Semantic Volume: Quantifying and Detecting Both External and Internal Uncertainty in LLMs

Xiaomin Li[1*], **Zhou Yu**[2], **Ziji Zhang**[2], **Yingying Zhuang**[2], **Swair Shah**[2], **Narayanan Sadagopan**[2], **Anurag Beniwal**[2]

[1]Harvard University
[2]Amazon

Abstract

Large language models (LLMs) have demonstrated remarkable performance across diverse tasks by encoding vast amounts of factual knowledge. However, they are still prone to hallucinations, generating incorrect or misleading information, often accompanied by high uncertainty. Existing methods for hallucination detection primarily focus on quantifying *internal uncertainty*, which arises from missing or conflicting knowledge within the model. However, hallucinations can also stem from *external uncertainty*, where ambiguous user queries lead to multiple possible interpretations. In this work, we introduce *Semantic Volume*, a novel mathematical measure for quantifying both external and internal uncertainty in LLMs. Our approach perturbs queries and responses, embeds them in a semantic space, and computes the Gram matrix determinant of the embedding vectors, capturing their dispersion as a measure of uncertainty. Our framework provides a generalizable and unsupervised uncertainty detection method without requiring internal access to LLMs. We conduct extensive experiments on both external and internal uncertainty detections, demonstrating that our Semantic Volume method consistently outperforms existing baselines in both tasks. Additionally, we provide theoretical insights linking our measure to differential entropy, unifying and extending previous sampling-based uncertainty measures such as the semantic entropy. Semantic Volume is shown to be a robust and interpretable approach to improving the reliability of LLMs by systematically detecting uncertainty in both user queries and model responses.

Code —
https://github.com/amazon-science/semantic-volume
Extended version — https://arxiv.org/abs/2502.21239

1 Introduction

Large language models encode extensive knowledge from massive training data and have shown remarkable achievements on diverse tasks (Brown 2020; Achiam et al. 2023; Touvron et al. 2023a; AI@Meta 2024; Anthropic 2023; Guo et al. 2025; Anil et al. 2023). Despite their success, LLMs still exhibit hallucination: generating information or conclusions that are incorrect, incomplete, fabricated, or misleading (Ji et al. 2023; Huang et al. 2023; Bang et al. 2023; Guerreiro et al. 2023; Chen et al. 2022; Bastounis et al. 2024). These hallucinations can propagate false information, undermine decision-making, and damage the credibility of AI systems. Detecting the hallucination is a challenging task, and a growing stream of research leverages the uncertainty in LLMs for hallucination detection (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024; Cole et al. 2023; Kadavath et al. 2022; Malinin and Gales 2020; Fomicheva et al. 2020; Kossen et al. 2024; Lin, Trivedi, and Sun 2023; Liu et al. 2024; Quevedo et al. 2024). Existing methods focus on **internal uncertainty**, which generally arises from missing relevant knowledge, conflicting information, or outdated data in the training corpus, and is assumed to reflect the model's intrinsic limitations (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024; Cole et al. 2023; Kossen et al. 2024). Nonetheless, such internal confusion, and consequently hallucinations, can also stem from **external uncertainty**, which occurs when the user's query is ambiguous, such as lacking context or having multiple possible interpretations due to typos, missing information, or ambiguous entities (Zhang et al. 2024; Min et al. 2020; Kuhn, Gal, and Farquhar 2022; Kim et al. 2024; Chi et al. 2024; Lee et al. 2023). External uncertainty cases should be handled by requesting clarification from the user (Zhang et al. 2024; Min et al. 2020; Kuhn, Gal, and Farquhar 2022; Lee et al. 2023). For example, when asked the ambiguous question "Who played Spiderman?", an LLM should ask the user to specify which movie they are referring to. It is important to note that internal uncertainty reflects true limitations of the model only after external uncertainty has been ruled out. To detect external uncertainty, current methods often rely on LLMs themselves to assess ambiguity via specialized prompting strategies (Kuhn, Gal, and Farquhar 2022; Kim et al. 2024; Chi et al. 2024; Zhang et al. 2024). In contrast, internal uncertainty (response uncertainty) detection follows two main paradigms: 1. **Probability-based** methods that utilize the token probabilities or entropy, requiring internal access to the model (Kadavath et al. 2022; Malinin and Gales 2020; Quevedo et al. 2024; Ji et al. 2024). 2. **Sampling-based** approaches, which sample multiple responses and propose measures to quantify uncertainty (Cole et al. 2023; Fomicheva et al. 2020; Zhang and Choi 2023; Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024; Kossen et al. 2024; Lin, Trivedi, and Sun 2023). A representative method in this category is the *Semantic Entropy*, which clusters sampled

*Correspondence to: Xiaomin Li (xiaominli@g.harvard.edu).

responses into semantic equivalence classes and computes entropy across these clusters (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024).

In this work, we introduce a unified sampling-based approach called the *Semantic Volume*, which can generally be applied to detect both internal and external uncertainty in LLMs without requiring access to the model weights (see Figure 1). Our method generates perturbations of queries and responses, obtains their semantic embedding vectors, and use a mathematical measure that essentially computes the determinant of the Gram matrix formed by the vectors, in order to quantify the semantic dispersion. Larger dispersion indicates higher uncertainty. More precisely, for external uncertainty, we prompt the LLM to generate multiple augmented versions of each query as perturbations, while for internal uncertainty, we sample multiple responses as perturbations. Then we take the normalized embedding vectors to be their representations. Putting these vectors as column vectors in a matrix $\boldsymbol{V}$, the value $\det(\boldsymbol{V}^\top \boldsymbol{V})$ mathematically measures the squared volume of the parallelepiped formed by these vectors. The log of this value essentially defines our Semantic Volume (see 3.1). The idea is that if uncertainty is low, all perturbations should be similar or close to each other, resulting in a small dispersion (i.e., a smaller volume).

We conduct comprehensive experiments on both query ambiguity detection and response uncertainty detection, demonstrating that our semantic volume method outperforms various baselines in both tasks. Additionally, we provide theoretical justification showing that our measure essentially captures the differential entropy of perturbation vectors, effectively quantifying overall semantic dispersion. Notably, the previously state-of-the-art *Semantic Entropy* method emerges as a special case of our approach, highlighting the broader generalization of Semantic Volume over existing sampling-based methods. Our findings suggest that Semantic Volume provides a robust, interpretable framework for improving LLM reliability by systematically detecting and addressing uncertainty in both user queries and model responses. Below is a list of our main contributions:

- We propose a novel mathematical measure for uncertainty detection, using the determinant of the Gram matrix (equivalently, parallelepiped volume) of embedding vectors to quantify semantic dispersion.
- Our method is **training-free** and **does not require internal access** to the model's hidden states or token probabilities).
- To the best of our knowledge, this is the **first framework to study both external and internal uncertainty** in LLMs.
- We validate our approach through comprehensive experiments on both external and internal uncertainty detection, demonstrating superior performance compared to various baselines.
- We provide **theoretical interpretations** of our semantic volume, linking it to differential entropy and generalizing existing sampling-based uncertainty measures.

2 Related Work

2.1 Hallucination

LLM hallucinations occur when the model generates incorrect, incomplete, fabricated, or misleading outputs (Ji et al. 2023; Huang et al. 2023; Bang et al. 2023; Guerreiro et al. 2023; Chen et al. 2022). Generally the hallucination can be caused by the lack of knowledge of the LLM itself (internal uncertainty), but could also originate from the ambiguity in the user's query (external uncertainty). Most LLMs are not trained to handle ambiguous queries and often respond incorrectly (Kim et al. 2024). Addressing external uncertainty requires a different approach, such as prompting the LLM to ask clarification questions before generating a response (Zhang et al. 2024; Min et al. 2020; Kuhn, Gal, and Farquhar 2022; Lee et al. 2023). In contrast, internal uncertainty caused by knowledge gaps can be mitigated through methods like retrieval-augmented generation (Lewis et al. 2020), reasoning (Wei et al. 2022; Guo et al. 2025; OpenAI 2024), or simply turning to stronger LLMs or human agents.

2.2 External Uncertainty

Query ambiguity detection is typically performed using LLMs with various prompting techniques (Kuhn, Gal, and Farquhar 2022; Kim et al. 2024; Min et al. 2020; Zhang et al. 2024; Chi et al. 2024; Yin et al. 2023; Lee et al. 2023). For instance, Kuhn, Gal, and Farquhar (2022) uses LLM prompting for both detecting ambiguity and generating clarification questions. Kim et al. (2024) prompts LLM to disambiguate question x itself, and then measure the difference between the x_{new} and x. Larger difference above a threshold indicates ambiguity. Min et al. (2020) introduced the AmbigQA dataset, which contains ambiguous queries and their answers. Following this, Zhang et al. (2024) proposed an ambiguity taxonomy and introduced the CLAMBER benchmark with binary ambiguity labels. They further evaluated various models on ambiguity detection under different settings, including zero-shot vs. few-shot and chain-of-thought (CoT) prompting vs. standard prompting.

2.3 Internal Uncertainty

Many studies propose uncertainty measures for hallucination detection (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024; Cole et al. 2023; Kadavath et al. 2022; Malinin and Gales 2020; Fomicheva et al. 2020; Kossen et al. 2024; Lin, Trivedi, and Sun 2023; Liu et al. 2024; Quevedo et al. 2024); similar to these, we focus on cases where LLM mistakes coincide with high uncertainty, and cases where an LLM hallucinates with high confidence are beyond the scope of this paper. There are generally two genres: **probability-based**, using information such as the token probability or entropy, and **sampling-based**, which samples more responses and measure the dispersion of the answers.

Probability-based. *Last Token Entropy*, which essentially uses the entropy of the vocabulary distribution at the last token, is a widely used measure of uncertainty (Malinin and Gales 2020). *Log Probabilities* average log conditional token probabilities (Malinin and Gales 2020; Quevedo et al. 2024). Quevedo et al. (2024) tries multiple ways to aggregate the

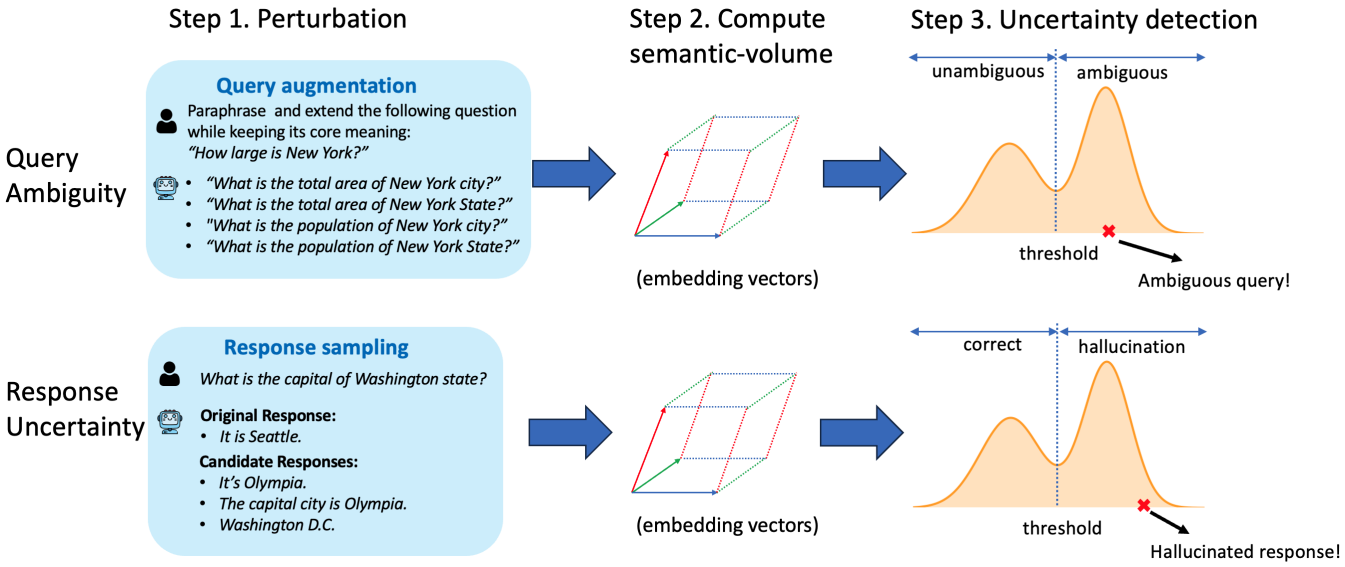


Figure 1: Pipeline for external and internal uncertainty detection using semantic volume. *Step 1.* Generate perturbations. For external uncertainty, we augment each query using an LLM, treating the augmentations as perturbations. For internal uncertainty, perturbations refer to multiple sampled candidate responses. *Step 2.* Compute semantic volume (essentially $\log\det(\boldsymbol{V}^\top\boldsymbol{V})$ where columns of $\boldsymbol{V}$ are normalized embedding vectors). *Step 3.* Cases with high semantic volume are predicted as ambiguous queries (external uncertainty) or hallucinated responses (internal uncertainty).

token probabilities, such as the minimal and averaged token probabilities.

Sampling-based. Kuhn, Gal, and Farquhar (2023) proposed *Semantic Entropy* to measure uncertainty in natural language and Farquhar et al. (2024) applied it to detect hallucinations in large language models. Essentially for each query, they generate multiple answers and then cluster them by the same semantic meanings. Then discrete entropy calculated from the sizes of different clusters is defined as the semantic entropy. In Kadavath et al. (2022) and Cole et al. (2023), they sample multiple answers and let LLM to judge the uncertainty based on these answers. this method is called *p(True)* in Kuhn, Gal, and Farquhar (2023). The *Lexical Similarity* method (Fomicheva et al. 2020; Grewal, Bonilla, and Bui 2024) considers the averaged similarity of the sampled answers, and lower similarity indicates higher dispersion.

3 Method

3.1 Definitions and Notations

Denote $[k] \overset{\text{def}}{=} \{1, 2, \ldots, k\}$ for any $k \in \mathbb{N}$. For the task of external uncertainty detection, we denote the query dataset by $\mathcal{D}_Q = \{q_i\}_{i\in[N_Q]}$, along with a tiny labeled subset $\mathcal{L}_Q \subseteq \mathcal{D}_Q$ (used to determine optimal semantic volume threshold and each query is assigned a binary label indicating whether it is ambiguous). For internal uncertainty detection, we define the query-response dataset as $\mathcal{D}_R = \{(q_i, r_i)\}_{i\in[N_R]}$ with a labeled subset $\mathcal{L}_R \subseteq \mathcal{D}_R$. Since our method and analysis apply to both tasks, we often drop the subscript and use the general notation $\mathcal{D} = \{s_i\}_{i\in[N]}$ and $\mathcal{L} \in \mathcal{D}$, where s_i represents a query for external uncertainty and a query-response pair for internal uncertainty.

Volume. Given normalized embedding vectors $\boldsymbol{V} = [\boldsymbol{v}_1\boldsymbol{v}_2\ldots\boldsymbol{v}_n]$ (each $\|\boldsymbol{v}_i\| = 1$), we define the squared volume as:

$$\mathbf{Vol}^2(\boldsymbol{V}) \overset{\text{def}}{=} \det(\boldsymbol{V}^\top\boldsymbol{V}). \tag{1}$$

The term "volume" originates from the fact that geometrically, $\sqrt{\det(\boldsymbol{V}^\top\boldsymbol{V})}$ represents the volume of the parallelepiped spanned by vectors $\{\boldsymbol{v}_i\}$. For example, in the three-dimensional case, where $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \boldsymbol{v}_3]$, it can be verified that $\sqrt{\det(\boldsymbol{V}^\top\boldsymbol{V})}$ precisely computes $|\boldsymbol{v}_1^\top(\boldsymbol{v}_2 \times \boldsymbol{v}_3)|$, which corresponds to the volume of the three-dimensional parallelepiped formed by $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$. A more detailed discussion on the geometric interpretation of this measure is provided in Appendix B.

Semantic Volume. To avoid numerical singularities from duplicate embeddings (e.g., two sampled responses or extended queries are identical), we add a small perturbation $\epsilon\boldsymbol{I}$ with $\epsilon = 10^{-10}$ and compute $\det(\boldsymbol{V}^\top\boldsymbol{V} + \epsilon\boldsymbol{I})$ to maintain numerical stability (in Appendix D, we show that ϵ is negligible to the spectral norm $\|\boldsymbol{V}^\top\boldsymbol{V}\|$ and hence it only serves to ensure numerical stability and does not affect the quantification). In practice, the absolute values of squared volumes are often small due to the nature of our perturbations. Therefore we take the logarithm, leading to the formulation

$$\log\mathbf{Vol}^2(\boldsymbol{V}) \overset{\text{def}}{=} \log\det\left(\boldsymbol{V}^\top\boldsymbol{V} + \epsilon\,\boldsymbol{I}_n\right) \tag{2}$$

Moreover, we apply Principal Component Analysis (PCA) to reduce dimensionality by projecting the vectors onto the top d principal components, obtaining $\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1\tilde{\boldsymbol{v}}_2\ldots\tilde{\boldsymbol{v}}_n] \in \mathbb{R}^{d\times n}$, where each $\tilde{\boldsymbol{v}}_i \overset{\text{def}}{=} \boldsymbol{P}_{PCA}\boldsymbol{v}_i$. Here, $\boldsymbol{P}_{PCA}$ is the projection matrix. This results in the general form of our final semantic

volume measure:

$$\begin{aligned}\text{SemanticVolume}(\boldsymbol{V}) &\stackrel{\text{def}}{=} \log \mathbf{Vol}^2(\boldsymbol{P}_{PCA}\boldsymbol{V}) \\ &= \log\det\left(\tilde{\boldsymbol{V}}^\top\tilde{\boldsymbol{V}} + \epsilon\,\boldsymbol{I}_n\right)\end{aligned} \tag{3}$$

3.2 Semantic Volume Uncertainty Detection Algorithm

Our algorithm using semantic volume to detect high uncertainty is outlined below (the overall pipeline is illustrated in Figure 1 and the detailed pseudocode is provided in Algorithm 1). Additionally, we present a Case Study in Appendix K, analyzing representative examples of queries and responses exhibiting high versus low uncertainty.

Step 1: Augmentation. For each $s \in \mathcal{D}$, we augment it with n perturbations. Precisely, for external uncertainty, we prompt an LLM to augment/paraphrase each query s to obtain n perturbed versions, while for internal uncertainty, we sample n candidate responses.

Step 2: Compute semantic volume. We obtain embedding vectors using Sentence-Transformer (Reimers 2019), normalize them, and apply PCA dimension reduction, yielding $\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1\tilde{\boldsymbol{v}}_2\ldots\tilde{\boldsymbol{v}}_n] \in \mathbb{R}^{d\times n}$ for n perturbations. Then compute the semantic volume according to (3).

Step 3: Uncertainty detection. A higher semantic volume indicates greater uncertainty. To determine the optimal threshold τ^*, we use a tiny labeled random subset $\mathcal{L} \subseteq \mathcal{D}$ with size 100 for threshold tuning [1] (the exact formula for τ^* is characterized in Proposition 1). Finally, we classify the entire dataset $\mathcal{D}$ by assigning binary uncertainty labels based on the semantic volume threshold.

4 Experiment: External Uncertainty

4.1 Experimental Setup

Data. We use two benchmark datasets. The first is CLAMBER (Zhang et al. 2024), a balanced dataset of 3K queries, each annotated with a binary label indicating whether it is ambiguous. The second is a balanced subset of 5K samples from the AmbigQA dataset (Min et al. 2020). Each question in AmbigQA is labeled as either unambiguous or ambiguous; ambiguous questions are further annotated with multiple disambiguated versions and corresponding answers, each reflecting a distinct plausible interpretation.

Models. For query augmentation, we use `Claude3.5-Sonnet` (Anthropic 2023) and the prompt is provided in Appendix J. `Qwen2-1.5B-instruct` (Yang et al. 2024) is used as the sentence-transformer to generate embeddings.

Evaluation. We conduct experiments for binary classification tasks on ambiguity of the queries. The performance is assessed by comparing the predicted binary labels against the ground truth labels, reporting both accuracy and F1 score.

Baselines. Note that the sampling-based methods discussed in Section 2.3 can be naturally extended to query ambiguity detection if we can generate analogous perturbations of the queries, similar to how candidate responses are sampled in response uncertainty detection. Below, we outline the baseline methods we consider. Some of these were originally designed for response uncertainty, but we readily adapt their methodology to query ambiguity.

- *Type 1: Prompting-based.* Directly prompt LLMs to determine whether a given query is ambiguous. We evaluate the following models: Vicuna-13B (LMSys 2023), Llama2-13B-Instruct (Touvron et al. 2023b), Llama2-70B-Instruct (Touvron et al. 2023b), Llama3.2-3B-Instruct (AI 2024), and ChatGPT (Achiam et al. 2023). We also include ChatGPT results from CLAMBER using few-shot and chain-of-thought prompting (Zhang et al. 2024).
- *Type 2: Probability-based*. Using the token probabilities to quantify uncertainty. These methods require access to the model's internal token probabilities. We consider the follow methods and use `Llama3.2-1B-Instruct` (AI 2024) to obtain the token probabilities. (a) **Last Token Entropy** (Kadavath et al. 2022; Arora, Huang, and He 2021; Malinin and Gales 2020): computes the entropy of the vocabulary distributions at the last token of the query. (b) **Log Probabilities** (Malinin and Gales 2020; Quevedo et al. 2024): measures uncertainty by computing the log of the product of conditional token probabilities, which is equivalent to summing the log conditional probabilities across all tokens in the query.
- *Type 3: Sampling-based*. These methods originally measure the variation of sampled responses to quantify response uncertainty. Here for query ambiguity, we adapt and apply them to perturbed variations of queries. (a) **p(True)** (Kadavath et al. 2022; Cole et al. 2023): the original p(True) method quantifies uncertainty based on the probability of the LLM's output. We adopt it to directly use the LLM's answer and compare it against the ground truth binary labels. (b) **Lexical Similarity** (Lin, Trivedi, and Sun 2023; Fomicheva et al. 2020): computes the averaged pairwise similarity of perturbed queries. (c) **Semantic Entropy** (Kuhn, Gal, and Farquhar 2023): clusters the perturbations into semantic equivalence classes and computes the entropy over the clusters.

4.2 Results

The performance of our method and baseline approaches on the query ambiguity classification task is presented in Table 1. Here we choose $n = 20$ for the augmentations for queries (the discussion on varying the perturbation size n is provided in Appendix F). The original CLAMBER dataset includes a diverse range of ambiguities, such as queries involving unfamiliar entities, self-contradictions, multiple meanings, and missing context. We observe that identifying ambiguous queries remains challenging for LLMs, even for powerful models like ChatGPT, despite various prompting strategies (few-shot and CoT). Among the baselines, probability-based methods achieve higher F1 scores but a critical limitation of them is that they require access to token probabilities, which is not provided for most of the close-sourced models. Among sampling-based methods, semantic entropy gen-

[1] Note that this labeled subset is only used for finding a more precise threshold. In practice, one can consider a completely unsupervised setting and use a simpler heuristic threshold such as the median.

erally performs better. Nonetheless, our semantic volume method significantly outperforms all three categories of baselines, demonstrating its effectiveness in detecting ambiguous queries. For AmbigQA task, results indicate that ambiguity detection on this dataset is generally more challenging compared to CLAMBER. Nonetheless, our Semantic Volume method continues to outperform the baseline methods.

5 Experiment: Internal Uncertainty

5.1 Experimental Setup

Data. We use subsets of the TriviaQA (Joshi et al. 2017) and SQuAD (Rajpurkar et al. 2016) datasets, which are reading comprehension benchmarks containing questions paired with reference answers. For each task, we generate responses using the evaluation LLMs (see model details below) with a temperature of 0. A response y is flagged as a hallucination if the ROUGE-L score with respect to the reference answer y_{ref} falls below 0.3 (i.e. the label is defined as $\mathbf{1}_{RougeL(y,y_{ref})<0.3}$), following the same metric used in Kossen et al. (2024) [2]. We retain 2500 data labeled as hallucinations and 2500 labeled as correct, constructing a balanced 5K dataset for each task.

Models. We test on these models: `Llama3.2-1B-Instruct`, `Llama3-8B-Instruct`, `Qwen2.5-1.5B-Instruct`, `Qwen3-8B`, `Qwen3-14B` and `Mistral-7B-Instruct`. For sampling-based methods, we generate candidate responses with temperature 1. For embeddings, we use the same sentence-transformer as in Section 4 (ablation study of different sentence-transformers can be found in Appendix G).

Evaluation. We compare the predicted binary hallucination labels against the ground truth labels and report both accuracy and F1 score. Furthermore, we also add the evaluation based on the AUROC (area under the receiver operator characteristic curve) metric, which compares the raw uncertainty scores against the ground truth labels. In fact, for a given uncertainty measure $m(\cdot)$, the AUROC score is equivalent to $\mathbb{P}[m(y_{hallucinated}) > m(y_{correct})]$, where $y_{hallucinated}$ and $y_{correct}$ are randomly chosen hallucinated and correct answers, respectively. Hence a higher AUROC (closer to 1) indicates that the uncertainty measure more effectively distinguishes hallucinated responses by assigning them higher uncertainty scores. AUROC is a widely used metric in many existing studies on response hallucination detection (Kuhn, Gal, and Farquhar 2023; Kossen et al. 2024; Kadavath et al. 2022).

Baselines. We adapt the baseline methods from Section 4.1, applying them to sampled responses instead of augmented queries.

5.2 Results

The performance results for `Llama3.2-1B-Instruct` under both TriviaQA and SQuAD tasks are presented in Tables 2 (results for other models can be found in Appendix H). For sampling-based methods, we set the response sampling size to $n = 20$. Notably, our semantic volume method significantly outperforms all baselines in both accuracy and F1 score. Furthermore, the AUROC results confirm that our semantic volume serves as a highly effective uncertainty signal for hallucination detection. Additionally, we observe that when comparing p(True), which includes sampled candidate responses as context, to direct prompting, the inclusion of context degrades the performance. Moreover, for methods that rely on prompting LLMs (including pTrue), we find that LLMs sometimes exhibit a strong bias toward answering almost all "Yes" or all "No". In fact in Table 2, both *Prompt Llama3.2-1B-Instruct* and *pTrue (Llama3.2-1B-Instruct)* exhibit this behavior, nearly predicting all responses as hallucinations (further discussions are provided in Appendix H). This instability highlights another drawback of such methods that rely on LLM prompting for uncertainty estimation. From both tables, we observe that sampling-based methods that measure the dispersion of sampled responses (particularly lexical similarity and semantic entropy) generally outperform probability-based methods, which aligns with the findings in Cole et al. (2023).

5.3 Distribution Separation

In this section, we compare the distributions of various uncertainty measures using visualization and the Kolmogorov–Smirnov (KS) test (Smirnov 1948). Specifically, we plot histograms for the hallucinated subset (label 1) and the correct subset (label 0) from our TriviaQA dataset. Ideally, a well-performing measure should yield two distinct bulks, with greater separation indicating stronger discriminative power. To quantitatively assess the separation, we perform a two-sample Kolmogorov–Smirnov test, a non-parametric test that compares two empirical distributions by measuring the maximum distance between their empirical cumulative distribution functions. A large KS statistic combined with a small p-value suggests that the two distributions are significantly different.

The plots and statistics are shown in Figure 2. Combined the KS statistics and the histograms, we observe that indeed *Last Token Entropy* and *Log Probabilities* struggle to effectively separate the two groups of data, while *Lexical Similarity*, *Semantic Entropy*, and our *Semantic Volume* exhibit stronger separation. Particularly, we note that the distribution of semantic entropy closely resembles that of our semantic volume measure. In fact, we will provide theoretical analysis in Section 7 showing that our semantic volume can be interpreted as the differential entropy of the semantic embedding vectors, and can be viewed as a more general and continuous version of semantic entropy.

6 Ablation Study and Hyperparameter Analysis

The original embedding dimension from the sentence-transformer is 1536. In the external uncertainty experiment, we reduce the dimensionality using PCA with $d = 10$, while in the internal uncertainty experiment, we set $d = 20$. Ablation studies confirm PCA-projected vectors $\tilde{\boldsymbol{V}}$ outperform

[2] We also conducted a human validation on 1,000 randomly sampled examples and found a 95.1% agreement with this labeling criterion.

Method	CLAMBER		AmbigQA	
	Acc.	F1	Acc.	F1
Vicuna-13B (zero-shot)	50.6	39.9	45.1	19.5
Llama2-13B-Instruct (zero-shot)	45.6	43.6	46.2	22.4
Llama2-70B-Instruct (zero-shot)	50.3	34.2	49.4	64.5
Llama3.2-3B-Instruct (zero-shot)	51.5	37.7	49.3	33.8
ChatGPT (zero-shot)	54.3	53.4	54.8	52.8
ChatGPT (few-shot)	51.6	49.2	54.9	53.1
ChatGPT (zero-shot + CoT)	57.3	56.9	54.2	55.1
ChatGPT (few-shot + CoT)	53.6	51.4	54.0	50.8
Last Token Entropy	52.2	67.3	48.1	63.2
Log Probabilities	45.5	66.0	50.3	62.8
pTrue (Llama3-8B-Instruct)	$52.6_{1.19}$	$53.2_{1.44}$	$50.1_{0.88}$	$28.1_{2.60}$
pTrue (Mistral-7B-Instruct)	$47.2_{1.16}$	$26.3_{2.28}$	$49.8_{0.35}$	$24.3_{3.02}$
Lexical Similarity	$52.8_{0.41}$	$53.7_{0.75}$	$48.9_{1.07}$	$24.9_{2.28}$
Semantic Entropy	$50.2_{0.33}$	$62.8_{0.56}$	$51.4_{0.97}$	$61.3_{2.59}$
Semantic Volume (ours)	$\mathbf{58.0}_{0.18}$	$\mathbf{69.1}_{0.32}$	$\mathbf{55.7}_{0.95}$	$\mathbf{67.6}_{0.83}$

Table 1: External uncertainty: Accuracy and F1 on CLAMBER and AmbigQA. Sampling-based methods report mean ± std. over three trials (std. as subscript).

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.5	66.8	N/A	50.6	22.7	N/A
Prompt Llama3-8B-Instruct	65.5	60.1	N/A	64.0	65.5	N/A
Prompt Mistral-7B-Instruct	68.7	61.8	N/A	60.7	57.3	N/A
Last Token Entropy	60.1	59.9	63.9	53.6	34.2	56.6
Log Probabilities	60.1	62.9	65.5	54.8	46.4	56.7
pTrue (Llama3.2-1B-Instruct)	$49.5_{0.37}$	$64.4_{0.81}$	$61.2_{0.91}$	$50.1_{0.81}$	$9.5_{3.26}$	$52.3_{1.11}$
pTrue (Llama3-8B-Instruct)	$63.7_{2.04}$	$45.4_{4.23}$	$56.9_{3.25}$	$58.2_{0.96}$	$45.4_{1.63}$	$53.5_{1.30}$
pTrue (Mistral-7B-Instruct)	$62.8_{0.86}$	$46.3_{1.12}$	$65.4_{1.26}$	$51.5_{1.22}$	$11.3_{3.65}$	$51.7_{0.89}$
Lexical Similarity	$64.6_{0.61}$	$72.2_{0.36}$	$73.3_{0.39}$	$58.3_{0.54}$	$59.9_{0.66}$	$62.4_{0.71}$
Semantic Entropy	$63.8_{0.98}$	$69.7_{1.08}$	$73.9_{0.75}$	$61.8_{1.51}$	$68.8_{1.75}$	$64.9_{1.17}$
Semantic Volume (ours)	$\mathbf{72.4}_{0.55}$	$\mathbf{75.5}_{0.34}$	$\mathbf{79.7}_{0.16}$	$\mathbf{64.7}_{0.49}$	$\mathbf{71.2}_{0.63}$	$\mathbf{68.8}_{0.56}$

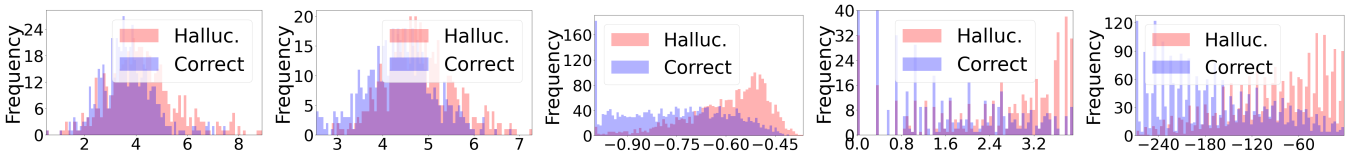
Table 2: Internal uncertainty: Accuracy, F1, and AUROC on TriviaQA and SQuAD. "N/A" indicates AUROC is not applicable (no probabilistic scores). For sampling-based methods, we report mean ± std. over three trials (std. shown as subscript).

raw embeddings (see Appendix E). This suggests that lower-dimensional projections help separate perturbation vectors more effectively. Furthermore, we explore various values of d to analyze the effect of dimensionality on performance. Our findings indicate that the optimal dimension d is task-dependent, with diminishing improvements beyond a certain point. Nonetheless, it is important to note that even without PCA dimension reduction, our method still outperforms various baselines in both external and internal uncertainty detection tasks (see Figure 7).

To demonstrate the generalizability of our method and study the effect of various hyperparameters and model choices. In Appendix F, we analyze the effect of varying n, the number of perturbations. As expected, larger n yields more accurate uncertainty estimation but increases computational cost, while smaller n reduces cost but may sacrifice accuracy. We choose $n = 20$ to balance performance and efficiency. In Appendices G and H, we examine the impact of different embedding models and response generation models. We find that larger embedding models provide little additional performance gain. Furthermore, when detecting hallucinations in responses generated by a larger LLM, the performance of most methods slightly declines. However, our Semantic Volume method still outperforms all baselines.

7 Theoretical Analysis

In this section, we present theoretical analyses, with proofs and additional supporting lemmas provided in Appendix C. First, we derive the exact formula for the optimal threshold τ^* in Proposition 1. Then in Theorems 1 and 2, we show that under Gaussian distribution assumptions of the perturbations, our semantic volume measure effectively computes the differential entropy of the embedding vectors, using entropy as a measure of uncertainty detection. This insight allows us to naturally view our method as a generalization of semantic entropy (Kuhn, Gal, and Farquhar 2023). Notably, semantic entropy involves a manual clustering step, and only considers the entropy between clusters while ignoring discrepancies



(a) Last Token Entropy (KS: 0.238, p-val: 1e-36) (b) Log Probabilities (KS: 0.226, p-val: 5e-27) (c) Lexical Similarity (KS: 0.428, p-val: 1e-26) (d) Semantic Entropy (KS: 0.376, p-val: 7e-32) (e) Semantic Volume (KS: 0.446, p-val: 3e-54)

Figure 2: Distribution separation across different uncertainty measures. "Halluc." denotes hallucination. Each subfigure shows histograms of model outputs for two classes with Kolmogorov–Smirnov statistics.

within clusters (since the responses in the same cluster are semantically similar but not exactly identical). In contrast, **our method more generally captures the overall semantic dispersion across all sampled perturbations**, providing a more comprehensive uncertainty measure.

Proposition 1 (Formula for optimal threshold τ^*)**.** *Denote $\mathcal{L} \in \mathcal{D}$ as a labeled subset with inputs $\{s_i\}$ and labels $\{y_i\}$:*

$$\mathcal{L} = \{(s_i, y_i)\}_{i=1}^{M} \quad \textit{with} \quad y_i \in \{0, 1\},$$

For each s_i, denotes its semantic volume as $m(s_i)$. Define a classification rule

$$\hat{y}_i(\tau) = \begin{cases} 1, & m(s_i) > \tau, \\ 0, & \textit{otherwise}. \end{cases}$$

Then the optimal threshold τ that maximizes the F_1 score on $\mathcal{L}$ is given by

$$\begin{aligned} \tau^* &\overset{def}{=} \arg\max_{\tau \in \mathbb{R}} F_1(\tau) \\ &= \arg\max_{\tau \in \mathbb{R}} \left(\frac{2 \sum_{i=1}^{M} \mathbf{1}_{\hat{y}_i(\tau) = y_i = 1}}{\sum_{i=1}^{M} \mathbf{1}_{\hat{y}_i(\tau) = 1} + \mathbf{1}_{y_i = 1}} \right). \end{aligned} \tag{4}$$

Note that the F_1 score can be replaced by other metrics, such as the accuracy.

Theorem 1. *Denote the embedding vector and normalized embedding vector of the original text (either a query or a response) as $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{v}}$, respectively. Denote the perturbation embeddings as $\boldsymbol{X} \overset{def}{=} [\boldsymbol{x}_1 \boldsymbol{x}_2 \dots \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ and the normalized perturbation embeddings as $\boldsymbol{V} \overset{def}{=} [\boldsymbol{v}_1 \boldsymbol{v}_2 \dots \boldsymbol{v}_n] \in \mathbb{R}^{d \times n}$ (i.e. $\boldsymbol{x}_i = \boldsymbol{v}_i / \|\boldsymbol{v}_i\|$ for each $i \in [n]$). Assume Gaussian distribution $\boldsymbol{x}_i \sim \mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$. Then in high-dimensional regime, $\log\det(\boldsymbol{V}^\top \boldsymbol{V})$ corresponds to the shifted differential entropy of the perturbations $\{\boldsymbol{x}_i\}_{k \in [n]}$. That is,*

$$\log\det(\boldsymbol{V}^\top \boldsymbol{V}) \doteq \mathcal{H}(\boldsymbol{X}) + C,$$

where $\mathcal{H}(\boldsymbol{X}) \overset{def}{=} -\mathbb{E}_{x \sim \boldsymbol{X}} [\log p_{\boldsymbol{X}}(x)]$ is the differential entropy and C is a constant offset term.

Then we obtain the following Theorem, which is a direct consequence of Theorem 1 and Lemma 1 in Appendix C.

Theorem 2. *Under the same setting and notations of the Theorem 1, our **Semantic Volume** method essentially generates same binary decisions compared to using **differential entropy** of the perturbation embedding vectors. That is, denote the Semantic Volume measure and differential entropy measure as $m(\cdot)$ and $\tilde{m}(\cdot)$ respectively. For the labels*

$$y_i \overset{def}{=} \mathbf{1}_{m(s_i) < \tau^*} \quad \textit{and} \quad \tilde{y}_i \overset{def}{=} \mathbf{1}_{\tilde{m}(s_i) < \tilde{\tau}^*},$$

where $\tilde{\tau}^*$ is the optimal threshold for the differential entropy measure, we have

$$\tilde{y}_i = y_i \quad \textit{for all } s_i \in \mathcal{D} \setminus \mathcal{L}.$$

8 Conclusion

One limitation of our current study is that our work considers the same scope as the references in Section 2.3: we focus on the situation where uncertainty aligns with incorrectness, without addressing *confidently wrong* responses, meaning an LLM hallucinates with high confidence (low uncertainty). We believe that addressing such cases requires different strategies, such as factuality checking or incorporating external knowledge for verification. We leave these explorations for future work.

In summary, we have introduced *Semantic Volume*, a novel and general-purpose measure for detecting both *external uncertainty* (query ambiguity) and *internal uncertainty* (response uncertainty) in large language models. By generating perturbations, embedding these perturbations as normalized vectors, and computing the determinant of their Gram matrix, we obtain a measure that captures the overall semantic dispersion. Extensive experiments on benchmark datasets showed that semantic volume significantly outperforms various types of existing baselines (prompting-based, probability-based, and sampling-based) for both ambiguous query classification and response hallucination detection. Furthermore, from a theoretical standpoint, we established that semantic volume can be viewed as the differential entropy of the embedding vectors, thereby unifying and extending prior sampling-based metrics (e.g., semantic entropy). This interpretation highlights why our measure is robust and comprehensive: unlike purely clustering-based approaches, we account for the *overall* dispersions in the embedding space. Moreover, our method is applicable even when the LLM is only accessible via an external API or when internal model details are unavailable, making it broadly practical across closed-source or API-based models. Overall, our findings suggest that semantic volume is a promising step toward more reliable, interpretable uncertainty detection for both external and internal uncertainty of LLMs.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI, M. 2024. LLaMA 3.2: Advancing AI for Vision, Edge, and Mobile Devices. Meta AI Blog. Accessed: Feb 11, 2025.

AI@Meta. 2024. Llama 3 Model Card.

Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Anthropic. 2023. Introducing the Claude 3 family of models. Accessed: 2025-01-02.

Arora, U.; Huang, W.; and He, H. 2021. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bastounis, A.; Campodonico, P.; van der Schaar, M.; Adcock, B.; and Hansen, A. C. 2024. On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of'I don't know'. *arXiv preprint arXiv:2408.02357*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, X.; Li, M.; Gao, X.; and Zhang, X. 2022. Towards improving faithfulness in abstractive summarization. *Advances in Neural Information Processing Systems*, 35: 24516–24528.

Chi, Y.; Lin, J.; Lin, K.; and Klein, D. 2024. CLARINET: Augmenting Language Models to Ask Clarification Questions for Retrieval. *arXiv preprint arXiv:2405.15784*.

Cole, J. R.; Zhang, M. J.; Gillick, D.; Eisenschlos, J. M.; Dhingra, B.; and Eisenstein, J. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; and Specia, L. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555.

Grewal, Y. S.; Bonilla, E. V.; and Bui, T. D. 2024. Improving Uncertainty Quantification in Large Language Models via Semantic Embeddings. *arXiv preprint arXiv:2410.22685*.

Guerreiro, N. M.; Alves, D. M.; Waldendorf, J.; Haddow, B.; Birch, A.; Colombo, P.; and Martins, A. F. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11: 1500–1517.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Harville, D. A. 1998. Matrix algebra from a statistician's perspective.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ji, Z.; Chen, D.; Ishii, E.; Cahyawijaya, S.; Bang, Y.; Wilie, B.; and Fung, P. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kim, H. J.; Kim, Y.; Park, C.; Kim, J.; Park, C.; Yoo, K. M.; Lee, S.-g.; and Kim, T. 2024. Aligning Language Models to Explicitly Handle Ambiguity. *arXiv preprint arXiv:2404.11972*.

Kossen, J.; Han, J.; Razzak, M.; Schut, L.; Malik, S.; and Gal, Y. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Kulesza, A.; Taskar, B.; et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3): 123–286.

Lee, D.; Kim, S.; Lee, M.; Lee, H.; Park, J.; Lee, S.-W.; and Jung, K. 2023. Asking clarification questions to handle ambiguity in open-domain qa. *arXiv preprint arXiv:2305.13808*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, X.; Gao, M.; Zhang, Z.; Yue, C.; and Hu, H. 2024. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Liu, L.; Pan, Y.; Li, X.; and Chen, G. 2024. Uncertainty Estimation and Quantification for LLMs: A Simple Supervised Approach. *arXiv preprint arXiv:2404.15993*.

LMSys. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. LMSys Blog. Accessed: Feb 11, 2025.

Malinin, A.; and Gales, M. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Min, S.; Michael, J.; Hajishirzi, H.; and Zettlemoyer, L. 2020. AmbigQA: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.

Quevedo, E.; Yero, J.; Koerner, R.; Rivas, P.; and Cerny, T. 2024. Detecting Hallucinations in Large Language Model Generation: A Token Probability Approach. *arXiv preprint arXiv:2405.19648*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250*.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Smirnov, N. 1948. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2): 279–281.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? *arXiv preprint arXiv:2305.18153*.

Zhang, M. J.; and Choi, E. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

Zhang, T.; Qin, P.; Deng, Y.; Huang, C.; Lei, W.; Liu, J.; Jin, D.; Liang, H.; and Chua, T.-S. 2024. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. *arXiv preprint arXiv:2405.12063*.