|---|---|---|---|---|---|---|
| TriviaQA | 8.432e-3 | 0.0 | 1.997e-13 | 2.762e-15 | 2.753e-18 | 0.0 | 1.040e-29 |
| EntityQuestions | 4.268e-28 | 0.0 | 8.000e-38 | 7.940e-49 | 0.0 | 0.0 | 2.885e-42 |

**Table 6: Examples and explanations of queries under different scenarios.**

| Scenario Name | Example | Description |
|---|---|---|
| **Original** | Which country was The Locker Room created in? | The original expression of the dataset |
| **Chat** | Do you know which country was The Locker Room created in? | The scenario of daily chat |
| **Academic** | Please elaborate on the country that the Locker Room was created. | The scenario of academic writing |
| **Research** | Which geographical region is associated with the creation of The Locker Room? | The scenario of academic discussion |
| **Report** | The Locker Room was created in which country? | The scenario of written report |
| **Podcast** | Can you tell me the country where The Locker Room was founded? | The scenario of chat in podcast |
| **Informal** | What country is home to The Locker Room? | The scenario of informal chat |

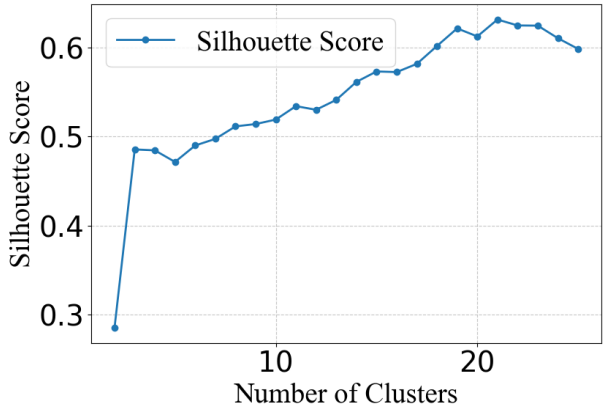


**Figure 6: The results of silhouette coefficients under different numbers of clusters. The horizontal and vertical axes denote the number of clusters and silhouette coefficients respectively. The silhouette coefficients first increase and then decrease with the increase of the number of clusters.**

relatively fewer clusters while guaranteeing the reliability of the clustering results. Consequently, we set the cluster number to 3 as the silhouette coefficient's growth began to decelerate to achieve a trade-off between K-means performance explainability.

# E Case Study

## E.1 Scenarios Interpret

We provide specific cases to illustrate the meaning of each scenario in Table 6.

## E.2 Cases of AUROC Change

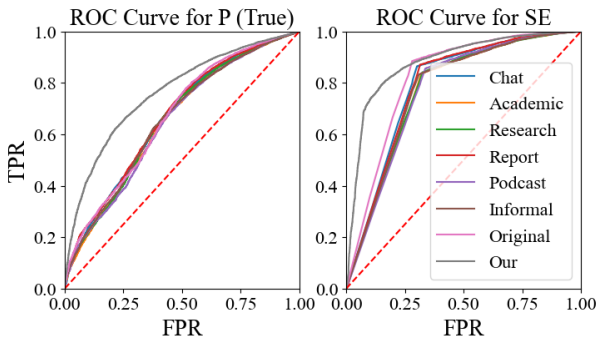


**Figure 7: The ROC curves for P (True) (left) and SE-UE (right) on the TriviaQA dataset. We examine how our method's ROC curve changes compared to different scenarios. The horizontal axis represents the proportion of negative instances incorrectly predicted as positive and the vertical axis represents the proportion of positive instances correctly predicted as positive.**

As shown in Figure 7, we select a strong baseline (SE-UE) and a weak baseline (P (True)) to demonstrate the changes in the ROC

**Table 7: Cases of UE scores. The table presents the uncertainty scores for various scenarios under P (True). The column "Score" represents the uncertainty score in each scenario. UE Result shows the UE prediction results. We consider a UE score exceeding 0.5 as the UE algorithm deeming the answer as TRUE, whereas a value below 0.5 indicates the answer is FALSE (hallucinated). The results in red indicate that the prediction results are inconsistent with the ground truth UE label.**

| QA Pair | Scenario | Score | UE Result | UE Label |
|---|---|---|---|---|
| Question: What kind of work does Peter Crisp do?<br>Gold answer: politician<br>Target LLM's answer: politician | Chat | 0.4159 | FALSE | TRUE |
| | Academic | 0.4667 | FALSE | |
| | Research | 0.8234 | TRUE | |
| | Report | 0.4867 | FALSE | |
| | Podcast | 0.4870 | FALSE | |
| | Informal | 0.6553 | TRUE | |
| | Original | 0.3297 | FALSE | |
| | Our | 0.5025 | TRUE | |
| Question: What kind of work does Ludwig Huber do?<br>Gold answer: biologist<br>Target LLM's answer: actor | Chat | 0.7645 | TRUE | FALSE |
| | Academic | 0.5948 | TRUE | |
| | Research | 0.5339 | TRUE | |
| | Report | 0.5591 | TRUE | |
| | Podcast | 0.5641 | TRUE | |
| | Informal | 0.5687 | TRUE | |
| | Original | 0.0479 | FALSE | |
| | Our | 0.4827 | FALSE | |

curve of different scenario-dependent algorithms after using our method.

Overall, the primary cause of our improvement is due to a significant reduction in the misclassification of positive examples as negative ones, while ensuring that as few negative examples as possible are incorrectly classified as positive. This represents a notable advancement over the limitations of previous uncertainty estimation methods.

### E.3 Cases of Estimate UE Score

We provide specific cases to illustrate the effectiveness of our method. As shown in Tab 7, we show the results of augmenting the P (True) algorithm with our framework. In Case 1, the LLM generation is correct, but in most scenarios, the judgments based on P (True) estimation are incorrect. Our method effectively extracts semantic information by combining other scenarios through FA and then evaluates the answer as correct, which is consistent with the UE label. In Case 2, the LLM generation is incorrect according to the gold answer, but in most scenarios, the UE results show that the answer is True. Our method successfully identifies the answer as False, which is consistent with its UE label.

## F Important Instructions

We provide important instructions involved in our experiments in Tab 8.

## G Limitations

Although our approach demonstrates improvements on black-box LLM (GPT4o), we only run our method on open-source LLMs with a relatively smaller number of parameters due to limitations of computational resources. Besides, our method require multiple paraphrases of the original query, which may incur additional computation cost.

## H Future Work

We will try to run our method on more open-source LLMs with larger sizes. In addition to GPT-4, we will also evaluate our method on more powerful black-box LLMs, for example, Claude. Besides, We observed that under certain scenario-dependent baselines, the UE scores across different scenarios exhibit manifold characteristics (see Sec. 4.2). We intend to investigate the interpretability of this finding in our future work.

Received 14 October 2024

**Table 8: Instruction templates in our experiments.**

| | Instruction templates |
|---|---|
| Scenario-Specific paraphrase | You need to rewrite the provided sentences into the scenario: [SCENARIO].<br>Here are three examples: [EXAMPLES]. You need to paraphrase this sentence: [SENTENCE]. |
| Paraphrase (– Scenario in Sec. 4.3) | You need to provide six synonymous sentences for the input sentence while maintaining the style of the original sentence. Here are three examples:[EXAMPLES].<br>You need to paraphrase this sentence: [SENTENCE]. |
| Sample answers | Question: [QUESTION], Answer: [ANSWER]. (repeat 5 times)<br>Question: [QUESTION], Answer: |
| P (True) | Question: [QUESTION], Possible answer: [ANSWER]. Is the possible answer:\n (A) True\n (B) False\n The possible answer is: [LABEL]. (repeat 5 times)<br>Question: [QUESTION], Possible answer: [ANSWER]. Is the possible answer:\n (A) True\n (B) False\n The possible answer is: |